Comptes rendus des 12-èmes Rencontres de la Société Francophone de Classification

Vladimir Makarenkov, Guy Cucumel et

François-Joseph Lapointe, Editeurs

Montréal, 30 Mai – 1er Juin 2005

UQÀM
Université du Québec à Montréal

sfc

Université de Montréal

12-èmes Rencontres de la Société Francophone de Classification

---

Comptes rendus

Vladimir Makarenkov
Guy Cucumel
François-Joseph Lapointe
Editeurs

Université du Québec à Montréal

COMITÉ DE PROGRAMME

Président : **Guy Cucumel** (Université du Québec à Montréal)

Jean-Pierre Asselin de Beauville (Agence universitaire de la Francophonie)
Cédric Chauve (Université du Québec à Montréal)
Marie Chavent (Université de Bordeaux 3)
François-Joseph Lapointe (Université de Montréal)
Vladimir Makarenkov (Université du Québec à Montréal)
Pascale Rousseau (Université du Québec à Montréal)

COMITÉ SCIENTIFIQUE

Président : **Vladimir Makarenkov** (Université du Québec à Montréal)
Vice-président : **François-Joseph Lapointe** (Université de Montréal)

Patrice Bertrand (ENST-Bretagne, Brest)
Hans Bock (Institut de la statistique, Aix-la-Chapelle)
Marc Bourdeau (École Polytechnique de Montréal)
Maria Paula Brito (Université de Porto)
Sergio Camiz (Université de Rome "La Sapienza")
Cédric Chauve (Université du Québec à Montréal)
Alain Guénoche (Institut de mathématiques de Luminy)
André Hardy (Facultés universitaires Notre-Dame de la Paix)
Georges Hébrail (ENST de Paris)
Pierre Legendre (Université de Montréal)
Giuseppe Melfi (Université de Neuchâtel)
Amedeo Napoli (LORIA, CNRS)
Marcel Rémon (Facultés universitaires Notre-Dame de la Paix)
Gilbert Saporta (CNAM, Paris)
Marylène Troupé (Université des Antilles-Guyane**)**

Préface

Après Bordeaux en 2004, Montréal, deuxième ville francophone du monde, est heureuse d'accueillir du 30 mai au 1er juin 2005 les 12èmes Rencontres de la Société Francophone de Classification.

Organisées pour la première fois en Amérique du Nord, ces rencontres vont permettre à des chercheurs venant des cinq continents de présenter leurs travaux dans les différentes branches de la recherche en Classification et d'échanger entre eux.

Nous tenons à remercier Monsieur Roch Denis, Recteur de l'Université du Québec à Montréal pour la réception qu'il organise pour les participants, ainsi que Monsieur Gérald Tremblay, Maire de Montréal, qui recevra les conférenciers invités, les membres du bureau de la SFC ainsi que les membres des comités d'organisation et scientifique dans les salons de l'Hôtel de Ville.

Nous exprimons notre reconnaissance à tous ceux qui nous ont soutenu financièrement et sans qui nous n'aurions pu organiser cette manifestation: l'Agence Universitaire de la Francophonie, l'Association des Statisticiennes et Statisticiens du Québec, la Coop UQAM, la Faculté des sciences de l'Université du Québec à Montréal, la Fondation « La Science Statistique », Génome Québec, Les Laboratoires Universitaires Bell, la Société Statistique de Montréal.

Nos remerciements s'adressent également à tous les orateurs de ces rencontres et, notamment, aux conférenciers invités : Francisco de A.T. de Carvalho, Pierre Hansen, Melvin F. Janowitz, Sabin Krolak-Schwerdt, Bruno Leclerc et Maurizio Vichi.

Enfin, nous remercions les membres des Comités scientifiques et d'organisation, ainsi que le personnel de l'UQAM qui ont contribué à la préparation de ces journées et plus particulièrement Jenny Desrochers du Service des communications de l'UQAM.

Montréal, le 30 mai 2005 Les éditeurs

Table des matières

Préface

Conférences plénières

Communications

Prix Simon Régnier

Conférences plénières

Classification Automatique et Fonctions de Proximités en Analyse des Données Symboliques

Francisco de A.T. de Carvalho

Centro de Informatica - CIn,
Universidade Federal de Pernambuco,
Av. Prof. Luiz Freire, s/n – Cidade Universitára
CEP : 50740-540, Recife-PE, Brésil
{ealn,fatc}@cin.ufpe.br

RÉSUMÉ. Dans ce travail nous présentons plusieurs fonctions de proximité qui peuvent être utilisées pour obtenir des partitions d'objets symboliques par des algorithmes de types « nuées dynamiques ». Ces méthodes de nuées dynamiques peuvent être appliquées directement au tableau individus - variables ou peuvent être appliquées sur des tableaux de proximité.

MOTS-CLÉS : Analyse des Données Symboliques, Données Symboliques, Fonctions de Proximités, Algorithme de type Nuées Dynamiques.

1 Introduction

Grâce à la technologie informatique, de vastes ensembles de données sont recueillis et il est nécessaire de les résumer. Actuellement plusieurs approches ont été proposées pour l'extraction de connaissances, la découverte de régularités et la simplification de ces données. Notre approche est l'approche symbolique en classification et en analyse de données. Son point de départ est l'extraction de connaissances de ces grandes bases de données, comme en "data mining". Ces connaissances sont modélisées par des objets plus complexes, appelées "objets symboliques", décrits par des variables symboliques qui peuvent prendre comme valeur non seulement une catégorie ou une valeur numérique comme dans les approches classiques mais aussi un ensemble de valeurs, un intervalle, une distribution de fréquence car ces objets peuvent correspondre à des groupes d'individus et il faut tenir compte de leur variabilité. L'étape suivante est l'extension des méthodes et algorithmes usuels de l'extraction de connaissances à ces données plus complexes, passant ainsi du "data mining" au "knowledge mining".

Dans ce travail nous présentons plusieurs fonctions de proximité qui peuvent être utilisés pour obtenir des partitions d'objets symboliques par des algorithmes de types « nuées dynamiques ». Ces méthodes de nuées dynamiques peuvent être appliquées directement au tableau individus - variables ou peuvent être appliquées sur des tableaux de proximité.
Dans le premier cas nous allons nous restreindre à des données de type quantitatives continue, de type intervalle ou un mélange de ces deux types: les méthodes de nuées dynamiques seront alors basées soit sur des distances adaptatives du type Mahalanobis ([SOU 04]), soit sur des distances adaptatives et non adaptatives du type city-block, euclidienne, Chebyshev, ou plus généralement du type Minkowsky ([CHA 02, CHA 03, DEC 04, DEC a, DEC b, SOU 04a, SOU04b]).

L'algorithme d'optimisation utilisé est de type Nuées Dynamiques ([DID 78]) et il consiste à utiliser alternativement une étape de *représentation*, où la partition est fixée et pour laquelle on cherche le meilleur représentant de chaque classe au sens de la distance choisie, et une étape d'*allocation*, où les représentants sont fixés et dans laquelle à affecter chaque individu à la classe dont le représentant lui est le plus semblable. On recommence ces étapes jusqu'à la convergence.
Dans le deuxième cas nous allons considérer des objets symboliques décrits par différents types de variables symboliques (catégoriques multi-valuées, de type intervalle ou de type modal). Lors du calcul de la proximité entre ces unités statistiques, il est nécessaire de tenir compte à la fois de la variabilité (disjonction des valeurs relatives à une variable) et de la connaissance du domaine (dépendance hiérarchique entre variables). Ces dépendances hiérarchiques sont exprimées par des règles ([BOC 00]).
Pour les données symboliques Booléennes (celles décrites par des variables catégoriques multi-valuées ou de type intervalle) nous allons considérer deux familles d'indices de proximités.
La première famille utilise pour chaque variable une fonction de comparaison pour mesurer à la fois la différence de contenu et la différence de position (dans le cas où les données seraient ordonnées, i.e., intervalles, ensemble de catégories ordonnées) et une fonction d'agrégation. Les fonctions de comparaison utilisent des opérateurs symboliques (union et intersection symboliques) et celles qui mesurent la différence de contenu peuvent être basées sur les indices usuels de comparaison des tableaux binaires. La fonction d'agrégation s'inspire de la métrique de Minkowsky ([CHA 03, DEC 94, DEC 98b, DEC 00]).
La seconde famille n'utilise pas de fonction d'agrégation. Elle utilise une fonction de comparaison globale, qui tient compte de toutes les variables à la fois, pour mesurer la différence de contenu et la différence de volume. Ces fonctions de comparaison utilisent aussi des opérateurs symboliques (union et intersection symboliques) et celles qui mesurent la différence de contenu peuvent être basées sur les indices usuels de comparaison des tableaux binaires ([DEC 98b]).
Le problème majeur lié à toutes ces approches est celui de l'aspect combinatoire du calcul lors de la prise en compte des dépendances hiérarchiques. Il est linéaire en fonction du nombre de variables et, malheureusement, exponentiel en fonction du nombre de règles. Cette difficulté nous a amenés à l'introduction de la Forme Normale Symbolique ([CSE 99, CSE 01, CSE 02]). La Forme Normale Symbolique (NSF), inspiré de la $3^{ème}$ Forme Normale des bases de données relationnelles, consiste à factoriser les objets symboliques selon les contraintes exprimées par des règles entre les variables de telle façon que, dans la plu part des cas, les calculs s'effectuent dans un temps qui n'est quasiment plus affecté par la présence des règles. La transformation elle-même étant effectuée en un temps polynomial en fonction du nombre d'objets.
Une variable modale Y définie sur un ensemble $E = \{\omega_1, \omega_2, \ldots\}$ de domaine $D = \{m_1, \ldots, m_k\}$ est une application $Y(\omega) = (U(\omega), q(\omega))$, où $\omega \in E$, et où $q(\omega)$ est une distribution de poids sur le domaine D et $U(\omega) \subseteq D$ est le support de $q(\omega)$ dans D ([BOC 00]).
Pour les données symboliques modales (celles décrites par des variables symboliques du type modal) les indices de proximités utilisent pour chaque variable une fonction pour comparer les supports et une autre fonction pour comparer les distributions de poids. Ces comparaisons sont ensuite réunies par une fonction d'agrégation. La fonction de comparaison des distributions peu être du type city-block, euclidienne ([CHA 03]), Chebyshev, issue du coefficient d'affinité ([BOC 00] p. 160) ou issue des mesures de généralité pour des données de type modal ([BRI 02]).
L'algorithme d'optimisation utilisé est encore de type Nuées Dynamiques ([DID 78]) et il consiste à utiliser alternativement une étape de *représentation* et une étape d'*allocation*. Lors de l'étape de représentation la partition est fixée et pour chacune des classes on cherche l'individu pour lequel la somme des distances aux individus de la même classe est minimum. Cet individu est le meilleur représentant de la classe au sens de la distance choisie. Lors de l'étape d'*allocation*, les représentants sont fixés et on cherche à affecter chaque individu à la classe dont le représentant lui est plus semblable. On recommence ces étapes jusqu'à la convergence.

2 Bibliographie

[BRI 02] BRITO, P., DE CARVALHO, F.A.T., «Symbolic Clustering of Constrained Probabilistic Data", *Exploratory Data Analysis in Empirical Research: Proceedings of the 25th Annual Conference of the German Classification Society, Gfkl-2001,* Munich (Germany), Schwaiger, M and Opitz, O. Eds., 2003, p.12—21, Springer, Berlin Heidelberg.

[BOC 00] BOCK H-H., DIDAY, E., *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data*, Springer, 2000.

[CHA 02] CHAVENT, M., LECHEVALLIER, Y, "Dynamical Clustering of Interval Data Optimization of an Adequacy Criterion Based on Haussdorf Distance", *Classification, Clustering, and Data Analysis: Proceedings of the 8th Conference of the International Federation of Classification Societies, IFCS-2002,* Krakow (Poland), Jajuga, K. et al Eds, 2002, p. 53—60, Springer, Berlin Heidelberg.

[CHA 03] CHAVENT, M., DE CARVALHO, F. A. T., LECHEVALLIER, Y., VERDE, R., "Trois nouvelles méthodes de classification automatique de données symboliques de type intervalle", *Revue de Statistique Appliquée,*, v.LI, n.4, 2003, p.5—29.

[CSE 99] CSERNEL, M., DE CARVALHO, F.A.T., "Usual operations with symbolic data under normal symbolic form", *Applied Stochastic Models in Business and Industry*, v 11, 1999, p.241—257.

[CSE 01] CSERNEL, M., DE CARVALHO, F.A.T., "On memory requirement with normal symbolic form", *Exploratory Data Analysis in Empirical Research: Proceedings of the 25th Annual Conference of the German Classification Society, Gfkl-2001,* Munich (Germany), Schwaiger, M and Opitz, O. Eds., 2003, p. 22—30, Springer, Berlin Heidelberg.

[CSE 02] CSERNEL, M., DE CARVALHO, F.A.T., "Modelling memory requeriment with normal symbolic form", *Classification, Clustering and Data Analysis Proceedings of the 8th Conference of the International Federation of Classification Societies, IFCS-2002,* Krakow (Poland), Jajuga, K. et al Eds, 2002*,* p.289—296, Springer, Berlin Heidelberg.

[DEC 94] DE CARVALHO, F.A.T, "Proximity coefficients between Boolean symbolic objects", *New Approaches in Classification and Data Analysis: Proceedings of the 4th Conference of the International Federation of Classification Societies, IFCS-1994,* Paris (France), Diday et al Eds, 1994, p.387—394, Springer, Berlin Heidelberg.

[DEC 98a] DE CARVALHO, F.A.T, "Extension based proximities between constrained Boolean symbolic objects", *Data Science, Classification and Related Methods: Proceedings of the 5th Conference of the International Federation of Classification Societies, IFCS-1996,* Tokyo (Japan), Hayashi, C. et al, 1998, p.370—378, Springer, Berlin Heidelberg.

[DEC 98b] DE CARVALHO, F.A.T., SOUZA, R.M.C.R., "Statistical proximity functions of Boolean symbolic objects based on histograms", *Advances in Data Science and Classification: Proceedings of the 6th Conference of the International Federation of Classification Societies, IFCS-1998,* Rome (Italy), Rizzi, A. et al Eds, 1998, p.391—396, Springer, Berlin Heidelberg.

[DEC 99] DE CARVALHO, F. A. T., VERDE, R., LECHEVALLIER, Y., "A dynamical clustering of symbolic objects based on a context dependent proximity measure", *Proceedings of the IX International Symposium on Applied Stochastic Models and Data Analysis, AMSDS-1999,* Lisboa (Portugal), Bacelar-Nicolau, H. et al Eds, p.237—242, LEAD, Universidade de Lisboa.

[DEC 00] DE CARVALHO, F.A.T., DIDAY, E., "Un Indice de Proximite entre Objets Symboliques qui tient compte des Contraintes dans l'Espace de Description", *Induction Symbolique et Numérique à partir des Données*, Diday, E. et al Eds, 2000, p.225—246, Cépadues Éditions, Toulouse.

[DEC 04] DE CARVALHO, F. A. T., LECHEVALLIER, Y., SOUZA, R. M. C. R, "A dynamic cluster algorithm based on adaptive L_r distances for quantitative data", *Classification, Clustering and Data*

Mining Applications: Proceedings of the 9th Conference of the International Federation of Classification Societies, *IFCS-2004*, Chicago (USA), Banks, D. et al Eds, 2004, p.33—42, Springer, Berlin Heidelberg.

[DEC a] DE CARVALHO, F.A.T., SOUZA, R.M.C.R, CHAVENT, M., LECHEVALLIER, Y. "Adaptive Hausdorff distance and dynamic clustering of interval data", soumis à Pattern Recognition Letters.

[DEC b] DE CARVALHO, F.A.T., BRITO, P., BOCK, H.-H., "Dynamic clustering for interval data based on L_2 distance", soumis à Pattern Recognition.

[DID 78] DIDAY, E., GOVAERT, G., LECHEVALLIER, Y., SIDI, J., "Clsutering in pattern recognition", *Proceedings of the 4th Joint International Conference on Pattern Recognition*, Kyoto, Japan.

[SOU 04a] SOUZA, R. M. C. R., DE CARVALHO, F. A. T., LECHEVALLIER, Y., "Dynamic cluster methods for interval data based on Mahalanobis distances", *Classification, Clustering and Data Mining Applications: Proceedings of the 9th Conference of the International Federation of Classification Societies*, *IFCS-2004*, Chicago (USA), Banks, D. et al Eds, 2004, p.351 – 360, Springer, Berlin Heidelberg.

[SOU 04b] SOUZA, R. M. C. R., DE CARVALHO, F. A. T., "Clustering of Interval Data based on City-Block Distances", *Pattern Recognition Letters*, v.25, n.3, 2004, p.353 – 365.

[VER 00] VERDE, R., DE CARVALHO, F. A. T., LECHEVALLIER, Y., "A dynamical clustering algorithm for multi-nominal data", *Data Analysis, Classification and Related Methods: Proceedings of the 7th Conference of the International Federation of Classification Societies*, *IFCS-2000,* Namur (Belgium), Kiers, H.A.L. et al Eds, 2000, p.387—394, Springer, Berlin Heidelberg.

[VER 01] VERDE, R., DE CARVALHO, F.A.T., LECHEVALLIER, Y., "A Dynamical Clustering Algorithm for Symbolic Data", *Tutorial on Symbolic Data Analysis*, *25th Annual Conference of the German Classification Society*, Munich (Germany), DIDAY, E., LECHEVALLIER, Y. Eds, 2001, p.59—72.

DISCRIMINATION LINÉAIRE ET PROGRAMMATION MATHÉMATIQUE

Pierre Hansen

GERAD et HEC Montréal,
École des Hautes Études Commerciales 3000,
ch. de la Côte-Sainte-Catherine Montréal (Québec)
Canada H3T 2A7

Résumé : Le problème de la discrimination linéaire s'énonce comme suit : étant donné deux ensembles A et B d'observations (ou de points ou d'entités) dans R^+ déterminer un hyperplan P qui sépare le mieux possible les points de A de ceux de B. Il est bien connu que si les fermetures convexes des points de A et des points de B ont une intersection vide, ce problème peut être résolu par programmation linéaire. Si ce n'est pas le cas, il faut prévoir une mesure de l'erreur. Les plus courantes sont le taux d'erreur et la somme des distances des points mal classés à l'hyperplan, mesurés dans une norme L_p. Ces erreurs peuvent être mesurées sur l'ensemble de points initial (ou ensemble d'entraînement) ou, de manière plus significative, sur un second ensemble de points (ou ensemble de test).
De nombreuses formulations de ces problèmes en terme de programmation mathématiques ont été proposées par divers auteurs. Elles sont souvent erronées; le cas des distances à l'hyperplan en norme L_p a été clarifié par Mangasarian [MAN 99].

Les buts de l'exposé sont les suivants :

i. Passer en revue les principaux modèles de discrimination linéaire ;
ii. Présenter des algorithmes exacts pour les critères des taux d'erreur et de la somme des distances en norme L_p avec p =2 et p=8 ;
iii. Présenter une heuristique de recherche à voisinage variable, voir [HAN 01] et [MLA 97], pour la discrimination de très grand ensembles selon le critère de la somme des distances pour une norme L_p quelconque avec $0 = p = 8$;
iv. Examiner empiriquement la performance des algorithmes proposés, et notamment l'effet du paramètre p.

Bibliographie

[HAN 01] P. HANSEN et N. MLADENOVIC, Variable Neighborhood Search: Principles and Applications, *European Journal of Operational Research* 130, 2001, 449-467.

[MAN 99] O. L. MANGASARIAN, Arbritrary norm separating plane, *Operations Research Letters* 24, 1999, 15-23.

[MLA 97] N. MLADENOVIC et P. HANSEN, Variable Neighborhood Search , *Computers and Operations Research*, 24, 1997, 1097-1100.

Cluster analysis based on posets [1]

M. F. Janowitz

DIMACS
Rutgers University
96 Frelinghuysen Road
Piscataway, NJ, USA 08854-8018

ABSTRACT. When dissimilarities are measured in a space other than the reals, it is argued that previous models for cluster analysis are not adequate. Possible new order theoretic models will be explored. It is also shown that formal concept analysis may be viewed as a special case of a Boolean dissimilarity coefficient.

KEYWORDS: cluster analysis, Boolean dissimilarity, formal concept analysis

---

1 Background

A basic knowledge of cluster analysis will be assumed at the outset. This can be obtained by consulting where needed one of the standard references [GOR 99, JAI 88, MIR 96]. Any needed background from the theory of partially ordered sets may be obtained from [BIR 67, DAV 90, SZA 63]. An interesting mathematical model for cluster analysis was presented in [JAR 71]. We shall not reproduce it here, but do point out that the current discussion has its origins in that text. The basic input to a clustering algorithm is a finite nonempty set P equipped with a finite collection of attributes that the members of P may possess. These attributes can be numerical, nominal or binary. The attributes are then converted to a *dissimilarity coefficient* (DC). This is a mapping $d\colon P\times P \to \Re_0^+$, the non-negative reals, that satisfies $d(a,b) = d(b,a)$, and $d(a,a) = 0$ for all $a, b \in P$. The DC d is an *ultrametric* if it also satisfies $d(a,b) \leq \max\{d(a,c), d(b,c)\}$ for all $a, b, c \in P$.

The T-transform: Let $\Sigma(P)$ denote the set of reflexive symmetric relations on P, ordered by set inclusion. Associated with any DC d, there is a mapping $Td\colon \Re_0^+ \to \Sigma(P)$ defined by

$$Td(h) = \{(a,b) : d(a,b) \leq h\} \text{ for all } h \in \Re_0^+.$$

[1] Preliminary versions of short portions of this talk were given at Ecole Nationale Supérieure des Télécommunications de Bretagne on October 30, 2004, and at DIMACS on March 9, 2005.

It is easy to show and well known that $Td(h)$ is an equivalence relation for all $h \in \Re_0^+$ if and only if d is an ultrametric. Thus ultrametrics yield nested sequences of equivalence relations. For that reason, a cluster algorithm may be viewed as a transformation $d \mapsto C(d)$ of a DC d into an ultrametric $C(d)$.

In [JAN 78], we replaced $\Re_0^+$ with a join semilattice L having a smallest member 0. We defined an L-dissimilarity coefficient to be a mapping $d\colon P \times P \to L$ such that $d(a,b) = d(b,a)$, and $d(a,a) = 0$ for all $a, b \in P$. The DC d is an *ultrametric* if it also satisfies $d(a,b) \leq d(a,c) \vee d(b,c)$ for all $a, b, c \in P$. The T-transform associated with an L-dissimilarity coefficient d is defined as expected by taking it to be the mapping $Td\colon L \to \Sigma(P)$ defined by $Td(h) = \{(a,b) : d(a,b) \leq h\}$ for all $h \in L$. This was the original setting, but it is not quite what is needed.

Single-linkage clustering is one of the standard clustering algorithms. Here is how it operates. If $u = C(d)$, we take $Tu(h) = \gamma \circ Td(h)$, where γ is the transitive closure operator. It was shown in [JAN 78] that if $h \wedge k$ exists in L, then $Tu(h \wedge k)$ must equal $Tu(h) \cap Tu(k)$. This says that $\gamma \circ Td(h \wedge k)$ must equal $\gamma \circ Td(h) \cap \gamma \circ Td(k)$. But γ defined on $\Sigma(P)$ does not have this property. Thus we must either abandon single linkage clustering as a cluster method or modify our model. This fact most certainly limited the algorithms that were available to us in conjunction with the percentile clustering model [JAN 89].

2 The Modified Model

We choose here to change our perspective a bit. First of all, we assume nothing past the fact that the place L in which dissimilarities are measured should be a partially ordered set. Thus we do not assume even that L has a smallest element 0. The idea behind the concept of a dissimilarity coefficient is that at each $h \in L$ there is a reflexive symmetric relation $S(h)$ that identifies the pairs of members of P that are candidates for grouping at level h. We want a dissimilarity coefficient to take $\{a, b\}$ and identify all those $h \in L$ for which $(a,b) \in S(h)$. When $L = \Re_0^+$, and we write $d(a,b) = h$, we really mean that $(a,b) \in S(k)$ if and only if $k \geq h$. It seems reasonable to assume that $S\colon L \to \Sigma(P)$ should have the property that $h \leq k \Longrightarrow S(h) \subseteq S(k)$. Note that we are *not* assuming that the existence of $h \wedge k$ in L implies that $S(h \wedge k) = S(h) \cap S(k)$. When L is a join semilattice or has a largest member, it also seems reasonable to assume the existence of $h \in L$ such that $S(h) = P \times P$.

Order filters of L will play an important role in what follows. Unfortunately, there is no uniform agreement about whether to require that they be nonempty. Here is the convention we shall follow. An *order filter* of L is a subset F of L such that $h \in \mathsf{F}$, $h \leq k \Longrightarrow k \in \mathsf{F}$. If L has a largest member, we require that any order filter be nonempty; otherwise, we allow the empty set to be an order filter. The set $\mathcal{F}(L)$ of order filters of L is ordered by the rule $\mathsf{F} \leq \mathsf{G} \Longleftrightarrow \mathsf{G} \subseteq \mathsf{F}$. This may seem strange but the point is that we want $x \mapsto \mathsf{F}_x$ to be an embedding. Here F_x is the principal filter generated by x, and is defined by $\mathsf{F}_x = \{y \in L : y \geq x\}$. Since $\mathsf{F} \vee \mathsf{G} = \mathsf{F} \cap \mathsf{G}$ and $\mathsf{F} \wedge \mathsf{G} = \mathsf{F} \cup \mathsf{G}$,

it is true that $\mathcal{F}(L)$ is a bounded distributive lattice with smallest element the order filter L, and largest element the empty filter or a filter consisting of the largest member of L.

Definition 1. We now define an *L-general* dissimilarity coefficient to be a mapping $D\colon P\times P \to \mathcal{F}(L)$ that satisfies

(GD1) $D(a,b) = D(b,a)$.

(GD2) $D(a,a) = L$ for all $a \in P$.

Note that we are using a capital letter D to clearly distinguish this type of DC from the usual $d\colon P\times P \to L$. It will sometimes be useful to replace (GD2) with

(GD2′) $D(a,a) = \bigwedge\{D(a,b) : b \in L,\ a \neq b\}$ (providing the needed meets exist), or

(GD2″) Calculate $D(a,a)$ using same formula as for $D(a,b)$ with $b \neq a$.

Of course (GD2′) and (GD2″) also make sense for an ordinary DC.

An ordinary DC $d\colon P\times P \to L$ has an associated general DC $D\colon P\times P \to \mathcal{F}(L)$ defined by $D(a,b) = \mathsf{F}_{d(a,b)}$, For that reason, we can be sloppy about terminology, and just use the notation d versus D to specify whether we are dealing with ordinary or general DCs.

We are in the classical setting now if we view D as taking values in $\mathcal{F}(L)$, and take $TD\colon \mathcal{F}(L) \to \Sigma(P)$ in the usual manner. But we want the map associated with D to go from L into $\Sigma(P)$. We therefore define $SD\colon L \to \Sigma(P)$ by the rule $SD(h) = \{(a,b) : h \in D(a,b)\}$. A moment's reflection should convince the reader that to say that $SD(h)$ is a transitive relation for all h is equivalent to saying that

(GD3) $D(a,b) \leq D(a,c) \vee D(b,c)$ for all $a,b,c \in P$.

The point is that $h \in D(a,c) \cap D(b,c)$ should force $h \in D(a,b)$. In terms of the binary relation $SD(h)$, we are saying that $(a,b),(b,c) \in SD(h)$ should imply that $(a,c) \in SD(h)$. This leads us to call a general DC an *ultrametric* if it satisfies (GD1), some variant of (GD2), and (GD3). A cluster method may now be taken as a mapping $D \mapsto C(D)$, where D is a general DC and $C(D)$ an ultrametric. Please recall that there is nothing that precludes the original DC D from having the property that each $D(x,y)$ is a principal filter. Parallel to the fact that $d(a,b) \leq h \iff (a,b) \in Td(h)$ is the fact that

$$h \in D(a,b) \iff (a,b) \in SD(h).$$

We mention the fact that it may be convenient to take as our input the values of $SD(h)$ for $h \in L$, rather than the associated general DC D. **Warning:** If we use (GD2′) or (GD2″) in place of (GD2), then $TD(h)$ as well as $SD(h)$ need not produce a reflexive relation. Thus we must replace $\Sigma(P)$ with the symmetric relations on P.

The design of useful clustering algorithms when faced with dissimilarities measured in an arbitrary poset L is of course a critical issue of concern, and will be an active aspect of future research efforts.

3 Boolean dissimilarities

We begin by mentioning that D is called a Boolean dissimilarity whenever L is a Boolean algebra. We now consider a rather special situation. Suppose the set P is equipped with k binary attributes. We want to use these attributes to define a DC taking values in $L = 2^k$. Let $a, b \in P$ with $a \neq b$. If a has attributes $(a_1, a_2, \ldots, a_k)$ and b has attributes $(b_1, b_2, \ldots, b_k)$, we want to define a dissimilarity $d(a, b) = (x_1, x_2, \ldots, x_k) \in 2^k$, where x_i is computed entirely from a_i and b_i. How shall we do this? Since a DC is supposed to be a measure of how dissimilar a and b are, it is clear that if $a_i \neq b_i$, we want $x_i = 1$. There are now only two remaining cases to consider: $a_i = b_i = 0$ and $a_1 = b_i = 1$. We end up with three distinct DCs of interest:

$d_1(a, b) = (x_1, x_2, \ldots, x_k)$ where $x_i = 1$ if $a_i \neq b_i$ and 0 otherwise.

$d_2(a, b) = (y_1, y_2, \ldots, y_k)$ where $y_i = 0$ if $a_i = b_i$ and 1 otherwise.

$d_3(a, b) = (z_1, z_2, \ldots, z_k)$ where $z_i = 0$ if $a_i = b_i = 0$ and 1 otherwise.

Note that we need not further consider d_3, as d_2 and d_3 are symmetric with respect to negation of attributes. Note further that the definitions of the output DC may vary from coordinate to coordinate of 2^k.

Theorem 2. *For $i = 1,\ 2,$ or 3 and $h = (h_1, h_2, \ldots, h_k)$, the symmetric relation $SD(h)$ is transitive.*

The two DCs d_1, d_2 have their analogues among the standard dissimilarities for binary data. In the literature, they are called the "simple matching coefficient" and "Russell and Rao coefficient". The simple matching coefficient for objects x and y just counts up the number of attributes in which x, y differ and divides by the total number of attributes. This is related to what we called d_1. Note that if we compute $d_1(a, b)$ and just add up the resulting vector and divide by k, we have the result of the simple matching coefficient. The Russell and Rao coefficient counts up the number of attributes in which x, y differ, and the number where they are both 0 and divides by the total number of attributes. This was our d_2 in the Boolean setting. There are of course other rather different ways of measuring a dissimilarity between binary attributes. Note that there was no need for a general DC here because of Theorem 2. No cluster method was needed to produce an ultrametric.

4 Formal concept analysis

It turns out that the notion of a Boolean dissimilarity fits nicely into a general theory designed to help gain insight into the structure of complicated data sets. This theory is called Formal Concept Analysis. An early reference to the underlying ideas can be found in [BIR 67] in the discussion of Galois connections. An elegant, formal treatment occurs in [GAN 99]. A concise lattice theoretic introduction may be found in [DAV 90], Chapter 11. It turns out that with an appropriate definition of $d_2(a, a)$, formal concept analysis may be viewed as a special case of d_2.

In no way does this obviate the usefulness of formal concept analysis as a tool for determining the structure of large, complex data sets, but it does put cluster analysis and formal concept analysis into a common framework. The DC d_1 suggests an alternate formulation of formal concept analysis based on the individual attributes forming bipartitions of the original data set.

References

[BIR 67] G. Birkhoff, *Lattice Theory, third ed.*, American Mathematical Society, Providence, 1967.

[BLY 72] T. S. Blyth and M. F. Janowitz, *Residuation Theory*, Pergamon Press, Oxford, 1972.

[DAV 90] B. A. Davey and H. A. Priestley, *Introduction to Lattices and Order*, Cambridge Univeristy Press, Cambridge, 1990.

[GAN 99] B. Ganter and R. Wille, *Formal Concept Analysis. Mathematical Foundations*, Springer-Verlag, Berlin, 1999.

[GOR 99] A. D. Gordon, *Classification, 2nd ed.*, Chapman & Hall, London, 1999.

[JAI 88] Anil K. Jain and Richard C. Dubes, *Algorithms for Clustering Data*, Prentice-Hall, Englewood Cliffsm 1988.

[JAN 78] M. F. Janowitz, *An order theoretic model for cluster analysis*, SIAM J. Applied Math. **34** (1978), 55-72.

[JAN 89] M. F. Janowitz and B. Schweizer *Ordinal and percentile clustering*, Mathematical Social Sciences **18** (1989), 135-186.

[JAR 71] N. Jardine and R. Sibson, *Mathematical Taxonomy,* Wiley, New York, 1971.

[MIR 96] B. Mirkin, *Mathematical Classification and Clustering*, Kluwer, Dodrecht, 1996.

[SCH 83] B. Schweizer and A. Sklar, *Probalistic Metric Spaces*, North-Holland, Amsterdam, 1983.

[SNE 72] P. H. A. Sneath and R. R. Sokal, *Numerical Taxonomy*, W. H. Freeman, San Francisco, 1973.

[SZA 63] Gabor Szasz, *Introduction to Lattice Theory*, Third Ed., Academic Press, Budapest, 1963.

Models and methods of two-mode cluster analysis

Sabine Krolak-Schwerdt

Department of Psychology,
Saarland University,
Postbox 151150
D-66041 Saarbrücken, Germany

ABSTRACT.
In this paper methods to cluster analyze two-mode data are discussed which assume that both objects and attributes contribute to the uncovering of meaningful clusters. Two-mode methods are reviewed and criteria are proposed which aim at a comparison and evaluation of the reviewed methods. An application to social cognition data illustrates how two-mode cluster analysis is superior to the classical approaches within this domain in understanding and interpreting clustering solutions.

Keywords: Classification, two-mode data, clustering methods

1 Introduction

This paper is concerned with the cluster analysis of two-mode data which consist of two sets of entities (e.g., objects and attributes). Methods to cluster analyze such data assume that both objects and attributes contribute to the uncovering of meaningful patterns of clusters.

The aim of two-mode techniques is to classify the objects and simultaneously to classify the attributes. Two-mode cluster analysis is thus a valuable device for any researcher who is interested in a grouping of objects as well as attributes. An example from applied research is the classification of mental diseases by sets of symptoms. Here, the analysis aims at a classification of mental diseases, a classification of symptoms and an indication of which groups of symptoms characterize which groups of diseases.

In this paper, a review of two-mode clustering methods is given. Subsequently, an application to social cognition data will be used to demonstrate the gain of information provided by two-mode approaches as compared to the usual one-mode methods. Finally, some criteria are proposed which may be helpful in comparing and evaluating the reviewed methods.

2 Review of two-mode clustering methods

Research on two-mode clustering frequently subdivides corresponding methods into three categories. The first category consists of methods which fit the mathematical structure of a two-mode ADCLUS generalization to the data by use of a global loss function. DeSarbo [DES 82] developed the first method within this category. Given a rectangular nonsymmetric similarity matrix S, the method aims at finding a best fitting matrix $\hat{S}$,

$$\hat{S} = P\ V\ Q' + C,$$

where the matrix P designates membership of the objects in k clusters, Q designates membership of the attributes in k clusters, V is a matrix of weights attached to the clusters and C is a matrix where each element is an additive constant c.

Based upon the model equation given above, ADCLUS generalizations have a sound psychological theory as to how the similarity between stimuli such as an object and an attribute is judged. Further characteristics of these methods are that they allow for overlapping as well as nonoverlapping clusters and that the input data are required to be interval or ratio scaled.

The second category contains methods which are fitting additive or ultrametric tree structures to two-mode data. Starting with a two-mode data matrix, the algorithms construct a sort of "grand matrix" in the first step which contains three types of similarities: between objects and attributes, between object pairs and between attribute pairs. Using the grand matrix, classical one-mode algorithms are applied in the second step. Methods within the second category are thus hierarchical methods representing the fusion process by a dendrogram. The clustering solution is always nonoverlapping. An example method within this category is the approach proposed by De Soete et al. [SOE 84].

These methods use heuristic criteria for the construction of clusters and no global loss function is defined. Advantages of these methods are that the input data are not restricted to dissimilarities. Second, using clustering algorithms with well-known properties such as the average-linkage procedure omits the use of cumbersome estimation procedures as in the first category and there is no a-priori knowledge of the number of clusters required to perform the analysis.

Methods belonging to the third category may be termed "reordering approaches". They may be characterized in the following way:
(1) Two-mode profile data are assumed as input and, depending on the specific method, the data values are interpreted as either categorical or they are restricted to a binary format. (2) The algorithms operate directly on the input data and, consequently, no intermediate steps such as the construction of a grand matrix is necessary. (3) The aim is to make clusters visible as blocks within the data matrix with objects showing identical values across the attributes after an appropriate reordering or permutation of the rows and columns of the data matrix. (4) Most of the methods within this category do neither involve a global loss function to be optimized nor a psychological foundation as to how clusters are formed, but instead use some heuristic rule. An example method within this category is the "bond energy algorithm" first developed by McCormick, Schweitzer and White [MCO 72] and improved by Arabie, Schleutermann and Hubert [ARA 88] as to the efficiency of the algorithm.

3 Application

In order to show how two-mode cluster analysis is superior to the classical one-mode approaches, experimental data on the cognitive representation of social stereotypes and their attributes were analyzed. The data sets consist of subjects' assignments of attributes from the domains of physical appearance, intelligence, attitudes and dominance to gender stereotypes, where the assignments were obtained under different judgment conditions. In the following, results from the analysis of one of the data sets will be presented as an example.

In order to derive a one-mode representation, Euclidean distances between stereotypes were computed from the data and subjected to the average-linkage procedure. The result was a grouping of the stereotypes into three clusters: (1) "intellectual", "soft type", (2) "typical woman", "housewife" and (3) "feminist", "senior citizen", "manager", "career woman".

In order to inspect which attributes have contributed to the clusters, a frequently used procedure within the one-mode framework is to compute the centroids of each cluster and subsequently to inspect the centroids to identify corresponding attributes. As compared to two-mode clustering, the disadvantage associated with this procedure is that it is neither possible to simultaneously obtain a clustering of the attributes nor to obtain any information as to how object clusters may be attached to specific attribute clusters. Thus, combining a one-mode approach by looking at cluster centroids does not substitute a two-mode procedure.

In the second step, the data were reanalyzed by two-mode approaches from each of the categories outlined above. All solutions provided by the two-mode approaches encompassed the stereotype classification suggested by the one-mode average-linkage procedure. However, the two-mode

representations made subjects' beliefs visible about how certain subsets of attributes tend to go together in specific types of persons. In the following, the entities of the attribute clusters associated with each object cluster are given in italics:

(1) "intellectual", "soft type", *"slovenly", "unfashionable"", "idealistic", "shows feelings"*
(2) "typical woman", "housewife", *"selfless", "reserved", "passive", "likable"*
(3) "feminist", "senior citizen", "manager", "career woman", *"strong", "dominant", "ambitious", "intelligent"*.

As an example, cluster 1 consisted of specific male stereotypes (e.g., "intellectual", "soft type") and showed that their specific characteristics were in the domains of physical appearance and attitudes (e.g., "slovenly", "idealistic" etc.). In sum, in this application the two-mode approaches yielded clustering solutions which were very similar as to the allocation of both stereotypes and attributes and which offer more information about the cognitive representation of social stereotypes than the one-mode solution.

4 Criteria for comparison and evaluation

In applied research, criteria which may serve as an aid in comparing and evaluating two-mode methods are the following: 1) interpretation of input data and their scale values, 2) the question, if the method involves major computational problems, 3) the derivation of a measure of goodness-of-fit, 4) the specification of different types of clustering and between-cluster relationships and 5) finally the derivation of a three-mode version of the method.

The first aspect involving interpretation of input data yields a separation of the group of ADCLUS generalizations and most ultrametric tree fitting approaches from reordering methods in that the former two may be applied to the special case of two-mode similarity data. Thus, the applicability of these approaches is restricted. The goal of classifying all types of `genuine' two-mode data is achieved only by reordering approaches.

The scale values criterion characterizes the ADCLUS generalizations and some ultrametric tree fitting approaches as having the drawback of requiring metric data while every other approach besides these is applicable to the broad range of non-metric input data.

Computational problems concern the major objection against the group of reordering approaches. The most serious drawback is that all approaches use heuristics and the most of them yield nonunique solutions. On the other hand, these methods are quite popular because they offer the advantage of being easy to handle for users: They operate directly on the input data and thus avoid additional transformations of the data into appropriate similarity values. Furthermore, the clustering solution may be interpreted directly on the input data in that clusters are visualized as object-attribute-blocks.

As to the issue of computational problems, the ultrametric tree fitting approaches appear promising for solving the two—mode classification problem at first sight. They use clustering algorithms with well-known properties such as average-linkage. Furthermore, there is no a-priori knowledge on the number of clusters required to perform the analysis. However, as the clustering solution is always nonoverlapping, clusters may be obtained which consist of objects (or attributes) only. Thus, the applicability of these methods may be restricted to specific research domains. That is, any research domain in which reasoning from a theory or hypothesis asks for overlapping clusters precludes the use of ultrametric tree fitting approaches. This is the case in any application where attributes may characterize more than one cluster and/or objects may belong to more than one cluster.

A very different view on two-mode approaches concerns the question, if their theoretical underpinnings allow for an extension to three-mode data which consist of objects, variables and conditions. There are three approaches which offer a three-mode version: the De Soete et al. [SOE 84] approach and two methods from the reordering category. The other methods have not been extended in this direction as yet.

Most of the reviewed approaches do not offer a measure of goodness-of-fit of the two-mode solution. There are some exceptions within the ADCLUS generalizations. These methods involve an estimation of the quality of the clustering solution in terms of variance accounted for by th representation or error variance not explained.

The final criteria, which involve the specification of different types of clustering and between--cluster relationships, qualify the group of ADCLUS extensions as the more interesting approaches. ADCLUS extensions are the only methods which allow for overlapping as well as nonoverlapping representations while the other methods construct a solution of either type. Furthermore, it is a distinct feature of the ADCLUS extensions to specify the relations between different clusters in terms of inter-cluster similarity or association. This is accomplished by use of the matrix V, where the off-diagonal weights indicate the similarity of corresponding row and column clusters.

5 Conclusions

The criteria and results from comparing and evaluating two-mode clustering procedures imply two sets of reccomendations. One offers recommendations for applied analyses in selecting a particular method. The other deals with topics of two-mode clustering that would benefit from continued research in the methodology. Both sets of recommendations will be discussed.

6 Bibliography

[ARA 88] ARABIE P., SCHLEUTERMANN S., DAWS J., HUBERT L. (1988), Marketing Applications of Sequencing and Partitioning of Nonsymmetric and/or Two-mode Matrices. In W. Gaul & M. Schader (Eds.), *Data, expert knowledge and decisions* (pp. 215-224). Berlin: Springer.

[DES 82] DESARBO, W.S. (1982), Gennclus: New Models for General Nonhierarchical Clustering Analysis. *Psychometrika*, 47, 449-475.

[SOE 84] DE SOETE G., DeSARBO W.S., FURNAS G.W. and CARROLL J.D. (1984), The Representation of Rectangular Proximity Matrices. In E. Degreef & J. Van Buggenhaut (Eds.), *Trends in Mathematical Psychology*. North-Holland, Amsterdam, 377-392.

[MCO 72] McCORMICK W.T., SCHWEITZER P.J., WHITE T.W. (1972), Problem Decomposition and Data Reorganization by a Clustering Technique. *Operations Research*, 20, 993-1009.

Implications, emboîtements et ajustement de classifications

Bruno Leclerc

École des Hautes Études en Sciences Sociales
Centre d'Analyse et de Mathématique Sociales
54 boulevard Raspail
75270 Paris cedex 06, france
leclerc@ehess.fr

RÉSUMÉ. Soit $\mathcal{C}$ un ensemble de classes d'éléments d'un ensemble S On lui associe (classiquement) une relation I d'implication et (moins habituellement) un ordre Œ d'emboîtement. Lees ensembles de classes $\mathcal{F}$ contenant S et stables par intersection (systèmes de fermeture) jouent alors un rôle particulier puisque l'on peut se ramener à ce cas. On note l'importance des classes $\wedge$–irréductibles de $\mathcal{F}$, et des emboîtements (M, M^+) associés, où M est $\wedge$–irréductible et M^+ est l'élément minimum de $\mathcal{F}$ contenant strictement M. L'exposé porte sur les possibilités d'inférer, à partir de ces éléments, un système de fermeture à partir d'une relation d'emboîtement partielle R.

MOTS-CLÉS : Classe, Système de fermeture, Famille de Moore, Implication, Emboîtement, Treillis Irréductible.

1 Introduction

Ce résumé présente les éléments de base pour l'exposé, qui portera sur les possibilités, formelles et algorithmiques, de l'ajustement d'un système de classes à une relation d'implication donnée. La prise en compte de ce type de données revient, d'une part, à considérer la cohérence des éléments d'une classe par rapport à ceux de l'extérieur (plutôt qu'intrinsèque) et, d'autre part, permet de s'intéresser à tous les ensembles d'objets considérés et non aux seules classes reconnues.

Soit S un ensemble fini, et $\mathcal{C} \subseteq \mathcal{P}(S)$ un ensemble de classes d'éléments de S. On définit à la section 2 la relation d'implication I correspondant à $\mathcal{C}$, ainsi que l'ordre d'emboîtement $Œ$ étroitement apparenté à I ; on travaillera en pratique sur cet ordre. On constate en section 3 qu'il revient au même, dans ce cadre, de considérer le système $\mathcal{C}$ ou le plus petit système de fermeture $\mathcal{F} = \mathcal{F}(\mathcal{C})$ contenant $\mathcal{C}$. On caractérise des couples de parties de S déterminants pour de tels systèmes, puis on présente à la section 4 quelques résultatrs sur le problème d'ajustement mentionné ci-dessus.

2 Classes, implications et emboîtements

Considérons un ensemble fini S et un ensemble $\mathcal{C}$ de classes d'éléments de S. Une classe $C \in \mathcal{C}$ est une partie de S dont les éléments ont été regroupés en vertu de propriétés communes [CP 13] ou d'un certain type de proximité entre eux. On associe à cet ensemble $\mathcal{C}$ deux relations binaires I et $Œ$ sur l'ensemble $\mathcal{P}(S)$ des parties de S.

- La relation d'*implication I* sur S associée à $\mathcal{C}$ correspond à l'idée que la partie B est systématiquement associée à la partie A dans $\mathcal{C}$, en ce sens que *A implique B* (ce qui est noté $A \Rightarrow B$, ou $(A, B) \in I$, ou $A\ I\ B$) si toute classe de $\mathcal{C}$ contenant A contient aussi B.
- La relation d'*emboîtement* $\mathrel{Œ}$ sur S associée à $\mathcal{C}$ correspond à l'idée que la partie B est plus générale que la partie A par rapport à $\mathcal{C}$, en ce sens que *A est emboitée dans B* (ce qui est noté $(A, B) \in Œ$, ou $A\ Œ\ B$) si $A \subset B$ (inclusion stricte) et s'il existe une classe de $\mathcal{C}$ contenant A et ne contenant pas B.

Les relations I et $Œ$ se déduisent alors l'une de l'autre ; par exemple, on a $Œ = \{(A, B) \in \mathcal{P}(S)^2 : A \subset B$ et $(A, B) \notin I\}$. La relation I est un *système implicatif complet* (SIC) sur S, c'est-à-dire qu'elle vérifie :
(I1) pour tous $A, B \subseteq S$, $B \subseteq A \Rightarrow A\ I\ B$;
(I2) pour tous $A, B, C \subseteq S$, $A\ I\ B$ et $B\ I\ C \Rightarrow A\ I\ B$;
(I3) pour tous $A, B, C, D \subseteq S$, $A\ I\ B$ et $C\ I\ D \Rightarrow A \cup C\ I\ B \cup D$.

De son côté, la relation $Œ$ vérifie les propriétés :
(E1) pour tous $A, B \subseteq S$, $A\ Œ\ B \Rightarrow A \subset B$;
(E2) pour tous $A, B, C \subseteq S$, $A \subset B \subset C \Rightarrow [A\ Œ\ C \Leftrightarrow A\ Œ\ B$ ou $B\ Œ\ C]$;
(E3) pour tous $A, B \subseteq S$, $A\ Œ\ A \cup B \Rightarrow A \cup B\ Œ\ B$.

Une relation d'implication est réflexive (selon (I1), on a $A\ I\ A$ pour tout $A \subseteq S$), tandis que (I2) signifie qu'elle est transitive. Il s'agit donc d'un préordre sur $\mathcal{P}(S)$. Une relation d'emboîtement est, d'après (E1), irréflexive ($(A, A) \notin Œ$, pour tout $A \subseteq S$) et asymétrique ($A\ Œ\ B \Rightarrow (B, A) \notin Œ$, pour tous $A, B \subseteq S$), et (E2) entraîne qu'elle est transitive. Les emboîtements sont donc des ordres irréflexifs sur $\mathcal{P}(S)$.

Si $\mathbf{I}$ est l'ensemble des SICs sur S, on observe que $I, I' \in \mathbf{I}$ entraîne $I \cap I' \in \mathbf{I}$, et que la relation pleine $\mathcal{P}(S)^2$ est le SIC correspondant à $\mathcal{C} = \varnothing$. Si $\mathbf{O}$ est l'ensemble des ordres d'emboîtement sur S, on a $O, O' \in \mathbf{O}$ entraîne $O \cap O' \in \mathbf{O}$, la relation vide correspondant à $\mathcal{C} = \varnothing$.

Les préordres d'implication ont été intensivement considérés dans la littérature (cf. Caspard et Monjardet [CM 03] pour des références commentées). Ainsi, dans le cas des bases de données relationnelles, $A\ I\ B$ signifie que toute requête satisfaite par les éléments de A l'est aussi par ceux de B ; ce type de considérations se retrouve dans de nombreux autres contextes. Les ordres d'emboîtement sont d'abord apparus dans le cas particulier où $\mathcal{C}$ est une hiérarchie sur S et à propos du consensus de classifications [ADA 86]. Leur récente extension à des ensembles de classes plus variés est présentée dans [DL 04a].

3 Systèmes et opérateurs de fermeture

Un *système de fermeture* (ou *famille de Moore*) sur S est une famille $\mathcal{F}$ de parties de S vérifiant (i) $S \in \mathcal{F}$, et (ii) pour tous $F, F' \in \mathcal{F}$, $F \cap F' \in \mathcal{F}$. On rencontre fréquemment de tels systèmes ; ainsi, nous venons de voir que l'ensemble $\mathbf{I}$ des SICs sur S est un système de fermeture sur $\mathcal{P}(S)^2$. L'ensemble de tous les systèmes de fermeture sur S, noté $\mathbf{F}$, est lui-même un système de fermeture sur $\mathcal{P}(S)$.
Un ensemble $\mathcal{C}$ de classes de S se complète de façon minimale en un système de fermeture $\mathcal{F}(\mathcal{C})$ le contenant, en posant $\mathcal{F}(\mathcal{C}) = \{S\} \cup \{\cap \mathcal{B} : \mathcal{B} \subseteq \mathcal{C}\}$ (on fait toutes les intersections de classes prises dans $\mathcal{C}$, et on ajoute S si nécessaire).
Inversement, pour un système de fermeture $\mathcal{F}$ sur S donné, il y a une plus petite famille $\mathcal{M}_\mathcal{F}$ de parties de S telle que $\mathcal{F} = \mathcal{F}(\mathcal{M}_\mathcal{F})$. Un élément M de $\mathcal{F}$ est dit $\cap$*-irréductible* s'il ne s'obtient pas comme intersection d'autres éléments de $\mathcal{F}$. De façon équivalente, M est $\cap$-irréductible si l'ensemble $\{F \in \mathcal{F} : M \subset F\}$ a un plus petit élément, noté M^+. On a alors $\mathcal{F} = \mathcal{F}(\mathcal{C}) \Leftrightarrow \mathcal{M}_\mathcal{F} \subseteq \mathcal{C} \subseteq \mathcal{F}$, $\mathcal{M}_\mathcal{F}$ étant l'ensemble des

$\cap$–irréductibles de $\mathcal{F}$. Parmi les éléments de $\mathcal{M}_\mathcal{F}$, les *coatomes* de $\mathcal{F}$ sont ceux pour lesquels $M^+ = S$. Les éléments de $\mathcal{F}$ s'obtiennent à partir de ceux de $\mathcal{M}_\mathcal{F}$ par $F = \cap \{M \in \mathcal{M}_\mathcal{F} : F \subseteq M\}$, pour tout $F \in \mathcal{F}$.

Une remarque importante est que le SIC I et l'emboîtement $\mathcal{E}$ associés à un ensemble de classes $\mathcal{C}$ quelconque ne changent pas lorsqu'on remplace $\mathcal{C}$ par le système de fermeture $\mathcal{F} = \mathcal{F}(\mathcal{C})$. De plus, il y a une correspondance biunivoque, établie par Armstrong [ARM 74], entre SIC et systèmes de fermeture, et ce résultat se transmet aux ordres d'emboîtements [DL 04a]. Si I' et $\mathcal{E}'$ sont le SIC et l'emboîtement associés à un autre système de fermeture $\mathcal{F}'$ sur S, on a de plus, pour les inclusions :

$$\mathcal{F} \subseteq \mathcal{F}' \Leftrightarrow I' \subseteq I \Leftrightarrow \mathcal{E} \subseteq \mathcal{E}'.$$

Les couples de la forme (M, M^+) devant avoir un rôle important dans la suite, nous commençons à les étudier. Il est d'abord clair que, si $\mathcal{E}$ est la relation d'emboîtement associée à $\mathcal{F}$, on a, pour tout $M \in \mathcal{M}_\mathcal{F}$, $M \mathcal{E} M^+$ (puisque $M \subset M^+$ et M est une classe de $\mathcal{F}$ ne contenant pas M^+). Le résultat suivant, où la condition (iii) est issue de la théorie des ensembles ordonnés (elle provient des "relations-flèches"), permet de reconnaître les couples (M, M^+) parmi ceux de $\mathcal{E}$ dès lors que l'on connaît tous les $\cap$-irréductibles contenant strictement M.

Proposition. Soient un système de fermeture $\mathcal{F}$ sur S et l'ordre d'emboîtement $\mathcal{E}$ associé à $\mathcal{F}$, et soit $(A, B) \in \mathcal{E}$. Alors, $A \in \mathcal{M}_\mathcal{F}$, et $B = A^+$ si et seulement si les trois conditions suivantes sont vérifiées :

(i) $B = \cap \{M \in \mathcal{M}_\mathcal{F} : B \subseteq M\}$,

(ii) A est maximal pour l'inclusion parmi les $A \subseteq S$ tels que $A \mathcal{E} B$,

(iii) Il n'existe pas d'élément $\cap$-irréductible M de $\mathcal{F}$ tel que $A \subseteq M$, $B \subseteq M^+$, et $B \not\subseteq M$.

4 Un problème d'ajustement

Nous nous plaçons maintenant dans le cas où on a la donnée d'une relation R sur les parties de S telle que $A\ R\ B$ signifie que l'on a $A \subseteq B$, avec de plus des raisons de douter de l'implication de B par A (par exemple, un caractère partagé par les éléments de A ne se retrouve pas dans tous ceux de B). Nous cherchons alors à ajuster un ordre d'emboîtement $\mathcal{E}$ sur S, tel que défini en section 2 ci-dessus, à la relation R. Si nous y parvenons, nous avons une classification sur S, dont les classes sont les éléments du système de fermeture $\mathcal{F}$ associé à $\mathcal{E}$.

Nous avons observé que l'ensemble $\mathcal{O}$ des emboîtements sur S contient la relation vide et est stable par union. On en déduit l'existence de l'ordre d'emboîtement $\mathcal{E}_{sd}(R) = \cup \{\mathcal{E} \in \mathcal{O} : \mathcal{E} \subseteq R\}$.

Proposition. L'ordre d'emboîtement $\mathcal{E}_{sd}(R)$ est sous-dominant de R (i.e. maximum pour l'inclusion parmi les ordres d'emboîtement inclus dans R).

Cette première solution est signalée dans [DL 04b], tout au moins dans le cas particulier de l'agrégation d'un profil (k-uple) $\mathcal{F}^* = (\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_k)$ de systèmes de fermeture sur S en un système de fermeture consensus $\mathcal{F}$. Avec les emboîtements $\mathcal{E}_i$ associés aux systèmes $\mathcal{F}_i$, $i = 1,\ldots, k$, et, pour un certain seuil q $(1 \le q \le k)$, on considère la relation $R = \{(A, B) \in \mathcal{P}(S)^2 : A \mathcal{E}_i B$ pour au moins q indices $i\}$. Choisir alors l'ordre d'emboîtement consensus $\mathcal{E}_{sd}(R)$ revient à prendre le système de fermeture consensus $\mathcal{F} = \mathcal{F}(\mathcal{C}_q)$, où $\mathcal{C}_q$ est l'ensemble des classes présentes dans au moins q des $\mathcal{F}_i$. Comme cette méthode risque de donner peu de classes dans le système $\mathcal{F}$, on s'intéresse aussi aux ordres d'emboîtement *contenant* la relation R.
En s'inspirant des propriétés définies par Adams [ADA 86] pour caractériser sa méthode de consensus de

hiérarchies (cf. aussi [DM 03]), nous demandons alors que le système de fermeture $\mathcal{F}$ et son ordre d'emboîtement $\mathcal{E}$ vérifient les deux propriétés suivantes :

(AR1) $R \subseteq \mathcal{E}$ (préservation de $\subseteq$),

(AR2) pour tout $M \in \mathcal{M}_{\mathcal{F}}$, $M\ R\ M^{+}$ (emboîtements certifiés).

La propriété (AR2) peut être vue comme une réciproque très partielle (réduite aux seuls couples de la forme (M, M^{+}), $M \in \mathcal{M}_{\mathcal{F}}$) de (A$R$1). On obtient le résultat suivant, qui permet de montrer en corollaire que, s'il existe, un ordre d'emboîtement vérifiant (AR1) et (AR2) est "supérieur maximal" par rapport à R.

Théorème. *Il y a au plus un système de fermeture $\mathcal{F}$ sur S qui, avec son ordre d'emboîtement $\mathcal{E}$, vérifie simultanément les conditions* (AR1) *et* (AR2).

Corollaire. *Soit R une relation binaire sur $\mathcal{P}(S)$ et $\mathcal{F}$ un système de fermeture sur S qui, avec son ordre d'emboîtement $\mathcal{E}$,, vérifie les conditions* (AR1) *et* (AR2)*. Alors, pour tout ordre d'emboîtement $\mathcal{E}'$, les inclusions $R \subseteq \mathcal{E}' \subseteq \mathcal{E}$ impliquent $\mathcal{E}' = \mathcal{E}$.*

La proposition donnée dans la section 3 permet de déterminer formellement, pour une relation R donnée, une procédure "descendante" pour construire un système de fermeture satisfaisant (AR1) et (AR2) ou pour conclure à l'inexistence d'un tel système. Nous indiquons simplement ici que la première étape consiste à déterminer les coatomes de $\mathcal{F}$ comme étant les parties A de S vérifiant $A\ \mathcal{E}\ S$ et A est maximale avec cette propriété.

On rend ainsi compte des cas où l'existence d'un ordre d'emboîtement $\mathcal{E}$ vérifiant (AR1) et (AR2) est attestée ([ADA 86] pour le consensus de hiérarchies avec $q = k$, et [SS 00] pour la recherche d'un "super-arbre" englobant des hiérarchies partielles). On cherchera dans l'exposé à généraliser ces cas "arborescents" et à explorer les questions algorithmiques soulevées.

5 Bibliographie

[ADA 86] ADAMS III E.N., " N-trees as nestings: complexity, similarity and consensus ", *Journal of Classification*, vol. 3, 1986, p. 299–317.

[ARM 74] ARMSTRONG W.W., " Dependency structures of data base relationships ", *Information Processing*, vol. 74, 1974, p. 580–583.

[CM 03] CASPARD N., MONJARDET B., " The lattices of Moore families and closure operators on a finite set: a survey ", *Discrete Applied Math.* , vol. 127, 2003, p. 241–269.

[CP 13] CAUMERY M., PINCHON J., *L'enfance de Bécassine*, Henri Gauthier, Paris, 1913.

[DM03] DAY W.H.E., McMORRIS F.R., *Axiomatic Consensus Theory in Group Choice and Biomathematics*, SIAM, Philadelphia, 2003.

[DL 04a] DOMENACH F., LECLERC B., " Closure Systems, Implicational Systems, Overhanging Relations and the case of Hierarchical Classification ", *Mathematical Social Sciences*, vol. 47, 2004, p. 349-366.

[DL 04b] DOMENACH F., LECLERC B., " Consensus of classification systems, with Adams' results revisited ". In D. Banks, L. House, F.R. McMorris, P. Arabie, and W. Gaul, editors, *Classification, Clustering and Data Mining Applications*, Springer, Berlin, pages 417-428, 2004.

[SS 00] SEMPLE C., STEEL M.A., " A supertree method for rooted trees ", *Discrete Applied Math.*, vol. 105, 2000, p. 147-158.

Clustering including dimensionality reduction: least-squares and maximum-likelihood approaches

Maurizio Vichi

Department of Statistics,
Probability and Applied Statistics
University "La Sapienza" of Rome
P.le Aldo Moro, 5
I-00185 Rome, Italy

ABSTRACT. In this paper new methodologies for clustering and dimensionality reduction of large data sets are illustrated using both a least-squares and maximum likelihood approach. The methodologies are described by both real applications and Monte Carlo simulations.

KEYWORDS: Clustering of objects and variables, dimensionality reduction, least-square partitioning, maximum likelihood clustering, mixture models.

1 Introduction

The analysis of proximity relationships within a set of objects can be obtained by identifying disjoint classes of objects which are perceived as similar to one another within each class. Such partitions can be obtained from the applications of cluster analysis methodologies. Nevertheless, cluster analysis is frequently used to partition variables instead of objects or both objects and variables. For example, marketers are interested to know how the market can be segmented into homogeneous classes of consumers according to their preference on products; at the same time, marketers may wish to know how products are clustered according to preferences of customers. This case will be referred to as *two-mode partitioning*. The basic idea is to identify *blocks*, i.e., sub-matrices of the observed data matrix, where objects and variables forming each block specify an *object cluster* and a *variable cluster*.

Of course, in applying two-mode partitioning, variables expressed on the same scale are required, so that entries are comparable among both rows and columns. If this is not the case, data need to be rescaled by an appropriate standardization method. The interested reader can find a very complete structured overview of two-mode clustering methods in Van Mechelen, Bock and De Boeck (2004).

In this paper we show the performance of some new methodologies for two mode partitioning of two way data recently proposed. In particular, the "double *k*-means" (Vichi, 2000; Rocci, Vichi 2004) for two-way data is discussed and compared with procedures that can be obtained by applying ordinary clustering techniques in repeated steps. The performance of double *k*-means has been tested by both a simulation study and an application to gene microarray data. Recently, Vichi and Martella, (2005) have studied the maximum likelihood clustering estimation of the double *k*-means parameters.

Clustering of objects and variables according to *double k-means* is particularly valuable when variables are not so discernible from objects as in the case above described when customers and products are considered. In this situation centroids for both objects and variables (e.g., mean profiles of customers and mean profiles of products) can be used to synthesize the observed data matrix. However, for a usual multivariate data matrix a reduction of the objects is generally given by means of centroids from partitioning methodology, while a reduction of the variables is achieved by a factorial methodology as PCA, hence by means of linear combinations that give different weights to the original variables. However, PCA, but also other factorial techniques, have often the drawback that different factors are characterized by the same original variables, so that the interpretation of these factors becomes a relevant and complex problem. In this situation it would be useful to partition objects into clusters summarized by centroids, but also to partition

variables into clusters of correlated variables, summarized by linear combinations of maximum variance as it is obtained in clustering and disjoint principal component analysis (CDPCA) (Vichi and Saporta, 2004). This methodology can be seen as a generalization of the double *k*-means.

2 The clustering and dimensionality reduction model

The double *k*-means model (Vichi, 2000) is formally specified as follows

$$\mathbf{X} = \mathbf{U}\,\overline{\mathbf{Y}}\,\mathbf{V}' + \mathbf{E}, \tag{1}$$

where $\mathbf{X}$ is a $(I \times J)$ data matrix, while matrix $\mathbf{E}$ is the error component matrix. Matrix $\mathbf{U}=[u_{ij}]$ is a $(I \times P)$ membership matrix, assuming values $\{0, 1\}$, specifying for each object i its membership to a class of the partition of objects in P classes. Matrix $\mathbf{V}=[v_{jq}]$ is a $(J \times Q)$ membership matrix, assuming values $\{0, 1\}$, specifying for each variable j its membership to a class of the partition of variables in Q classes. Matrix $\overline{\mathbf{Y}}=[\bar{y}_{pq}]$ is the $(P \times Q)$ *centroid* matrix where $\bar{y}_{pq}$ denotes the mean of values corresponding to object and variable clusters p and q, respectively. The first term in model (1) pertains to the portion of information of $\mathbf{X}$ that can be explained by the simultaneous classification of objects and variables.

Of course, matrix $\mathbf{X}$ is supposed to be column standardized if the variables are not commensurate. In the papers by Vichi, 2000 and Rocci and Vichi, 2004 fast alternating least-square algorithms are proposed in the case the model is estimated with the least-squares approach, while recently Vichi and Martella (2005) estimate parameters of the model according to a model-based likelihood approach.

The double *k*-means model can be modified to assess a partition of the objects along a set of centroids, as above, but also a partition of the variables along a set of linear combinations of maximum variance. Thus the model (1) is written (Vichi and Saporta, 2004)

$$\mathbf{X} = \mathbf{U}\,\overline{\mathbf{Y}}\,\mathbf{V}'\mathbf{B} + \mathbf{E}, \tag{2}$$

where matrix $\mathbf{B}$ is a diagonal matrix defined so that $\mathbf{V}'\mathbf{B}\mathbf{B}\mathbf{V} = \mathbf{I}_Q$ and $tr(\mathbf{B}\mathbf{B}) = Q$. An efficient alternating least-square algorithm is given.

3 Application

The short-term scenario of September 1999 on macroeconomic performance of national economies of twenty countries, members of the Organization for Economic Co-operation and Development (OECD) has been considered in the paper by Vichi and Kiers (2001) to test the ability of the factorial *k*-means analysis (which allows a simultaneous classification of objects and a component reduction for variables) in identifying classes of similar economies and help to understand the relationships within the set of observed economic indicators. The performance of the economies reflects the interaction of six main economic indicators: Gross Domestic Product (GDP), Leading Indicator (LI), Unemployment Rate (UR), Interest Rate (IR), Trade Balance (TB), Net National Savings (NNS). Variables have been standardized.

The classification, obtained by the tandem analysis, i.e. *k*-means applied on the first two principal components scores, when the number of clusters for the objects is equal to three and the number of components for the variables is equal to two, is displayed in figure 1.

The first PCA component is characterized mainly by net national savings, gross domestic product, whereas the second PCA component, by interest rate and trade balance. The unemployment rate characterizes both dimensions, as it can be observed from Table 1. The first component explains 28% of the total variance, while the second PCA dimension explains the 23%.

The classification of the countries is given below:

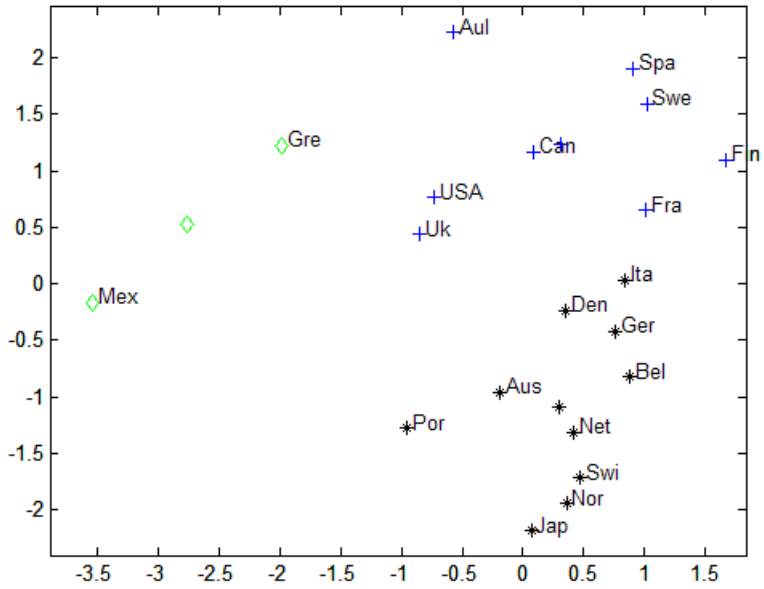


Figure 1. Tandem Analysis, i.e, *K*-means clustering computed on the first two principal components.

First class: Australia, Canada, Finland, France, Spain, Sweden, United Kingdom, United States;
Second class: Greece, Mexico,
Third class: Austria, Belgium, Denmark, Germany, Italy, Japan, Portugal, Netherlands, Norway, Switzerland.

Table 1: Component loadings defined by PCA

	GDP	IR	LI	UR	NNS	TB
Comp 2	-0.065	**-0.696**	**-0.229**	**0.367**	-0.092	**0.563**
Comp 1	**-0.567**	-0.175	-0.192	**-0.489**	**0.607**	0.059

Clustering and disjoint PCA (CDPCA) has been applied on the same data set by fixing the number of clusters for the objects and variables equal to three and two respectively. The component loadings matrix is shown in table 2, while the classification of the countries is given below:

First class: Australia, Canada, Denmark, Finland, France, Germany, Italy, Spain, Sweden, United Kingdom, United States;
Second class: Greece, Mexico, Portugal;
Third class: Austria, Belgium, Japan, Netherlands, Norway, Switzerland.

The first dimension of CDPCA is still characterized mainly by net national savings, and less strongly by gross domestic product, whereas the second CDPCA dimension by interest rate and trade balance. However, this time unemployment rate characterizes the first dimension only, as it can be observed from Table 2.
The first CDPCA dimension explains 26% of the total variance, while the second CDPCA dimension accounts for 21%.Thus, the loss of variance with respect to the PCA is irrelevant.

Table 2: Component loadings defined by Disjoint PCA

	GDP	IR	LI	UR	NNS	TB
Dim 2	0	-0.697	0.229	0	0	0.679
Dim 1	-0.383	0	0	-0.498	0.778	0

Comparing the two graphical representations in Figure 1 and 2, it can be observed that the CDPCA, more clearly shows three homogeneous classes, mainly representing the same countries of the tandem analysis with some relevant differences. These are mainly due to the role on the unemployment rate in the two analyses and less strongly by the leading indicator. In CDPCA UR characterizes the first dimension only, while it influences both dimensions in Figure 1. In Figure 2, Italy and Germany are positioned higher in the plot with respect to Figure 1 to better represent the higher unemployment rate they have.
In Figure 2 Mexico and Portugal also are located much closer because they have very similar values of GDP, LI and TB, which describe the first dimension of CDPCA.

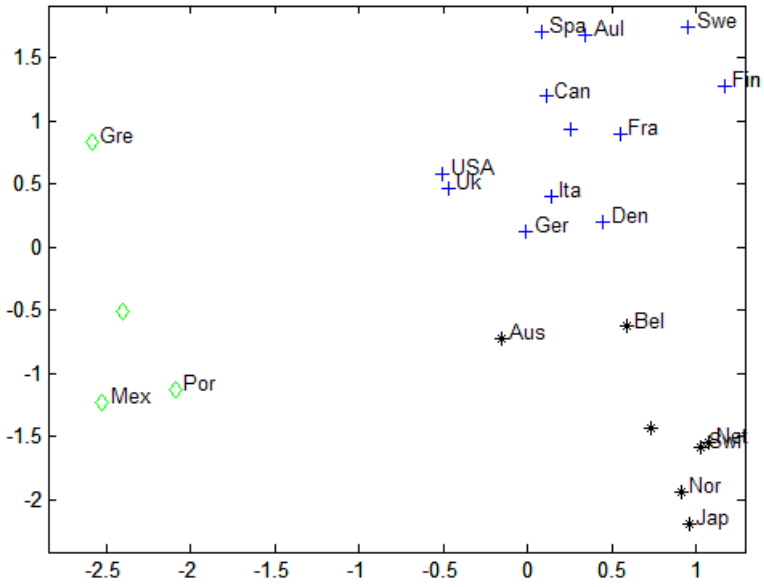


Figure 2. Clustering and Disjoint PCA.

4 Bibliography

ROCCI R, VICHI M., *Multimode partitioning*, 2004, submitted.

VAN MECHELEN, I., BOCK H. H. & DE BOECK, P., Two-mode clustering a structured overview. *Statistical Methods in Medical Research*, 2004, to appear.

VICHI, M., Double *k*-means Clustering for simultaneous classification of Objects and Variables. In Borra et al. (eds): *Advances in Classification and Data Analysis*, 43-52, 2000, Springer.

VICHI, M., KIERS, H.A.L, Factorial *k*-means analysis for two way data, *Computational Statistics and Data Analysis,* 37, 49-64, 2001.

VICHI, M, MARTELLA, F. *Model-based clustering for block-data*, 2005, submitted.

VICHI, M., SAPORTA G., *Clustering and Disjoint Principal Component Analysis*, 2004 submitted.

Communications

Couplage d'un problème de classification et d'estimation de densité par des noyaux gaussiens

Aaron Catherine

SAMOS-MATISSE
Université Paris 1
90 rue de Tolbiac Paris, France, 75013
catherine_aaron@hotmail.com

RÉSUMÉ. La classification et l'estimation de densité d'un nuage de point issu du tirage d'un mélange de lois sont deux problèmes intimement liés. En effet la connaissance de la densité induit une classification naturelle dans laquelle le nombre de classe est connu (et correspond au nombre de modes), d'autre part la connaissance de la classification permet de localiser dans l'espace les points correspondant à chacune des composantes du mélange et simplifie le problème de l'estimation de densité. Dans la pratique aucune de ces deux données n'est disponible. Dans ce papier on propose une méthode permettant de résoudre conjointement ces deux problèmes.

MOTS-CLÉS : Classification, estimation de densité, noyaux gaussiens, taille de fenêtre, cross-validation

1 Introduction

1.1 Problématique

On dispose de N observations dans $\mathbf{R}^p$ qui correspondent, par hypothèse, aux réalisations d'un mélange de k lois uni-modales et on se propose de résoudre le double problème : estimation de la densité du nuage de point (par une méthode à noyau) et segmentation des données. Dans un premier temps on va montrer en quoi ces deux problèmes sont implicitement liés.

1.2 Classification sous hypothèse de densité connue

Si on suppose que la densité totale du nuage est connue, Wishart a proposé, en 1969, une méthode de classification des données autour des domaines d'attraction des modes. Il y a autant de groupes que de modes et chaque point est affecté à la classe du mode qui « l'attire » (on peut lier le point au mode par un chemin croissant de densité). Visuellement cette classification est relativement naturelle (voir figure 1)

1.3 Densité sous hypothèse de classification connue

La principale difficulté pour l'estimation de densité par une méthode à noyau consiste à trouver une « bonne » taille pour la fenêtre des noyaux. Dans le cas d'une densité uni-modale on peut utiliser la cross-validation qui donne des résultats pertinents. Dans le cas de densités multi-modales avec hétérogénéité des dispersions autour des modes la recherche d'une taille unique (sur tout le nuage) de fenêtre est vouée à l'échec (voir figure 2). Dans ce cas on cherchera des tailles de fenêtres dépendant des points. Si la

classification des données est connue, une méthode naturelle serait de chercher une taille de fenêtre par classe.

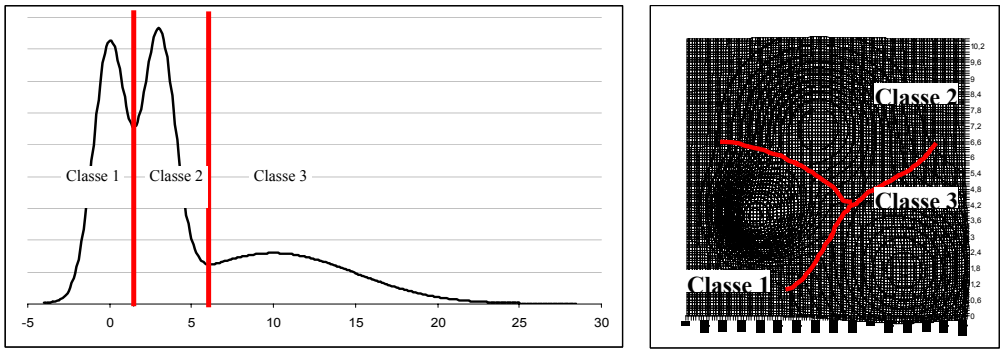


Figure 1 : classification sous hypothèse de densité connue

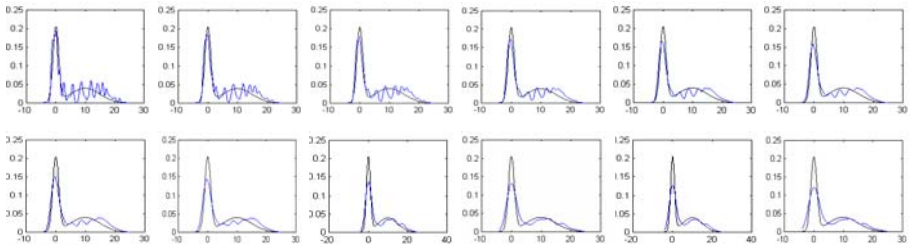


Figure 2 : Densité multi-modale avec hétérogénéité de dispersion et taille de fenêtre unique on n'estime jamais les deux composantes à la fois

2 Estimation de densité par noyaux et cross-validation

2.1 Estimation de densité par noyaux gaussiens

Dans toute la suite nous ne traiterons que le cas des variables uni-dimensionnelles pour simplifier l'écriture des équations mais la généralisation au cas multidimensionnel est aisé.

On note $\varphi(x,y,h)=\frac{1}{\sqrt{2\pi h^2}}\exp\left(-\frac{(x-y)^2}{2h^2}\right)$

L'estimation de la densité d'un nuage de N points x_i par noyaux gaussiens de taille de fenêtre h est donnée par : $\hat{f}_h(x)=\frac{1}{N}\sum_{i=1}^{N}\varphi(x,x_i,h)$ tout le problème se résumant à déterminer un h « correct ». Pour cela nous avons choisi la méthode dite de « cross-validation » exposée ci-dessous

2.2 Dans le cas d'une seule taille de fenêtre

On note $\hat{f}_{-i}^{h}(x)=\frac{1}{N-1}\sum_{j\neq i}\varphi(x,x_j,h)$ la densité estimée si on avait toute la base sauf le point i et on cherche à maximiser en h la pseudo-vraisemblance $L(h)=\prod_i \hat{f}_{-i}^{h}(x_i)$. L'annulation de la dérivée en h mène à : $h^2=\frac{1}{N}\sum_i\frac{\sum_{j\neq i}\varphi(x_i,x_j,h)(x_i-x_j)^2}{\sum_{j\neq i}\varphi(x_i,x_j,h)}$

2.3 Dans le cas de plusieurs tailles de fenêtres

Dans le cas où il y aurait K classes avec $\sigma(j)$ la classe de j la densité estimée est :

$\hat{f}_h(x) = \frac{1}{N}\sum_{i=1}^{N}\varphi(x, x_i, h_{\sigma(i)})$ et la maximisation de la pseudo-vraisemblance donne :

$$h_k^2 = \sum_i \frac{\sum_{j\neq i,\sigma(j)=k}\varphi(x_i,x_j,h_k)(x_i-x_j)^2}{\sum_{j\neq i}\varphi(x_i,x_j,h_{\sigma(j)})} \Bigg/ \sum_i \frac{\sum_{j\neq i,\sigma(j)=k}\varphi(x_i,x_j,h_k)}{\sum_{j\neq i}\varphi(x_i,x_j,h_{\sigma(j)})}$$

3 Algorithme de classification

L'algorithme proposé est un algorithme stochastique. En effet, ici, l'aspect stochastique a un double avantage d'une part on évite de converger vers le maximum de la vraisemblance en N classes avec une taille de fenêtre tendant vers 0 et, d'un point de vue pratique, on diminue notoirement le temps de calcul.

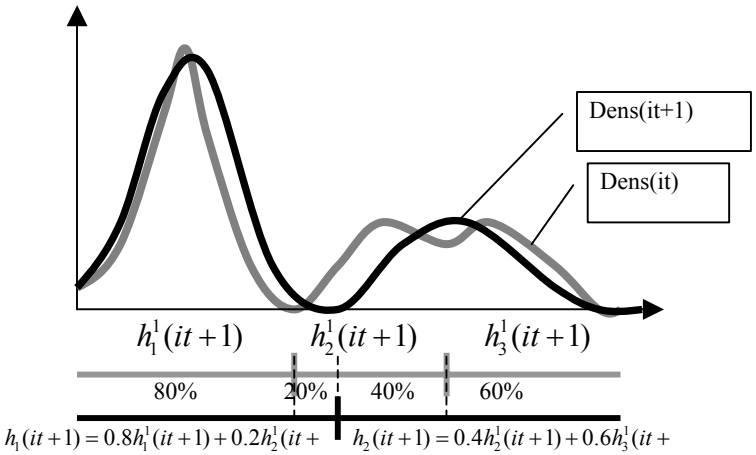
A l'état initial on a une seule classe et une taille de fenêtre h_0 puis, on itère Nit_1 :

- On tire $N_1 < N$ points y_j sans remise qui serviront de points sur lesquels on « posera » les noyaux
- On transforme les tailles de fenêtres pour maximiser la pseudo-vraisemblance de l'ensemble de la base en en effectuant Nit_2 fois :

$$h_k^1(it+1) := \sqrt{\sum_i \frac{\sum_{y_j\neq x_i,\sigma_{it}(j)=k}\varphi(x_i,y_j,h_k(it))(x_i-y_j)^2}{\sum_{y_j\neq x_i}\varphi(x_i,y_j,h_{\sigma_{it}(j)}(it))}} \Bigg/ \sum_i \frac{\sum_{y_j\neq x_i,\sigma_{it}(j)=k}\varphi(x_i,y_j,h_k(it))}{\sum_{y_j\neq x_i}\varphi(x_i,y_j,h_{\sigma_{it}(j)}(it))}$$

- On classe les données autour des modes observés pour obtenir la fonction σ_{it+1}
- Enfin obtient les nouvelles tailles de fenêtres par moyenne des valeurs de $h_k^1(it+1)$ sur la nouvelle classification : $h_{k'}(it+1) = \frac{\sum_{\sigma_{it+1}(i)=k'} h^1_{\sigma_{it}(i)}(it+1)}{\sum_{\sigma_{it+1}(i)=k'} 1}$

Pour finir, on effectue un dernier tour d'opération sur l'ensemble de la base et non uniquement sous un sous-ensemble tiré aléatoirement



on optimise les 3 tailles de fenêtres correspondants à dens(it) et on obtient trois tailles de fenêtres intermédiaires $h_1^1(it+1), h_2^1(it+1), h_3^1(it+1)$

Le regroupement autour de mode donne 2 classes

On calcule $h_1(it)$ et $h_2(it)$ comme moyennes des fenêtres intermédiaires

Figure 3 : Etapes de l'algorithme

4 Quelques résultats

Les résultats suivants sont le résultat d'estimation de densité et de classification sur des bases simulées en dimension 1 (ce qui permet de visualiser la différence entre les densités estimées et les « vraies » densités du tirage). Les lois simulées sont toutes des mélanges de gaussiennes. Les paramètres de l'algorithme sont, dans tous les cas : $N_1 = N/2$, $N_{it_1} = 10$ et $N_{it_2} = 3$

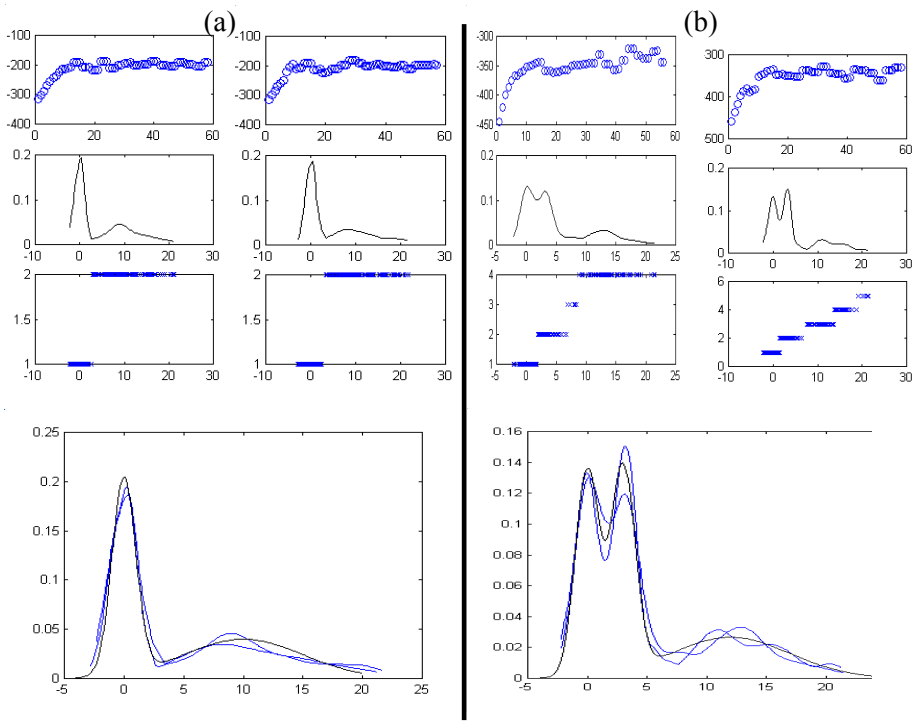


Figure 4 : Quelques résultats (a) : 200 points tirés pour moitié sur $\aleph(0,1)$ et pour moitié sur $\aleph(10,5)$ 2 exemples (pseudo-vraisemblance, estimation de densité et segmentation) et comparaison de la densité estimée à la vraie densité. (b) 300 points tirés pour tiers sur $\aleph(0,1)$, pour tiers sur $\aleph(3,1)$ et pour tiers sur $\aleph(12,5)$

5 Conclusion-perspectives

La méthode semble prometteuse mais nécessite encore des améliorations. En particulier, on aimerait construire un indicateur nous permettant de déterminer quand un mode est « significatif » afin de ne pas scinder en un trop grand nombre de classe (cf. figure4 exemple (b) où des modes annexes apparaissent qui, visuellement ne semblent pas « importants » mais qui numériquement induisent des erreurs de classification).

6 Bibliographie

.[BIC 0.3] BICEGO M, CRISTANI M., FUSIELLO A., MURINO V., *Watershed-based unsupervised clustering*, document de travail. http://profs.sci.univr.it/~bicego/bicego_murino_emmcvpr03.pdf.

[MIC 01] MICHALIS K., TITSIAS., ARISTIDIS C LIKAS., (2001), *Shared Kernel Models for Class Conditional Density Estimation*., IEE Transaction on Neural Network Vol 12 N°5 987-997

[SAI 96] SAIN S., SCOTT W., *On Locally Adaptive Density Estimation*, Journal of the American Statistical Association Vol 91 N°436 ftp://ftp.stat.rice.edu/pub/scottdw/Tech.Reps/adapt.ps.

.[STU 03] STUETZLE W., *Estimation of the cluster tree of a density by analysing the minimal spanning tree of a sample*, http://www.stat.washington.edu/wxs/Learning-papers/mst.pdf

[WIS 69] WISHART D., *Mode Analysis : A Generalization of Nearest Neighbor which Reduce Chaining Effect*, Numerical Taxinomy, Ed A.J. Cole, Academic Press,282-311

Analyse dissymétrique de la variance multivariée

Rafik Abdesselam

CREM - UMR-CNRS 6211, Université de Caen,
Esplanade de la paix, 14032 Caen, France
abdesselam@econ.unicaen.fr

RÉSUMÉ. Le modèle relationnel proposé peut être considéré comme un complément utile pour l'analyse de la variance multivariée (MANOVA). Il est en effet basé sur l'analyse statistique des individus muni d'un produit scalaire relationnel. Ainsi, une décomposition en deux sous-espaces orthogonaux associés aux valeurs moyennes et résiduelles, est présentée dans l'espace des individus. Pour analyser l'association symétrique ou dissymétrique entre les ensembles de variables à expliquer et explicatives (détermination des moments principaux et représentations graphiques), des analyses factorielles sont proposées pour décrire l'effet principal et résiduel du facteur contrôlé. Un modèle d'analyse dissymétrique est proposé puis comparé au modèle symétrique MANOVA. Enfin, un exemple sur données réelles (Iris de Fisher) est présenté.

MOTS-CLÉS : Modèle euclidien relationnel, analyse factorielle, MANOVA.

1 Introduction

Dans le modèle géométrique proposé, les vecteurs représentent les unités statistiques ou individus. On peut décrire différents nuages d'individus situés dans quatre sous-espaces, associés respectivement aux variables à expliquer (indépendantes), explicatives (indicatrices associées aux niveaux du facteur contrôlé), moyennes et résiduelles. Ce modèle est dit euclidien relationnel, car il est basé sur la notion de produit scalaire relationnel défini dans l'espace des individus. Cet espace est alors enrichi avec l'apport des informations contenues dans la structure des relations observées entre les variables dans l'espace des variables. Dans la section 2, on donne un bref rappel du produit scalaire relationnel, puis on montre qu'a priori le modèle euclidien relationnel général peut être simplifié. Quelques propriétés d'application pratique sont également présentées dans le cas de l'analyse symétrique ou dissymétrique de la MANOVA. Les deux analyses sont évaluées et comparées sur la base de données réelles dans la section 3.

2 Modèle euclidien relationnel pour la MANOVA

Soit $\{x^j ; j = 1,p\}$ et $\{y^k ; k = 1,q\}$ deux ensembles de variables centrées, observées sur la même population d'unités statistiques-individus. On note :
- $E_x = IR^p$ [resp. $E_y = IR^q$] est le sous-espace des individus, associé par dualité aux p variables continues $\{ x^j \}$ [resp. aux q variables indicatrices centrées $\{ y^k \}$ associées aux niveaux du facteur explicatif y],
- $\mathbf{X}_{(n,p)}$ [resp. $\mathbf{Y}_{(n,q)}$] est la matrice des données associée à l'ensemble des variables à expliquer $\{ x^j \}$ [resp. explicatives $\{ y^k \}$],
- Inx [resp. Iny] est l'injection canonique de E_x [resp. E_y] dans l'espace des individus $E = E_x \oplus E_y$,
- $N_x = \{ x_i ; i = 1,n \} \subset E_x$ [resp. $N_y = \{ y_i ; i = 1,n \} \subset E_y$] est le nuage des individus associé aux lignes de la matrice **X** [resp. **Y**],
- $\mathbf{M_x}$ [resp. $\mathbf{M_y}$] est la matrice du produit scalaire dans le sous-espace E_x [resp. E_y].

On pose $\mathbf{M_x} = {}^t\text{Inx}\ \mathbf{M}\ \text{Inx}$ et $\mathbf{M_y} = {}^t\text{Iny}\ \mathbf{M}\ \text{Iny}$, où $(\mathbf{M_x}, \mathbf{M_y})$ est un couple de produits scalaires euclidiens, **M** est un produit scalaire relationnel (Schektman 78) dans l'espace des individus $E = E_x \oplus E_y$, relativement aux ensembles de variables $\{ x^j \}$ et $\{ y^k \}$, si et seulement si :

$$\mathbf{M_{xy}} = {}^t\mathrm{Inx}\ \mathbf{M}\ \mathrm{Iny} = \mathbf{M_x}\,[(\mathbf{V_x M_x})^{1/2}]^+\,\mathbf{V_{xy}}\,\mathbf{M_y}\,[(\mathbf{V_y M_y})^{1/2}]^+$$

où, $\mathbf{V_x} = {}^t\mathbf{X\ D\ X}$, $\mathbf{V_y} = {}^t\mathbf{Y\ D\ Y}$, $\mathbf{V_{xy}} = {}^t\mathbf{X\ D\ Y}$ désignent les matrices de covariances, $\mathbf{D} = (1/n)\ \mathbf{I_n}$ est la matrice diagonale des poids des individus dans l'espace des variables noté F, $\mathbf{I_n}$ la matrice unité d'ordre n, et où $[(\mathbf{V_x M_x})^{1/2}]^+$ est l'inverse généralisée de Moore-Penrose de $(\mathbf{V_x M_x})^{1/2}$, relativement à $\mathbf{M_x}$.
On note $\mathbf{G}_{(n,p)}$ [resp. $\mathbf{R}_{(n,p)}$] la matrice des données associée aux variables moyennes $\{g^j = Q_y(x^j)\}$ [resp. résiduelles $\{r^j = x^j - g^j\}$], où, Q_y est l'opérateur de projection orthogonale sur l'image de $\mathbf{X}$, notée Im$\mathbf{Y} \subset$ F. On obtient les résultats classiques suivants :

- les variables $\{ g^j \}$ et $\{ r^j \}$ sont centrées.
- Im$\mathbf{G} \subset$ Im$\mathbf{Y}$, Im$\mathbf{R} \perp$ Im$\mathbf{Y}$, Im$\mathbf{X} \subset$ Im$\mathbf{G} \oplus$ Im$\mathbf{R}$.
- $\mathbf{V_{ry}} = \mathbf{V_{rg}} = 0$, $\mathbf{V_g} = \mathbf{V_{xg}} = \mathbf{V_{gx}}$, $\mathbf{V_{gy}} = \mathbf{V_{xy}}$, $\mathbf{V_r} = \mathbf{V_{xr}} = \mathbf{V_{rx}} = \mathbf{V_x} - \mathbf{V_g}$.

Comme pour les variables $\{x^j\}$ et $\{y^k\}$, un nuage de points-individus, noté N_g [resp. N_r], est associé aux variables $\{g^j\}$ [resp. $\{r^j\}$]. Le modèle relationnel proposé satisfait les hypothèses suivantes :

H_1) $E = E_x \oplus E_y \oplus E_g \oplus E_r$.
H_2) M est un produit scalaire relationnel dans E pour chacun des couples d'ensembles de variables.
H_3) Les produits scalaires dans les sous-espaces E_g et E_r sont égaux au produit scalaire euclidien M_x dans E_x. En effet, il est raisonnable de « voir » N_g et N_r de la même manière que N_x .

La construction statistique et géométrique du nuage $N_x^y = \{ P_y(x_i) ; x_i \in N_x \} \subset E_y \subset E$ à analyser, joue un rôle fondamental dans notre approche. E_y est le sous-espace explicatif sur lequel est projeté orthogonalement le nuage N_x. A priori $\mathbf{M_y}$ pourrait être quelconque ; le choix $\mathbf{M_y} = \chi^2_y$ (distance du khi-deux) simplifie les calculs. Pour $\mathbf{M_x}$ dans E_x, on utilise des générateurs de produits scalaires $\mathbf{M_x(\alpha)}$ afin de rechercher le « meilleur » produit scalaire, noté $\mathbf{M_x(\alpha^*)}$, qui maximise au mieux le critère du pourcentage d'inertie expliquée. Dans le contexte de notre approche, nous suggérons les expressions suivantes pour $\mathbf{M_x(\alpha)}$:

$$^1\mathbf{M_x(\alpha)} = [\ \alpha\mathbf{I_x} + (1-\alpha)\mathbf{V_x}\ ]^+ \qquad \text{et} \qquad ^2\mathbf{M_x(\alpha)} = \alpha\mathbf{I_x} + (1-\alpha)\mathbf{V_x} \qquad \text{avec } \alpha \in [0\ ,\ 1]$$

Ces générateurs vont évoluer de la position symétrique $^1\mathbf{M_x(0)} = \mathbf{V_x}^+$ (distance de Mahalanobis) vers la position dissymétrique $^2\mathbf{M_x(0)} = \mathbf{V_x}$ en passant par la position dissymétrique classique $^1\mathbf{M_x(1)} = \mathbf{I_x} = {}^2\mathbf{M_x(1)}$, où $\mathbf{I_x}$ est la matrice unité d'ordre p et $\mathbf{V_x}^+$ est l'inverse généralisée de Moore-Penrose de $\mathbf{V_x}$.

Le lemme suivant [SCH 00] permet de simplifier le modèle euclidien relationnel de la MANOVA.

Lemme

a) $\forall g_i \in E_g \ \| g_i - P_y(g_i) \| = 0$.
b) $\forall x_i \in E_x \ \| P_g(x_i) - P_y(x_i) \| = 0$.
c) $\forall x_i \in E_x \ \| x_i - (P_g + P_r)(x_i) \| = 0$.

Il découle du lemme b) que les représentations euclidiennes des nuages N_x^g et N_x^y sont identiques. Ainsi, le modèle relationnel peut être simplifié en prenant $E = E_x \oplus E_g \oplus E_r$, vu que les variables $\{y^k\}$ ne servent qu'à calculer les variables $\{g^j\}$: le sous-espace E_g remplace le sous-espace E_y. A noter que E_g est naturellement plus « riche » que E_y vu que $E_g \supset N_g$. Cette simplification est confirmée analytiquement.
De même, d'après le lemme c) les représentations euclidiennes de $N_{g+r} = \{P_g(x_i) + P_r(x_i) / x_i \in N_x\}$ et N_x sont identiques ; ainsi, le modèle peut encore se simplifier en prenant $E = E_g \oplus E_r$ et en remplaçant N_x par N_{g+r}. En d'autres termes, les deux matrices partitionnées suivantes,

$$X\ '(\ [\ (V_x M_x(\alpha^*)\ )^{1/2}\ ]^+\ )\ {}^t\begin{pmatrix} (\mathrm{V_g} M_x(\alpha^*)\ )^{1/2} \\ (V_r M_x(\alpha^*)\ )^{1/2} \end{pmatrix} \qquad \begin{pmatrix} M_x(\alpha^*) & 0 \\ 0 & M_x(\alpha^*) \end{pmatrix}$$

sont associées respectivement au nuage N_{g+r} et au produit scalaire relationnel M dans $E = E_g \oplus E_r$.

Propriété

Le modèle géométrique relationnel proposé, qui produit une analyse symétrique et dissymétrique de la MANOVA, consiste à effectuer les deux analyses en composantes principales (ACP) suivantes :

$$ACP\ [\ P_g(N_{g+r}) = N_x^{\ g}\ ;\ M\ ;\ D\ ] \qquad (1) \qquad ; \qquad ACP\ [\ P_r(N_{g+r}) = N_x^{\ r}\ ;\ M\ ;\ D\ ] \qquad (2)$$

On notera que lorsque $M_x(\alpha*) = V_x^{+}$, le modèle est symétrique, et est équivalent à la MANOVA : l'inertie expliquée, $I[N_x^{\ y}] = I[N_x^{\ g}]$ = trace $[V_g\ M_x]$, est égale à la trace de Pillai, un des indices d'association multivariés utilisés pour tester l'hypothèse d'égalité des moyennes. Dans les autres cas, on analyse un coefficient d'association dissymétrique correspondant à une MANOVA dissymétrique, nommée DMANOVA. Les moments et vecteurs axiaux principaux de $N_x^{\ g}$ et N_g [resp. $N_x^{\ r}$ et N_r] sont identiques ; de plus, les composantes principales correspondantes de $N_x^{\ g}$ [resp. $N_x^{\ r}$] appartiennent à Im$\mathbf{X} \subset$ F.

L'effet principal du facteur est décrit par la représentation simultanée de la projection orthogonale du nuage $N_x^{\ g} \cup N_g$ sur les plans principaux de l'ACP(1). Ainsi, le modèle relationnel permet naturellement d'enrichir ces premiers résultats avec ceux de la représentation simultanée de $N_x^{\ r} \cup N_r$ sur les plans principaux de l'ACP(2), c'est-à-dire, la description de l'effet résiduel du facteur contrôlé.

L'intérêt fondamental du modèle euclidien relationnel, dans le contexte de la MANOVA, est de proposer la décomposition orthogonale $x_i = P_g(x_i) + P_r(x_i)$ de chaque vecteur individu, relativement aux sous-espaces factoriel et résiduel. On obtient ainsi dans l'espace des individus E la décomposition classique de chaque variable $x^j = Q_g(x^j) + Q_r(x^j) = g^j + r^j$, qui existe dans l'espace des variables F. De plus, on montre que les espaces $E = E_g \oplus E_r$ et F sont liés par une isométrie, ce qui permet d'enrichir la représentation des individus sur les plans principaux tout en respectant la structure des relations observées entre les variables dans F.

3 Exemple numérique – Iris de Fisher

Pour évaluer l'intérêt du modèle, nous reprenons l'exemple analysé par Fisher (1936), concernant des observations relatives à trois espèces d'iris (*setosa*, *versicolor* et *virginica*) étudiées par le botaniste Edgar Anderson en Gaspésie (Québec). Quatre variables ont été observées sur 25 fleurs de chacune des espèces (longueur et largeur des pétales et des sépales).

MANOVA Test Criteria and F Approximations for the Hypothesis of No Overall species Effect
H = Type III SSCP Matrix for species E = Error SSCP Matrix S=2 M=0.5 N=33.5

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.01903064	107.79	8	138	<.0001
Pillai's Trace	1.30920004	33.17	8	140	<.0001
Hotelling-Lawley Trace	34.29934843	293.35	8	96.276	<.0001
Roy's Greatest Root	33.78890026	591.31	4	70	<.0001

Tableau 1. MANOVA - Critères multivariés

Le Tableau 1 donne les résultats des tests statistiques, traités par le logiciel SAS, habituellement utilisés en MANOVA. Ces quatre tests conduisent tous au rejet de l'hypothèse nulle d'égalité des moyennes.

Les représentations simultanées de l'effet du facteur (Fig. 1) et de l'effet résiduel (Fig. 2) permettent de comparer les résultats de la MANOVA et de la "meilleure" DMANOVA dont les pourcentages d'inertie expliquée, $I[N_x^{\ g}] / I[N_x]$ pour l'effet du facteur, sont respectivement de 32.73% et 87.73%. Pour les deux analyses, l'effet significatif du facteur-espèce contrôlé est représenté sur l'unique plan principal de la Figure 1. Les espèces sont bien séparées, l'espèce G_1 (I. *setosa*) est la plus éloignée des deux autres ; les espèces G_2 (I. *versicolor*) et G_3 (I. *virginica*) sont les plus rapprochées.

Sur le premier plan principal de la MANOVA, Figure 2, les points-individus projetés $N_x^{\ r}$ et les points-résidus correspondants N_r, notés R_1 (I. *setosa*), R_2 (I. *versicolor*) et R_3 (I. *virginica*), sont parfaitement confondus, cela signifie qu'il y a encore une partie non expliquée linéairement par le facteur. Ainsi, l'effet résiduel de la MANOVA qui compte quatre facteurs principaux, est réellement décrit sur le deuxième plan principal alors que l'effet résiduel de la DMANOVA, qui ne compte que trois facteurs, est représenté sur le premier plan principal. L'effet résiduel est ici jugé par l'ampleur des écarts (distances) entre les points projetés et les points résidus correspondants. On peut ainsi repérer les points-individus qui contribuent le plus à la création de l'effet résiduel. Cette vision géométrique du modèle peut être considérée comme un complément des résultats classiques de la statistique paramétrique.

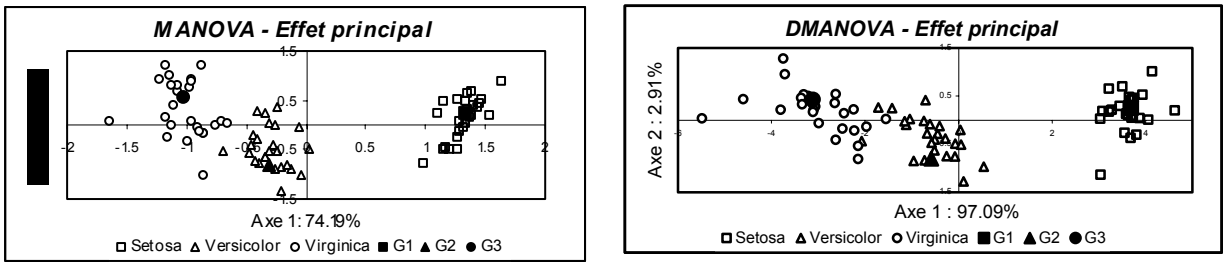


Fig. 1. ACP(1) : Représentation simultanée Nxg ∪ Ng

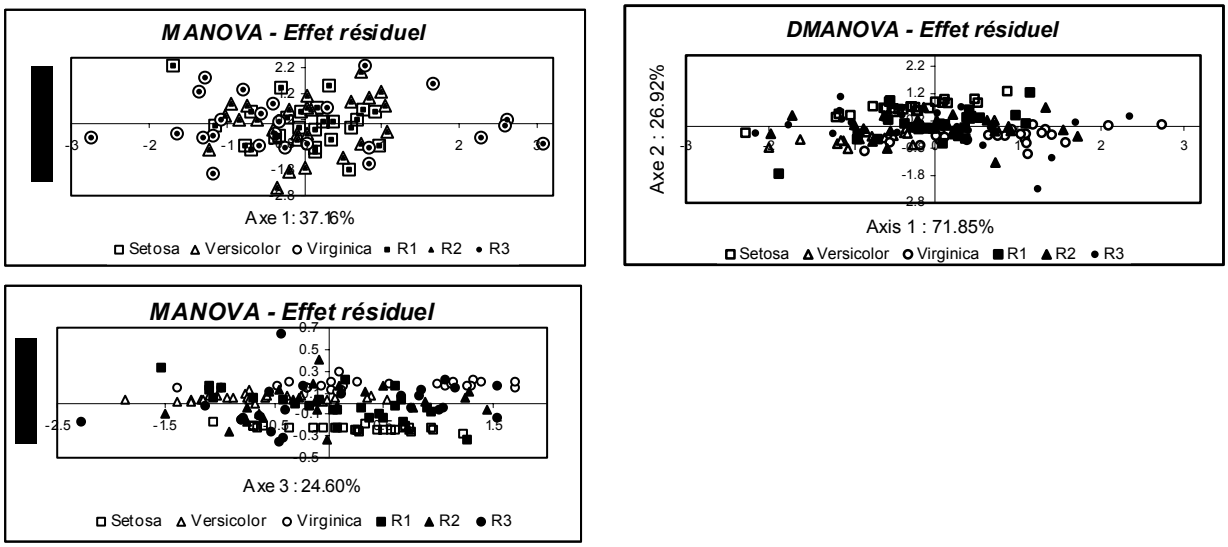


Fig. 2. ACP(2) : Représentations simultanées $N_x^r \cup N_r$

4 Conclusion et perspectives

Le modèle relationnel proposé est une règle formelle utile pour enrichir les résultats classiques de la MANOVA par ceux fournis par les deux analyses factorielles, décrivant l'effet principal et l'effet résiduel du facteur contrôlé. Il permet aussi de proposer une analyse dissymétrique DMANOVA, laquelle est souvent plus appropriée à la réalité observée. Ce modèle peut être utilisé pour synthétiser d'autres méthodes factorielles d'analyse des données [SCH 00, ABD 96]. Enfin, il peut évidemment être utilisé dans un contexte de classification ou de classement, lorsque les variables explicatives $\{y^k\}$ sont quantitatives et les variables $\{x^j\}$ sont les indicatrices d'une variable à expliquer x.

5 Bibliographie

[ABD 96] ABDESSELAM R., SCHEKTMAN Y., "Une analyse Factorielle de l'Association Dissymétrique entre deux variables qualitatives" *Revue Statistique Appliquée*, XLIV(2), 1996, p. 5-34.

[DAN 75] DAGNELIE P., "Analyse statistique à plusieurs variables" Gembloux, Pres. Agro., 1975, 362p.

[PAL 99] PALM R., "L'analyse de la variance multivariée et l'analyse canonique discriminante : principes et applications" *Notes de statistique et informatique*, Gembloux, Presses Agro., 1999, 40p.

[SCH 00] SCHEKTMAN Y., ABDESSELAM R., "A Geometrical Relational Model for Data Analyses", *Studies in Classification, Data Analysis and Knowledge Organization*. Data Analysis. Publisher W.Gaul, O.Opitz, M.Schader Editors, Springer, 2000, p. 359-368.

Optimisation de ressources dans la sélection de modèle des Machines à Vecteurs de Support

Mathias Adankon, Mohamed Cheriet, Nedjem Eddine Ayat

LIVIA, École de Technologie Supérieure
1100 Notre-Dame Ouest, Montréal, H3C 1K3, Canada
mathias@livia.etsmtl.ca, mohamed.cheriet@etsmtl.ca, nedjem@livia.etsmtl.ca

RÉSUMÉ : Le choix des paramètres du noyau d'une machine à vecteurs de support est très important et très déliquat parce que la performance du classifieur en dépend. Les méthodes développées pour réaliser l'ajustement automatique de ces paramètres nécessitent l'inversion de la matrice de Grahm-Schmidt ; ce qui requiert un espace mémoire important et un temps de calcul prohibitif. De plus, l'apprentissage des SVMs demande des ressources importantes en temps et en stockage au fur et à mesure que la base de données devienne large. Dans cet article, nous proposons une méthode accélérée de sélection de modèle des SVMs en utilisant une approximation du gradient de l'erreur empirique et une technique d'apprentissage incrémental des SVMs. Cette méthode appliquée à des problèmes disposant d'une quantité importante de données s'est avérée très efficace avec des résultats encourageants.

MOTS-CLÉS : SVM, apprentissage incrémental, sélection de modèle, optimisation du noyau .

1 Introduction

Les machines à vecteurs de support sont des machines lineaires particulières basees sur le critère de la maximisation de la marge qui leur donne un excellent pouvoir de generalisation. Elles sont les premières à utiliser les methodes à noyau [BA01], une technique qui consiste à projeter les donnees de l'espace des caracteristiques dans un autre espace de dimension plus elevee où les donnees qui etaient non lineairement separables peuvent le devenir. Les SVMs utilisent donc cette technique connue sous le nom de « kernel trick » et l'hyperplan definissant la frontière entre les classes est construit dans ce nouvel espace. La fonction permettant de realiser cette projection est appelee noyau.
Malgre que la capacite de generalisation est incorporee dans la structure des SVMs avec la minimisation du risque structurel [Vap98], le choix des paramètres du noyau affecte aussi la performance de ces derniers. Des travaux ont ete effectues pour realiser une bonne selection de paramètres [ACS02, CVBM01]. Dans ce papier, nous proposons une methode basee sur l'erreur empirique developpee en [ACS02] et que nous avons amelioree en utilisant deux techniques qui sont : l'approximation du gradient de l'erreur et l'apprentissage incremental des SVMs.
Cet article est structure de la manière suivante. En section 2, nous rappelons la technique d'optimisation des paramètres du noyau basee sur le critère de l'erreur empirique developpee en [ACS02]. Puis en sections 3 et 4, nous presentons respectivement la simplification du gradient de l'erreur empirique et la technique de l'apprentissage incremental. Enfin, la section 5 sera consacree à la presentation des resultats des experimentations et la section 6 à la conclusion.

2 Erreur Empirique

Dans cette section, nous allons decrire l'optimisation des paramètres du noyau des SVMs basee sur l'erreur empirique [ACS02]. Considerons un problème biclasse où les observations sont $\{(x_1, y_1), \ldots, (x_l, y_l)\}$ avec $x_i \in \mathbb{R}^d$ et $y_i \in \{-1, 1\}$. Une machine à vecteurs de support permet de determiner l'hyperplan optimal de separation entre les deux classes qui est defini par :

$$f(x_i) = \sum_{j=1}^{NVS} \alpha_j y_j k(x_j, x_i) + b \tag{1}$$

où α_j et b sont determines en resolvant le problème quadratique d'optimisation maximisant la marge, $j = 1, \ldots, NVS$ designe l'indice des observations associes aux reels α_j non nuls et $k : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ est la fonction noyau [SBS99]. En posant $t_i = (y_i + 1)/2$, l'erreur empirique pour chaque observation est definie par : $E_i = |t_i - \hat{p}_i|$.

Dans cette equation, $\hat{p}_i$ represente un estime de la probabilite a posteriori associe à l'observation x_i, et il est determine en utilisant la fonction logistique proposee par Platt [Pla00] : $\hat{p}_i = \frac{1}{1+exp(A.f_i+B)}$ avec $f_i = f(x_i)$ et les paramètres A et B sont evalues par le principe decrit par Platt.
En supposant que la fonction noyau depend d'un ou plusieurs paramètres, nous notons ces paramètres par le vecteur $\theta = (\theta_1, \ldots, \theta_n)$. L'optimisation des paramètres du noyau est realisee par l'algorithme de descente de gradient avec minimisation de $E = \sum E_i$ sur l'ensemble de validation, distinct de l'ensemble d'apprentissage.

3 Méthode approchée du gradient de l'erreur empirique

Soit N le nombre d'observations formant l'ensemble de validation, la derivee de l'erreur empirique est exprimee par :

$$\frac{\partial E}{\partial \theta} = \frac{\partial}{\partial \theta}\left(\frac{1}{N}\sum_{i=1}^{N} E_i\right) = \frac{1}{N}\sum_{i=1}^{N}\frac{\partial E_i}{\partial \theta} \tag{2}$$

avec

$$\frac{\partial E_i}{\partial \theta} = \frac{\partial E_i}{\partial f_i}.\frac{\partial f_i}{\partial \theta}$$

La première partie de la la derivee vaut : $\frac{\partial E_i}{\partial f_i} = \frac{\partial E_i}{\partial \hat{p}_i}.\frac{\partial \hat{p}_i}{\partial f_i}$

avec $\frac{\partial E_i}{\partial \hat{p}_i} = \frac{\partial |\hat{t}_i - \hat{p}_i|}{\partial \hat{p}_i} = \begin{cases} -1 & \text{si } \hat{t}_i = 1 \\ +1 & \text{si } \hat{t}_i = 0 \end{cases}$ et $\frac{\partial \hat{p}_i}{\partial f_i} = -A\hat{p}_i(1-\hat{p}_i)$

Ainsi, nous avons :

$$\frac{\partial E_i}{\partial f_i} = Ay_i\hat{p}_i(1-\hat{p}_i) \tag{3}$$

La seconde partie de la derivee est donnee par :

$$\frac{\partial f_i}{\partial \theta} = \frac{\partial}{\partial \theta}\left(\sum_{j=1}^{NVS}\alpha_j y_j k(x_j, x_i) + b\right) = \sum_{j=1}^{NVS} y_j \frac{\partial}{\partial \theta}\left[\alpha_j k(x_j, x_i)\right] + \frac{\partial b}{\partial \theta}$$

$$\frac{\partial f_i}{\partial \theta} = \sum_{j=1}^{NVS} y_j \left[\frac{\partial k(x_j, x_i)}{\partial \theta}\alpha_j + \frac{\partial \alpha_j}{\partial \theta}k(x_j, x_i)\right] + \frac{\partial b}{\partial \theta} \tag{4}$$

Nous constatons que le gradient $\frac{\partial f_i}{\partial \theta}$ est constitue essentiellement de deux principaux termes. Et lorsque nous posons $\alpha = (\alpha_1, \ldots, \alpha_{NVS}, b)$,nous pouvons utiliser l'approximation ci-dessous proposee par Chapelle et al. [CVBM01].

$$\frac{\partial \alpha}{\partial \theta} = -H^{-1}\frac{\partial H}{\partial \theta}\alpha^T \tag{5}$$

où

$$H = \begin{pmatrix} K^Y & Y \\ Y^T & 0 \end{pmatrix} \tag{6}$$

Dans l'equation (5), H represente la matrice hessienne de la fonction objective du SVM appelee matrice Grahm modifiee dont la taille est $(NVS+1) \times (NVS+1)$. Les composants K_{ij}^Y valent $y_i y_j k(x_i, x_j)$ et Y est le vecteur de taille $NVS \times 1$ contenant le label y_i des vecteurs de support.
Au cours des experimentations, nous avons note que les termes $\frac{\partial \alpha_j}{\partial \theta}k(x_j, x_i)$ sont negligeables par rapport aux termes $\frac{\partial k(x_j, x_i)}{\partial \theta}\alpha_j$. Ainsi, nous avons approche l'equation (4) par :

$$\frac{\partial f_i}{\partial \theta} = \sum_{j=1}^{NVS} y_j \alpha_j \frac{\partial k(x_j, x_i)}{\partial \theta} \tag{7}$$

Cette approche a l'avantage de nous dispenser du calcul de l'inverse de la matrice H de complexite minimale $O((NVS+1)^2)$ pour estimer le gradient de l'erreur empirique. Pour valider cette approximation, nous avons effectue plusieurs tests tant sur des donnees synthetiques que sur des donnees reelles. La figure 1, montre les courbes de variation de l'erreur empirique en fonction des iterations et celles du taux de l'erreur en validation au cours de la procedure d'optimisation d'un problème XOR, avec chevauchement des donnees. Nous remarquons que les courbes sont presque les mêmes aussi bien pour le gradient total (les courbes à gauche) que pour le gradient approche(les courbes à droite).

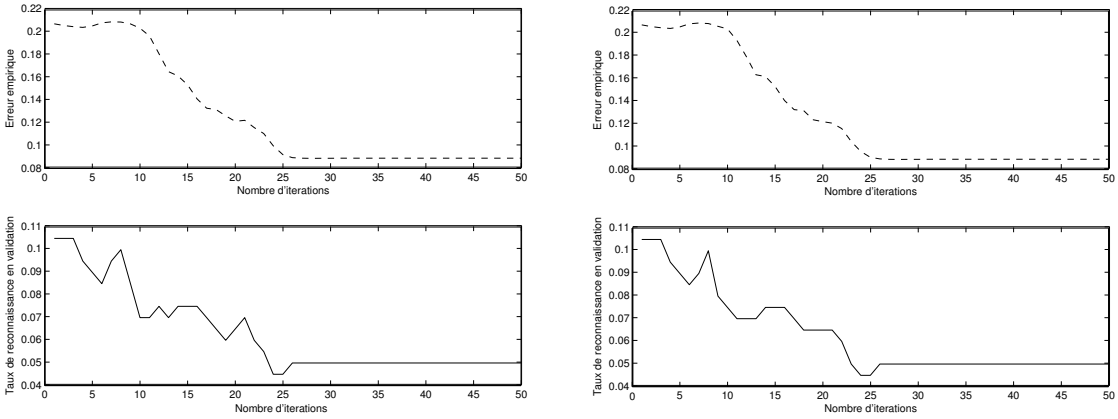


FIG. **1. Variation de l'erreur empirique et de l'erreur de test en validation au cours de l'optimisation : avec le gradient total(les courbes à gauche) et avec le gradient simplifie(les courbes à droite)**

4 Optimisation avec apprentissage incrémental

Cette methode a ete developpee en exploitant le pouvoir distinctif des SVMs de generaliser avec peu de donnees. Elle consiste à commencer l'optimisation avec un sous ensemble des donnees d'apprentissage que nous appelons ensemble actif. Et puis, au cours des iterations suivantes , on fait ajouter des donnees restantes, une idee introduite dans [SLS99]. Dans le souci de maintenir la taille de l'ensemble actif faible, nous supprimons de la base d'apprentissage les donnees qui sont loin de la marge. Ainsi nous gardons dans l'ensemble actif uniquement les exemples qui sont succeptibles d'être des vecteurs de support (observations situees dans et au voisinage de la marge provisoire) auxquels nous ajoutons de nouvelles observations pour les prochaines iterations.

1. *Initialiser les paramètres du noyau*
2. *Initialiser l'ensemble actif S*
3. *Répéter jusqu'à convergence*
  3.1 *Apprendre le SVM avec l'ensemble S*
  3.2 *Estimer les paramètres de la sigmoïde*
  3.3 *Calculer le gradient de l'erreur*
  3.4 *Corriger les paramètres du noyau*
  3.5 *Supprimer de S les données éloignées de la marge*
  3.6 *Ajourter à S une partie ΔS des données restantes*

FIG. **2. Algorithme d'apprentissage incremental jumele avec l'optimisation du noyau**

5 Résultats expérimentaux et discussions

Nous avons realise les experiences avec la base de donnees benchmark MNIST qui est un problème de reconnaissance de chiffres manuscrits, donc un problème à 10 classes. MNIST (Modified NIST) issu de la base NIST ,est constitue de 50000 observations pour l'apprentissage, 10000 pour la validation et 10000 pour le test.
Nous avons entraîne 45 classifieurs de base, en optant pour la technique "un contre un" , qui sont couples pour la decision finale. Les paramètres du noyau de chacun des 45 SVMs sont optimises localement. Nous avons utilise le noyau RBF en prenant C=100. Comme pour chaque apprentissage, nous avons environ 5000 exemples de chaque classe, nous initialisons S avec les 2500 premiers exemples et la taille de ΔS des donnees restantes ajoutee à chaque iteration est choisie dynamiquement.

Au cours du test de reconnaissance, la probabilite pour qu'un element appartienne à une classe $\omega_i (i = 1, ..., 10)$ est donee par : $p_i = \frac{1}{45} \sum_{i \neq j} \sigma(p_{ij})$ où $p_{ij} = P(x \in \omega_i / x \in \omega_i \cup \omega_j)$ et σ est la fonction de couplage, pour les details consulter [MM98].

Dans le tableau 1, nous presentons les differents resultats selon le couplage utilise. Ces resutats sont identiques pour les trois algorithmes d'optimisation que nous avons testes, à savoir l'algorithme sans approximation du gradient[ACS02], celui utilisant la valeur approchee du gradient et celui base sur l'apprentissage incremental avec le gradient approche. De plus nous avons eu ces mêmes resultats du tableau 1 avec C=10.
Au cours de nos experimentations, nous avons remarque que les trois algorithmes fournissent à la fin les mêmes

Modes de Couplage/$\sigma(x) =$	$\begin{cases} 1 & \text{si } x > 0.5 \\ 0 & \text{sinon} \end{cases}$	x	$\frac{1}{1+e^{-12(x-0.5)}}$	$\begin{cases} 1 & \text{si } x > 0.5 \\ x & \text{sinon} \end{cases}$	$\begin{cases} x & \text{si } x > 0.5 \\ 0 & \text{sinon} \end{cases}$
Taux d'erreur(%)	1.6	1.5	1.7	1.6	1.5

TAB. **1. Resultats obtenus sur le benchmark MNIST**

valeurs de paramètres de noyau à 10^{-3} près. Ainsi, à la phase de test, nous avons les mêmes taux d'erreur. Mais ils se distinguent surtout par le temps de calcul requis. Nous avons note que la difference de temps de calcul devient très interessante lorsque la taille de la base des donnees est importante, car le calcul du gradient approche est de complexite $O(N.NVS.n)$ tandis que celui du gradient total est $O(N.NVS.NVS.n)$. En figure 3, nous avons trois courbes ; chacune d'elles montre la variation du temps de calcul en fonction de la taille de la base de donnees. Sur cette figure nous pouvons notifier la reduction du temps obtenu lorsque nous passons d'une technique à une autre. Avec le gradient approche, nous reduisons le temps de calcul de 5% à 25% et en considerant en plus la technique d'apprentissage incremental, nous avons un gain accru variant entre 20% et 40%. Le non stockage de la matrice H de taille $(NVS + 1) \times (NVS + 1)$ permet entre autre de sauver de l'espace memoire .

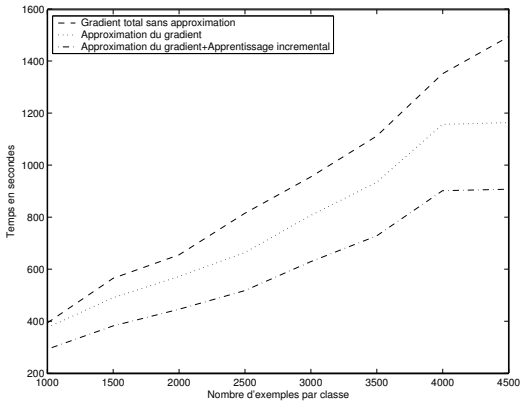


FIG. **3. Comparaison entre le temps de calcul requis par les trois methodes en fonction de la taille de l'ensemble d'apprentissage**

6 Conclusion

Dans cette communication, nous avons presente notre algorithme d'optimisation des paramètres de noyau des SVMs base sur l'apprentissage incremental des SVMs jumele avec une simplification de calcul du gradient de l'erreur empirique. Nous avons teste cette methode qui a donne de bons resultats, confirmant ainsi notre approche pour reduire le temps de calcul et l'espace memoire requis tout en preservant la qualite des paramètres optimises.

Références

[ACS02] N. E. Ayat, M. Cheriet, and C. Y. Suen. Empirical error based optimization of svm kernels : application to digit image recognition. In *Proc. of the Int.Workshop on Handwriting Recognition*, pages 292–297, Niagara, 2002.

[BA01] G. Baudat and F. Anouar. Kernel-based methods and function approximation. In *International Joint Conference on Neural Networks*, pages 1244–1249, Washington, 2001.

[CVBM01] O. Chapelle, V. Vapnik, O. Bousquet, and S. Mukherjee. Choosing multiple parameters for support vector machines. In *Machine Learning*, 2001.

[MM98] M. Moreira and E. Mayoraz. Improved pairwise coupling classification with correcting classifiers. In *Proceedings of the 10th European Conference on Machine Learning*, pages 160–171, 1998.

[Pla00] J. Platt. Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. pages 61–74. A.J. Smola, P. Bartlett, B. Schoelkopf, D. Schuurmans, 2000.

[SBS99] B. Schlkopf, C. J. C. Burges, and A. J. Smola. *Advances in Kernel Methods*. The MIT Press, Cambridge, Massachusetts, 1999.

[SLS99] N.A. Syed, H. Liu, and K.K. Sung. Incremental learning with support vector machines. In *International Joint Conference on Artificial Itelligence*, 1999.

[Vap98] V. Vapnik. *Statistical learning theory*. John Wiley and Sons, New York, 1998.

Méthode de suppression des règles d'association symboliques redondantes par la régression linéaire

Filipe AFONSO

Ceremade et Lamsade
Université Paris-Dauphine,
Place du Maréchal de Lattre de Tassigny
75775 Paris Cedex 16, France

RÉSUMÉ. Les règles d'association symboliques que l'on considère sont des règles extraites au niveau des concepts par un algorithme Apriori étendu. Notamment, pour l'exemple connu du panier de la ménagère, nous n'extrayons pas les règles au niveau des transactions. Nous découvrons des règles au niveau des clients, considérés comme des concepts dont l'extension sont leurs transactions, afin d'étudier leurs comportements d'achat. Dans cet article, nous proposons une méthode basée sur la régression linéaire afin de sélectionner les règles d'association symboliques intéressantes et non redondantes.

MOTS-CLÉS : règles d'association, données symboliques, régression linéaire

1 Introduction

Si $I=\{i_1,\ldots,i_n\}$ est un ensemble de n items et $T=\{t_1, \ldots, t_m\}$, $t_i \in P(I)-\varnothing$, un ensemble de m transactions alors une règle d'association est une règle telle que $A \rightarrow B$, $A \subset I$, $B \subset I$ (sous-ensembles d'items), $A \cap B = \varnothing$. Dans ([AGR 94]) , l'algorithme Apriori extrait ces règles d'association pour des supports et des confiances supérieurs à des seuils minimaux *minsup* et *minconf* où :

$$Sup\ (A \rightarrow B) = \frac{card\ (t \in T\ /\ A \cup B \subseteq t)}{card\ (T)}, \quad conf\ (A \rightarrow B) = \frac{\sup(A \rightarrow B)}{\sup(B)}.$$

Depuis, de nombreux travaux ont été menés afin de réduire la base de règles d'association. Notamment, Guigues et Duquenne [GUI 86] définissent une base minimale pour les règles d'association exactes (confiance=100%) alors que Luxenburger [LUX 91] définit une base minimale pour les règles d'association partielles (confiance<100%). Aussi, de nombreux indicateurs, outre le support et la confiance, ont été proposés afin d'évaluer la qualité des règles obtenues : confiance centrée, Loevinger, conviction, gain d'entropie, taux informationnels, lift, Piatetsky-Shapiro, Laplace.... Pour la plupart, ces indicateurs sont corrélés au support et à la confiance. Nous pouvons notamment nous référer à [BAY 99] et [BLA 04] qui reprennent les différents indicateurs. Dans [AFO 04], nous étendons l'algorithme Apriori afin d'extraire des règles au niveau des concepts décrits par des variables symboliques modales (voir [BOC 00]). Nous utilisons l'exemple connu du panier de la ménagère où au lieu d'extraire des règles d'association au niveau des transactions, nous découvrons des règles au niveau des concepts clients. Nous considérons l'exemple réduit (tableau 1) où pour chaque client, nous agrégeons dans la matrice les articles (items) achetés sous forme d'un diagramme construit avec la proportion de chaque article par rapport aux

achats totaux du client. Nous nous intéressons notamment aux associations entre les catégories d'items v=viande, p=poisson, c=céréales, f=fruits et légumes et l=produits laitiers. Nous avons alors une matrice symbolique où chaque ligne définit la "description" d'un client et chaque colonne est associée à une variable symbolique (une seule variable dans l'exemple). Les règles d'association extraites sont du type :

$$1/5 < P_v \leq 2/5 \wedge 0 < P_c \leq 1/5 \rightarrow 0 < P_p \leq 1/5 \qquad (*)$$

Qui se lit : « Si pour un client, la fréquence d'achat de viandes par rapport aux achats totaux est comprise entre 1/5 ouvert et 2/5 et la fréquence d'achat de céréales est comprise entre 0 ouvert et 1/5 alors la fréquence d'achat de poissons est comprise entre 0 ouvert et 1/5» avec un support et une confiance supérieurs à deux seuils minimaux. Les règles obtenues à partir de la matrice de données symboliques (tableau 1), pour *minsup*=50% et *minconf*=70%, sont données (tableau 2). Nous ne montrons pas ici comment nous les obtenons (voir [AFO 04]).

Ainsi, dans le cas de ces règles symboliques, nous proposons d'utiliser la régression linéaire afin d'étudier la qualité de nos règles et d'éliminer les règles d'association redondantes.

Concepts=Client	X=achats	Concepts=Client	X=achats
1	1/2v,1/6p,1/6c,1/6f	3	2/3v,1/3p
2	1/2v,1/3p,1/6c	4	2/3p,1/6c,1/6l

Tableau 1: Concepts clients décrits par une variable modale résumant les achats

N°	Règle	Sup%	Conf%	Régression	R²	F-test
1	$1/3<P_v\leq 2/3 \rightarrow 0<P_p\leq 1/3$	75	100	$P_p = 0.16+0.25P_v$	0.25	0.33
2	$0<P_p\leq 1/3 \rightarrow 1/3<P_v\leq 2/3$	75	100	$P_v = 0.25+P_p$	0.25	0.33
3	$0<P_c\leq 1/3 \rightarrow 0<P_p\leq 2/3$	75	100	$P_p = -0.1+2P_c$	0.57	1.33
4	$0<P_p\leq 2/3 \rightarrow 1/3<P_v\leq 2/3$	75	75	$P_v = 0.25+P_p$	0.25	0.33
5	$0<P_p\leq 2/3 \rightarrow 0<P_c\leq 1/3$	75	75	$P_c = 0.13+0.3P_p$	0.57	1.33

Tableau 2 : Règles d'association symboliques et les régressions linéaires associées

2 Etude des règles d'association symboliques par la régression linéaire

Nous remarquons que les règles symboliques contiennent de la variation. Si nous regardons la règle 1 (tableau 2), $1/3<P_v\leq 2/3 \rightarrow 0<P_p\leq 1/3$, nous ne pouvons pas savoir si lorsque P_v est proche de 2/3 alors P_p est plutôt proche de 1/3, ou proche de 2/3 ou bien varie dans l'intervalle]1/3,2/3] sans distinction. Pour étudier ces variations, nous utilisons la régression linéaire symbolique.

En effet, à l'aide de la régression linéaire, nous pouvons discriminer les règles symboliques de la forme:

$$\bigwedge_{i,u} [\underline{x}_{i,u} < P_{X\,i,u} \leq \overline{x}_{i,u}] \rightarrow [\underline{y}_v < P_{Yv} \leq \overline{y}_v]$$

où $P_{Xi,u}$ (P_{Yv}) désigne la fréquence de la catégorie u (v) de la variable modale X_i (Y) et $\underline{x}_{i,u}$($\underline{y}_v$), $\overline{x}_{i,u}$($\overline{y}_v$) sont les bornes des intervalles de fréquences. En effet, pour des règles avec une seule propriété en conclusion (i.e. un seul intervalle de fréquences), nous calculons la régression des fréquences en prémisse sur les fréquences en conclusion en ne conservant uniquement que les individus dans « l'extension de la règle ». Nous obtenons alors une équation linéaire:

$$P_{Yv} = \beta_0 + \sum_{i,u} \beta_{i,u} P_{X\,i,u} + \varepsilon$$

Dans le hyper-rectangle: $\bigwedge_{i,u} [\underline{x}_{i,u} < P_{X\,i,u} \leq \overline{x}_{i,u}] \wedge [\underline{y}_v < P_{Yv} \leq \overline{y}_v]$

où ε désigne le résidu de la régression, β_0 la constante et les $\beta_{i,u}$ désignent les coefficients des variables $P_{Xi,u}$ dans la régression.

Nous pouvons alors mesurer la qualité des règles grâce aux indicateurs de la régression linéaire: le coefficient de détermination R² (mesure la proportion de la variation de la variable à expliquer prise en compte par le modèle) et le test de Fisher-Snedecor (*F*-test) de validité de la régression (voir [PRU 96]).

Nous donnons (tableau 2) les régressions linéaires associées à chaque règle d'association. Par exemple, pour la règle 1, $1/3 < P_v \leq 2/3 \rightarrow 0 < P_p \leq 1/3$, nous faisons la régression des poids P_v sur les poids P_p en conservant uniquement les individus avec un poids P_v dans l'intervalle]1/3,2/3] et un poids P_p dans l'intervalle]0,1/3], c'est-à-dire les individus 1, 2 et 3 (voir tableau 3). Nous obtenons $P_p = 0.16 + 0.25 P_v$ lorsque P_v est dans l'intervalle]1/3,2/3] avec une part de variation de P_p expliquée par P_v de R^2=25%. Ce résultat est donné à titre d'exemple étant donné que le nombre d'individus est ici trop faible.

Individus de la régression	P_v	P_p	P_c	Individus	P_v	P_p	P_c
1	1/2	1/6	1/6	3	2/3	1/3	0
2	1/2	1/3	1/6	4	0	2/3	1/6

Tableau 3: Matrice où les fréquences des diagrammes deviennent les variables de la régression linéaire

3 Réduction du nombre de règles à l'aide du test de Student

Le test de Student (voir [PRU 96]) est un test de nullité des paramètres ($\beta_{i,u}$) dans l'équation de régression en supposant que les résidus normalisés suivent une loi normale centrée réduite. Il permet donc de supprimer d'une régression linéaire des variables qui ont une faible capacité prédictive.

Par conséquent, nous pouvons appliquer ce test aux régressions linéaires calculées pour nos règles symboliques. Ainsi, le test de Student supprime, pour chaque règle, les prémisses qu'il considère non significatives. En fait, nous remarquons que si la règle d'association $r = C \rightarrow D$, où C est une conjonction de k intervalles de fréquences et D un intervalle de fréquences unique, appartient à notre base de règles *R* alors la règle d'association $E \rightarrow D$ est fréquente pour tout E conjonction de k-1 propriétés telles que $E \subset C$. Ceci implique que si le test de Student supprime une prémisse alors nous retombons sur une règle déjà dans *R* ou de confiance inférieure au seuil minimum. Par conséquent, dès que le test de Student supprime une variable prédictive dans une régression linéaire alors nous supprimons la règle correspondante.

Par exemple, supposons que nos concepts soient décrits par deux variables modales X défini avec 3 modalités (a, b, c) et Y défini avec 3 modalités (e, f, g). Nous notons P_{Xa}, P_{Xb}, P_{Xc}, P_{Ye}, P_{Yf}, P_{Yg}, les fréquences des modalités a, b, c, e, f, g respectivement. Supposons que nous ayons extrait les 3 règles :

- Règle 1 : $0 < P_{Xa} \leq 1/10 \rightarrow 0 < P_{Yg} \leq 1/2$
- Règle 2 : $0 < P_{Xa} \leq 1/10 \wedge 3/10 < P_{Xb} \leq 7/10 \rightarrow 0 < P_{Yg} \leq 1/2$
- Règle 3 : $0 < P_{Xa} \leq 1/10 \wedge 3/10 < P_{Xb} \leq 7/10 \wedge 0 < P_{Ye} \leq 1/2 \rightarrow 0 < P_{Yg} \leq 1/2$

Supposons que nous calculions comme précédemment la régression des fréquences en prémisse de la règle 3 (P_{Xa}, P_{Xb}, P_{Ye}) sur les fréquences en conclusion (P_{Yg}). Nous obtenons un équation linéaire du type :

$$P_{Yg} = \beta_0 + \beta_a P_{Xa} + \beta_b P_{Xb} + \beta_e P_{Ye} + \varepsilon.$$

Nous appliquons le test de Student à cette équation. Supposons que le test de Student supprime la variable P_{Ye}. Nous supprimons alors la propriété $0 < P_{Ye} \leq 1/2$ de la règle 3. Nous obtenons la règle 2 et nous supprimons la règle 3 de la base de règles. Supposons maintenant que le test de Student supprime aussi la variable P_{Xb}. Nous supprimons alors la propriété $3/10 < P_{Xb} \leq 7/10$ de la règle. Nous obtenons la règle 1 et nous supprimons les règles 2 et 3 de notre base. Ainsi, grâce à la régression linéaire et à son test de Student, nous supprimons les règles redondantes et nous réduisons substantiellement notre base de règles.

Remarque : Avant d'appliquer les tests de Fisher et de Student, un test de normalité des résidus (Shapiro et Wilk) est nécessaire si le nombre d'individus est trop faible pour utiliser le théorème central limite.

4 Application

Nous prenons comme exemple la base de données d'une société comptoir fournie avec le logiciel Access de Microsoft. Cette base répertorie 2155 enregistrements pour 830 transactions qui sont des sous-ensembles de 8 catégories différentes de produits (items). Nous construisons les concepts clients, soient 89 concepts. Après la construction des concepts, chaque client est décrit par un diagramme résumant sa consommation. Nous appliquons l'algorithme « Apriori symbolique » aux 89 concepts clients avec quatre paires (*minsup*, *minconf*) différentes ((10%,70%), (8%,70%), (6%,70%), (6%,25%)). Nous donnons (tableau 4), le nombre de règles d'association obtenues avant et après l'application du test de Student.

L'application du test de Student de la régression linéaire a permis d'élaguer substantiellement notre base de règles en supprimant les règles d'association redondantes. Pour le test (*minsup*=6%, *minconf*=25%), nous obtenons 12555 règles respectant le seuil *minconf* et seulement 688 règles après l'application du test de Student. De même pour une confiance>70%, nous obtenons 360 contre 7249 règles après et avant Student.

	minsup=10%,*minconf*=70%	8%,70%	6%,70%	6%,25%
Règles *minconf*	580	1180	7249	12555
Règles *Student*	58	157	360	698

Tableau 4: Evolution du nombre de règles d'association obtenues avant et après le test de Student

De plus, nous avons, par exemple, extrait les règles symboliques suivantes (où P_i est le poids du produit i dans les achats du client) pour lesquelles nous donnons les régressions linéaires associées (tableau 5) : règle 1 = $\{0<P_7\leq 1/6 \wedge 0<P_3\leq 1/3 \wedge 0<P_1\leq 1/3 \wedge 0<P_8\leq 1/2 \rightarrow 0<P_4\leq 1/3\}$; règle 2 = $\{0<P_7\leq 1/3 \wedge 0<P_4\leq 1/2 \wedge 0<P_3\leq 1/2 \rightarrow 0<P_1\leq 1/3\}$; règle 3 = $\{1/6<P_4\leq 1/3 \rightarrow 0<P_2\leq 1/6\}$. Ces règles ont de bonnes confiances (de 70% à 98%) mais les coefficients de détermination varient plus fortement (de 14% à 48%) et le test de Fisher ne rejette pas les deux premières régressions au risque 5% (F>f(0.05) quantile de la loi de Fisher) et rejette la troisième. Ainsi, nous nous servons de ces indicateurs pour distinguer les meilleures règles.

Règle	Confiance	Régression	R^2	*F*	f(0.05)
1	98%	$P_4=0.45-0.48P_7-0.56P_3-0.49P_1-0.45P_8$	48%	10.3	2.69
2	93%	$P_1=0.33-0.46P_7-0.37P_4-0.33P_3$	34%	8.5	2.92
3	70%	$P_2=0.17-0.36P_4$	14%	4.0	4.26

Tableau 5: Régressions linéaires associées au règles 1, 2, 3 avec le coefficient R^2 et le test de Fisher F

5 Conclusion

Nous avons proposé une méthode afin d'étudier et de sélectionner les règles d'association symboliques non redondantes à partir de la régression linéaire. Cette méthode nous semble efficace et enrichissante et présente également l'avantage de ne pas être basée sur la définition du support.

6 Bibliographie

[AFO 04] AFONSO F., "Extension de l'algorithme Apriori et des règles d'association au cas des données symboliques diagrammes et sélection des meilleures règles par la régression linéaire symbolique", *Revue RNTI, Classification et fouilles de données*, Zighed D.A. et Venturini G. eds, 2004, Cépadues.

[AGR 94] AGRAWAL R., SRIKANT R., "Fast algorithms for mining association rules", *Proc. of the 20th Int'l Conf. on Very Large Databases*, 1994.

[BAY 99] BAYARDO Jr R.J., AGRAWAL R., "Mining the most interesting rules", *Proc. of the Fifth ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining*, 1999, p. 145-154.

[BLA 04] BLANCHARD J., GUILLET F., GRAS R., BRIAND H., "Mesurer la qualité des règles et de leurs contraposées avec le taux informationnel TIC", *revue RNTI*, *Extraction et Gestion des Connaissances,* Vol.1, 2004, p. 287-298, Cépadues, Paris.

[BOC 00] BOCK H-H., DIDAY E., *Analysis of Symbolic Data, Exploratory methods for extracting statistical information from complex data*, Springer Verlag, Heidelberg, 2000.

[GUI 86] GUIGUES J.L., DUQUENNE V., "Famille minimale d'implications informatives d'un tableau de données binaires", *Mathématiques et sciences humaines*, année 24, 95, 1986, p. 5-18.

[LUX 91] LUXENBURGER M., "Implications partielles dans un contexte", *Mathématiques informatique et sciences humaines*, année 29, 113, 1991, p. 5-18.

[PRU 96] PRUM B., *Modèle linéaire : Comparaison de groupes et régression*, Editions INSERM, 1996.

Fouille visuelle de dissimilarités à l'aide de matrices de scatterplots pseudo-euclidiennes

S. Aupetit, N. Monmarché, M. Slimane

Laboratoire d'Informatique,
Université François-Rabelais, Tours,
64, Avenue Jean Portalis
37200 Tours, France

RÉSUMÉ. Nous présentons une nouvelle approche de fouille visuelle de dissimilarités utilisant les espaces pseudo-euclidiens et nous l'appliquons à la visualisation de dissimilarité sur des modèles de Markov cachés.

MOTS-CLÉS : fouille visuelle de données, dissimilarité, matrice de scatterplots, pseudo-euclidien, modèle de Markov caché, analyse en composantes principales à noyau indéfini

1 Introduction

Il existe de nombreuses techniques pour la fouille de données [CAR 99] [SPE 01]. Cependant, peu sont adaptées à la visualisation de dissimilarités. Le multi-dimensional scaling [COX 01] est une des méthodes permettant la visualisation de dissimilarités (VD). La plupart des techniques de VD ont cependant un gros inconvénient : elles ne permettent pas de représenter les informations sans déformation. Bien que la déformation soit négligeable pour certaines dissimilarités et puisse même être une source d'information [CAM 02], ils en existent pour lesquelles la déformation est si importante que l'analyse en est imprécise voire erronée. Des travaux récents tels que [PEK 01] et [ONG 04] ont mis en évidence qu'il était possible d'exploiter des dissimilarités pour l'apprentissage et la classification en se plaçant dans des espaces pseudo-euclidiens (EPE). Dans cet article, nous montrons comment les EPE peuvent être exploités pour la VD tout en permettant à l'utilisateur de choisir le niveau de précision souhaitée. Pour cela, nous introduirons brièvement l'analyse en composantes principales à noyau indéfini, nous présenterons comment la technique de la matrice de scatterplots peut être étendue pour la VD dans les EPE. Finalement, nous nous servirons de cette représentation pour comparer des dissimilarités pour des modèles de Markov cachés.

2 Analyse en composantes principales à noyau indéfini

Soit ε un espace vectoriel de dimension N et q une forme quadratique sur ε^2 telle que $\forall(x,y)\in\varepsilon^2$, $q(x,y)=x'My$ avec M une matrice symétrique de dimension $N\times N$. Le couple (ε,q) définit alors un espace pseudo-euclidien (EPE) [PEK 01]. Soit w une dissimilarité sur l'ensemble fini d'éléments $E=\{x_1,...,x_T\}$. Soit $\psi:E\mapsto\varepsilon$ une fonction de plongement des éléments. Plonger l'ensemble des

points de E dans l'espace ε, tout en conservant les dissimilarités entre les points, consiste à trouver q et ψ tels que $\forall(x,y)\in E$, $w^2(x,y)=q(\psi(x)-\psi(y),\psi(x)-\psi(y))$. On sait [PEK 01] et [ONG 04] que q et ψ existent toujours. Il est possible de trouver une solution à ce problème en imposant une contrainte supplémentaire. Trouver q et ψ revient en pratique à trouver la matrice M et des coordonnées pour les points images de E par ψ. On peut montrer que la matrice M peut toujours être mise sous la forme $\begin{bmatrix} I_a & 0 \\ 0 & -I_b \end{bmatrix}$ avec I_k la matrice carrée identité de dimension $k\times k$ et avec $1\le a,b\le N$. Soit $c\in\varepsilon$ un point de l'espace ε. On note $\psi_c(x)=\psi(x)-c$, $\forall x\in E$, le vecteur colonne partant du point c et se terminant au point $\psi(x)$. Soient la matrice $G=\left(q(\psi_c(x_i),\psi_c(x_j))\right)_{1\le i,j\le T}$, $\{\lambda_1,...,\lambda_p\}$ les p valeurs propres strictement positives de G et $\{\lambda_{p+1},...,\lambda_{p+n}\}$ les n valeurs propres strictement négatives de G. On note alors $\{V_1,...,V_p,V_{p+1},...,V_{p+n}\}$ les vecteurs propres associés tels que $\forall i=1..p+n$, $V_i'V_i=1$. Si $n=0$ alors il est possible d'utiliser une Analyse en Composantes Principales à Noyau (ACPN) [SCH 99] (également connus depuis les travaux de Gower en 1958 dans le domaine de l'analyse de données comme la décomposition de la matrice de Torgerson (1966) associée à w) afin d'obtenir des coordonnées pour les points en projetant les points sur les axes principaux. Cependant, pour une dissimilarité quelconque, n est généralement non nul. Pour effectuer le même type de démarche que l'ACPN, nous définissons l'Analyse en Composantes Principales à Noyau Indéfini (ACPNI). Pour cela, on impose que la matrice M soit de la forme $\begin{bmatrix} I_p & 0 \\ 0 & -I_n \end{bmatrix}$. En notant $\{U_i\}_{1\le i\le p+n}$ les axes principaux du nuage, on a $U_i=1/\sqrt{|\lambda_i|}\sum_{j=1..T}V_{i,j}\psi_c(x_j)$ et $q(\psi_c(x),U_i)=1/\sqrt{|\lambda_i|}\sum_{j=1..T}V_{i,j}q(\psi_c(x),\psi_c(x_j))$. On remarque alors que $\forall i=1..p+n$, $q(U_i,U_i)=\pm1$ car $V_i'V_i=1$ donc les U_i expliquent au mieux le nuage. Pour garantir une explication maximale du nuage par les axes principaux, il est nécessaire que le point c soit le centre du nuage. En notant $a=\psi(x_1)$, le centrage s'effectue grâce à la formule $q(\psi_c(x),\psi_c(y))=q(\psi_a(x),\psi_a(y))-T^{-1}\sum_{i=1}^{T}q(\psi_a(x),\psi_a(x_i))-T^{-1}\sum_{t=1}^{T}q(\psi_a(y),\psi_a(x_i))+T^{-2}\sum_{i=1}^{T}\sum_{j=1}^{T}q(\psi_a(x_i),\psi_a(x_j))$.

3 Matrices de scatterplots pseudo-euclidienne

A partir de toute dissimilarité, nous pouvons donc à l'aide de l'ACPNI plonger les points, sans perte d'information, dans un EPE. Il ne reste plus qu'à représenter cet EPE. Pour cela, nous proposons d'étendre la technique de la matrice de scatterplots [CAR 99]. Il est donc nécessaire d'étudier les EPE de dimension 2. On nomme signature $\sigma(q)$ de q le couple (p,n) associé à la matrice M obtenu par ACPNI. Les propriétés associées à (ε,q) dépendent alors de $\sigma(q)$. Si $\sigma(q)=(2,0)$ ou si $\sigma(q)=(0,2)$ alors, respectivement, (ε,q) et $(\varepsilon,-q)$ sont deux espaces euclidiens dans lesquels les proximités apparentes des coordonnées correspondent aux dissimilarités réelles. Mais si $\sigma(q)=(1,1)$ alors deux points ayant des coordonnées proches peuvent être très dissimilaires [PEK 01]. Pour comprendre pourquoi cela peut se produire, il suffit d'étudier le cône isotrope de q (i.e. l'ensemble des points $x\in\varepsilon$ tel que $q(x,x)=0$) et les iso-lignes de la dissimilarité (i.e. les ensembles de points $x\in\varepsilon$ tel que $q(x,x)$ a une valeur fixée). Le cône isotrope et les iso-lignes sont décrites par la figure 1. La matrice de scatterplots pseudo-euclidienne (MSPE) s'obtient à partir de la matrice de scatterplots à laquelle s'ajoute les fonctionnalités suivantes. A la demande de l'utilisateur, par un clic de souris, les iso-lignes sont superposées à la représentation en prenant comme centre la position actuelle de la souris. Pour appréhender plus facilement la contribution des différents axes, les scatterplots sont mis à l'échelle des valeurs propres associées (cf. figure 2). Les points de la matrice sont représentés par des carrés pouvant être remplis à la demande par une couleur dépendante des données (par exemple la classe) ou par une image (par exemple une photographie).

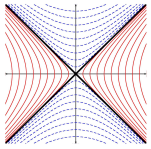


Figure 1 Iso-lignes d'un espace pseudo-euclidien de dimension 2 et de signature (1,1)

4 Visualisation de dissimilarités pour des MMC

Les Modèles de Markov Cachés (MMC) [RAB 89] sont des modèles stochastiques. Ils sont utilisés dans de nombreux domaines [CAP 01]. Un MMC λ est défini par trois matrices (A, B, Π). $A = (a_{i,j})_{1 \le i,j \le N}$ définit les probabilités de transition entre les N états cachés du système, $B = (b_i(j))_{1 \le i \le N, 1 \le j \le M}$ définit les probabilités d'émission des M symboles dans chacun des états cachés et $\Pi = (\pi_i)_{1 \le i \le N}$ définit les probabilités d'initialisation. Bien qu'il existe de nombreuses dissimilarités entre deux MMC dans la littérature [RAB 89] [FAL 95] [VIH 02], aucune ne permet la comparaison rapide de deux MMC ayant un nombre d'états cachés différents. Soient les deux distributions d'états cachés $\gamma_e = (1/N, ..., 1/N)$ et $\gamma^{\lambda}_{\infty} = \lim_{n \to \infty} A^n \Pi$ pour un modèle λ. Soient les distributions de symboles $\mu^{\lambda}(\gamma)$ et $\rho^{\lambda}(\gamma)$ pour le modèle λ définies par $\mu_k = \sum_{i=1}^{N} \gamma_i b_i(k)$ et $\rho_{k,l} = \sum_{i=1}^{N} \sum_{j=1}^{N} \gamma_i b_i(k) a_{i,j} b_j(l)$. Soit MER la moyenne de l'entropie relative entre les distributions u et v. Les dissimilarités suivantes permettent alors de comparer deux MMC x et y avec des nombres d'états cachés différents.

$$w_e^{(1)}(x,y) = MER(\mu^x(\gamma_e), \mu^y(\gamma_e)) \qquad w_e^{(2)}(x,y) = MER(\rho^x(\gamma_e), \rho^y(\gamma_e))$$

$$w_\infty^{(1)}(x,y) = MER(\mu^x(\gamma^x{}_\infty), \mu^y(\gamma^y{}_\infty)) \qquad w_\infty^{(2)}(x,y) = MER(\rho^x(\gamma^x{}_\infty), \rho^y(\gamma^y{}_\infty))$$

Le choix de ces dissimilarités ne sera pas discuté dans ce papier par manque de place. Nous possédons maintenant quatre dissimilarités mais nous ne savons pas comment elles structurent l'espace des MMC ni même si elles permettent d'effectuer de la classification. Nous avons donc choisi de les visualiser sur un ensemble de MMC appris. Pour cela, nous considérons la base d'images ORL [SAM 94]. Les 10 visages des 5 premières personnes de la base ont été appris via l'algorithme génétique décrit dans [SLI 96]. Les modèles obtenus possèdent des nombres d'états cachés différents. Les dissimilarités sont représentées via la MSPE (cf. figure 2).

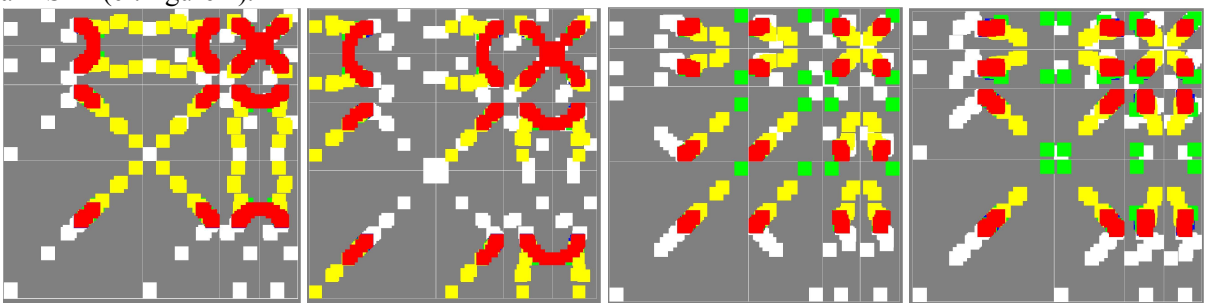


Figure 2 MSPE des quatre dissimilarités $w_e^{(1)}, w_e^{(2)}, w_\infty^{(1)}, w_\infty^{(2)}$.

5 Analyse et conclusion

Seules les 4 composantes correspondantes aux valeurs propres de plus grandes amplitudes ont été sélectionnées. Leurs signes sont (+,-,+,-). Ces 4 composantes expliquent entre 70% et 80% des composantes positives et négatives du nuage de points. On peut remarquer sur la figure 2 que les deux dernières dimensions sont quasiment soit identiques, soit symétriques. De plus, les valeurs propres associées sont similaires en amplitude et de signe opposé. Par conséquent, ces deux dimensions s'annulent

presque totalement et leur contribution à la dissimilarité peut donc être négligée. De plus, bien que les valeurs propres des composantes ne soient pas les mêmes, les structures de l'espace induit par $w_e^{(1)}$ et $w_\infty^{(1)}$ sont similaires à celles induites par $w_e^{(2)}$ et $w_\infty^{(2)}$. Par conséquent, le surplus en temps de calcul des dissimilarités $w_e^{(2)}$ et $w_\infty^{(2)}$ n'apporte pas de réel différence. Si on considère les MSPE non plus remplis avec les couleurs des classes (i.e. les personnes) mais plutôt avec les images des visages associés, on remarque que le premier axe principal effectue la séparation des images en fonction de la luminosité globale de l'image (du plus clair au plus foncé). Le coude correspond parfaitement avec la séparation effective entre l'ensemble des images plutôt claires et l'ensemble des images plutôt foncées. On note également que l'ensemble des visages d'une personne représente une zone bien localisée de l'espace sauf pour $w_\infty^{(1)}$ pour laquelle un point aberrant apparaît (un des visages de la personne est totalement détachés des autres visages de la personne). Ceci nous fait penser que $w_\infty^{(1)}$ et $w_\infty^{(2)}$ sont de moins bonnes dissimilarités que $w_e^{(1)}$, $w_e^{(2)}$. La dissimilarité qui semble la plus viable est $w_e^{(1)}$ car elle est rapide à calculer et que la structure de l'espace induit est correct bien que l'on puisse regretter que les classes ne soient pas plus séparées mais cela est certainement dû au critère de maximum de vraisemblance. Comme nous venons de la voir la MSPE est une méthode viable pour la comparaison et l'analyse de dissimilarités.

6 Bibliographie

[CAM 02] CAMIZ S., *Contribution, à partir d'exemples d'application, à la méthodologie en analyse des données*, Thèse doctorale, Paris, Université Paris-IX Dauphine, CEREMADE, 2002

[CAP 01] CAPPE O., *Ten years of HMMs*, www.tsi.enst.fr/~cappe/docs/hmmbib.html, 2001

[CAR 99] CARD S. K., MACKINLAY J. D., SHNEIDERMAN B., *Readings in Information Visualization: using vision to think*, Morgan Kaufmann Publishers, 1999

[COX 01] COX T. F., COX M. A. A., *Multidimensional scaling*, second edition, Chapman & Hall, 2001

[FAL 95] FALKHAUSEN M., REINIGER H., WOLF D., *Calculation of distance measures between hidden Markov models*, in Processing of the Eurospeech'95, 1995

[ONG 04] ONG C. S., MARY X., CANU S., SMOLA A. J., *Learning with non-positive kernels*, Proceedings of the 21st International conference on machine learning, 2004

[PEK 01] PEKALSKA E., PACLIK P., DUIN R P., *A generalized kernel approach to dissimilarity-based classification*, Journal of machine learning research, 2:175-211, 2001

[RAB 89] RABINER L. R., *A tutorial on hidden Markov models and selected applications in speech recognition*, Proceedings of IEEE, pages 257-286, 1989

[SAM 94] SAMARIA F. S., HARTER A., *Parameterization of a stochastic model for human face identification*, 2nd IEEE Workshop on application of computer vision, 1994

[SCH 99] SCHOLKOPF B., SMOLA A. J., MULLER K.-R., *Advances in kernel methods – support vector learning*, pages 327-352, MIT Press, Cambridge, MA, 1999

[SLI 96] SLIMANE M., VENTURINI G. ASSELIN DE BEAUVILLE J. P., BROUARD T., BRANDEAU A., *Optimizing hidden Markov models with a genetic algorithm*, LNCS, Springer, 1063 :384-396, 1996

[SPE 01] SPENCE R., *Information Visualisation*, Addison-Wesley, ACM Press, 2000

[VIH 02] VIHOLA M., HARJU M., SALMELA P., SUONTAUSTA J., SAVELA J., *Two dissimilarity measures for hmms and their application in phoneme model clustering*, ICASSP'02, 993-936, 2002

Classification de données par automate cellulaire

H. Azzag[*], F. Picarougne[*], C. Guinot[**], G. Venturini[*]

[*]*Université François-Rabelais de Tours, Laboratoire d'Informatique (EA 2101),*
64, Avenue Jean Portalis, 37200 Tours, France
{hanene.azzag, fabien.picarougne,venturini}@univ-tours.fr

[**]*CE.R.I.E.S., 20, rue Victor Noir, 92521 Neuilly-sur-Seine Cédex, France*
christiane.guinot@ceries-lab.com

RÉSUMÉ. Nous présentons dans cet article un premier algorithme utilisant un automate cellulaire pour résoudre un problème de classification. Nous commençons par faire des rappels sur les concepts constituant un automate cellulaire. Nous montrons ensuite comment ces concepts peuvent être appliqués à la classification : les cellules réparties sur une grille 2D peuvent soit être vides soit contenir une donnée. La fonction locale de transition des cellules favorise la constitution de regroupement d'états (données) similaires pour des cellules voisines. Nous présentons ensuite les résultats visuels de notre méthode sur des données classiques.

MOTS-CLÉS : Automates cellulaires, Classification non supervisée, Méthodes biomimétiques

1 Introduction

Parmi toutes les méthodes et problématiques liées au domaine de la classification [JAI 1999], des chercheurs s'intéressent plus spécialement aux méthodes inspirées de systèmes ou de phénomènes biologiques. Nous avons présenté récemment un survol de ce type de méthodes dans [AZZ 2004]. Une des conclusions de ce survol était le fait qu'à notre connaissance, aucun algorithme de classification utilisant les automates cellulaires n'a été défini à ce jour. Pourtant le modèle des automates cellulaires est connu depuis longtemps [NEU 1966] et possède de nombreuses propriétés intéressantes comme celles que l'on retrouve notamment dans le célèbre "jeu de la vie" [GAR 1970]: l'émergence de comportements complexes à partir de règles locales plus simples. Nous allons donc montrer dans la suite de cet article que ce modèle, utilisé dans de nombreux domaines [GAN 2003], peut apporter sa contribution au problème de la classification, comme cela a déjà été exposé pour les algorithmes génétiques, les algorithmes à base de fourmis artificielles ou encore les réseaux immunitaires artificiels (voir un survol dans [AZZ 2004]).
La section 2 décrit succinctement les propriétés des automates cellulaires. La section 3 présente notre algorithme et les différents choix que nous avons du effectuer. La section 4 présente des résultats expérimentaux sur des jeux de données classiques ainsi que les conclusions et perspectives liées à ce travail.

2 Principes des automates cellulaires

Nous rappelons quelques principes des automates cellulaires (AC par la suite) [GAN 2003] que nous avons repris pour définir notre algorithme. Un AC est défini par la donnée d'un quadruplet (C, S, V, δ). $C=\{c_1, ..., c_{NCell}\}$ représente un ensemble de cellules où $NCell$ est constant au cours du temps. $S=\{s_1, ...,s_k\}$ est l'ensemble fini d'états que va pouvoir prendre chaque cellule. L'état de la cellule c_i est noté $c_i(t)$. V

représente le voisinage entre cellules qui va structurer l'ensemble des cellules. Pour chaque cellule c_i, on définit $V(c_i)$ comme l'ensemble des cellules voisines de c_i. Nous allons nous intéresser dans ce travail à une structuration en 2D des cellules qui sont placées sur une matrice ou grille de dimension $N \times N$ (le nombre de cellules vaut donc $NCell = N^2$). Chaque cellule possède un voisinage carré de coté v centré sur elle-même. Ce voisinage est tel que la grille est toroïdale (le haut est relié au bas, le coté droit au coté gauche). Une cellule a donc toujours un voisinage de v^2 cellules. La fonction de transition locale δ détermine le nouvel état d'une cellule en fonction des états perçus. Enfin, on appelle configuration de l'AC à l'instant t le vecteur d'états $AC(t) = (c_1(t), ..., c_{NCell}(t))$. Un AC évolue de $AC(t)$ à $AC(t + 1)$ en appliquant δ à chacune des cellules en parallèle.

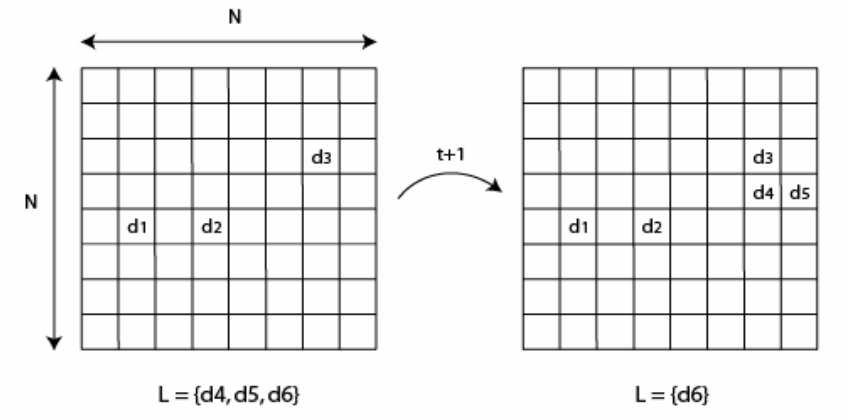


Figure 1 – Représentation de notre automate cellulaire 2D et de la liste *L* d'états

3 Modélisation pour la classification et description de l'algorithme

Nous notons dans la suite par $d_1, ..., d_n$ les n données à classer et par $Sim(i, j) \in [0, 1]$ la mesure de similarité entre deux données d_i et d_j. Nous avons considéré un automate 2D où les $NCell$ cellules sont réparties sur une grille carrée (voir figure 1). La motivation ici est d'obtenir des résultats visuels permettant à l'utilisateur d'explorer directement la classification et permettant également dans le futur de visualiser d'autres informations comme des images représentant chaque donnée.

L'ensemble des états des cellules est $S = \{vide, d_1, ..., d_n\}$. Autrement dit, chaque cellule sera vide ou bien contiendra une (et une seule) donnée. À chaque itération de l'algorithme, les états de toutes les cellules vont être (éventuellement) modifiés selon des règles locales qui vont tendre à faire apparaître des états (données) similaires pour des cellules voisines sur la grille. La taille de la grille est fixée empiriquement [LUM 1994] en fonction de n avec la formule $N = \sqrt{2n}$ afin de laisser de la place ($2n$ cellules au lieu de n) pour l'organisation spatiale des classes. La taille v du carré définissant le voisinage a été fixée également empiriquement à $v = \sqrt{2n}/10$.

Nous allons utiliser les notations/définitions suivantes : une cellule est isolée si son voisinage immédiat comporte moins de 3 cellules non vides. Nous avons choisi d'obtenir des classifications non recouvrantes : un état d_i donné ne pourra apparaître qu'une seule fois dans la grille. Nous utilisons donc une liste L qui représente la liste des données qui n'apparaissent pas sur la grille (et qui restent à placer). Initialement, L contient toutes les données et les états de toutes les cellules sont à "*vide*".

Les règles locales de changement d'état sont les suivantes, pour une cellule C_{ij} vide :

- R_1 Si C_{ij} est isolée, Alors $C_{ij}(t+1) \leftarrow d_k$ où d_k est une donnée choisie aléatoirement dans L
- R_2 Si C_{ij} est non isolée, Alors $C_{ij}(t+1) \leftarrow d_k$ où d_k est soit une donnée choisie aléatoirement dans L (probabilité $P = 0.01$), soit la donnée de L la plus similaire à celle du voisinage de C_{ij} (probabilité $1 - P = 0.99$)

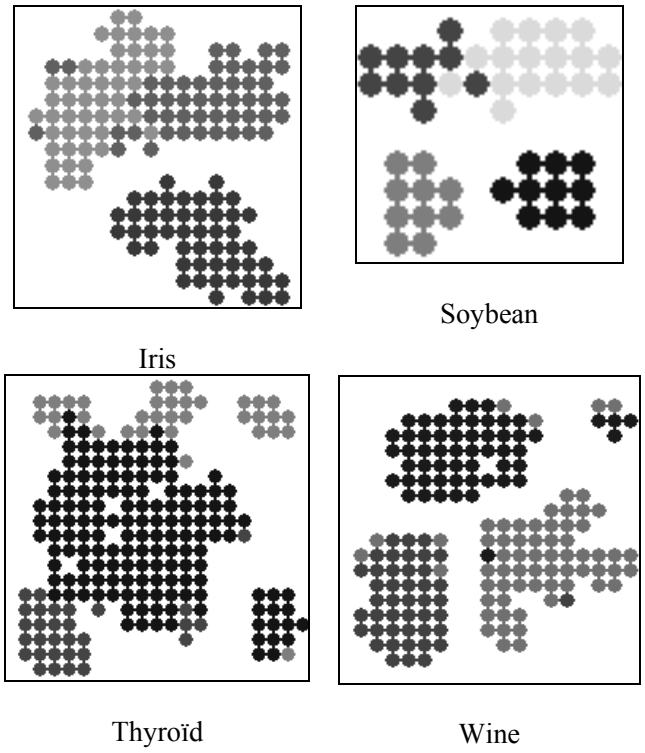


Figure 2 – Résultats visuels obtenus (les couleurs indiquent les classes réelles)

Pour une cellule C_{ij} contenant une donnée d_k (i.e. $C_{ij}(t) = d_k$) :

- R_3 Si $\overline{Sim}_{d_{k'} \in V(C_{ij})}(d_k, d_{k'}) < Seuil(t)$, alors $C_{ij}(t+1) \leftarrow$ *vide* et d_k est remise dans L
- R_4 Si C_{ij} est isolée Alors $C_{ij}(t+1) \leftarrow$ *vide* avec une probabilité $P' = 0,75$ (et d_k est remise dans L).

Dans les autres cas, la cellule reste inchangée ($C_{ij}(t+1) \leftarrow C_{ij}(t)$).

Pour appliquer ces règles sur les cellules et éviter les conflits d'affectation des données présentes dans L, nous avons testé plusieurs ordres de parcours de la grille afin de décider quelles cellules accèdent à la liste en premier. L'ordre que nous avons sélectionné est de parcourir aléatoirement les cellules (une permutation des N^2 cellules est générée aléatoirement au début de l'algorithme).
La valeur de *Seuil(t)* est initialisée à la similarité maximum entre les données, puis va décroître progressivement. Initialement, les données placées côte à côte seront donc très similaires. À chaque itération de l'algorithme, ce seuil est décrémenté d'un pas constant (égal à un millième de l'écart type observé dans les similarités). La diminution de ce seuil fait que l'algorithme va converger puisque les données une fois mises en place ne bougeront plus lorsque *Seuil (t)* sera faible.

4 Résultats et conclusion

Nous avons appliqué notre algorithme sur des bases de données classiques issues du *Machine Learning Repository* [BLA 1998]. Nous avons utilisé le même jeu de paramètre pour toutes les bases (voir section précédente). Les résultats présentés sur la figure 2 illustrent les classifications trouvées pour ces données. Nous remarquons que la disposition des classes correspond aux propriétés connues des bases, comme par

exemple pour les bases Iris et Wine. Les temps de convergence de l'automate sont inférieurs à 1s (en Java sur PC P4 à 2.6 GHz). Une analyse des résultats en terme de pureté et de nombre de classes est en cours.
En conclusion, nous avons proposé un premier algorithme de classification utilisant les automates cellulaires. Nous avons montré expérimentalement que cet algorithme est capable de regrouper de manière pertinente des données de bases classiques. Il est de plus capable de produire une visualisation des résultats et peut contribuer ainsi à la problématique de la fouille visuelle. Nous avons effectué une spécialisation du modèle général présenté dans la section 2 : la fonction locale de transition utilise une information globale (la liste *L* des données disponibles ainsi que le seuil *Seuil(t)*). Nous souhaitons relâcher cette contrainte en autorisant des données à apparaître plusieurs fois sur la grille (classification recouvrante) et en rendant le seuil de similarité local à chaque donnée (ce qui permettra à l'algorithme d'être incrémental). Cet algorithme peut être comparé à un algorithme utilisant les fourmis artificielles [LUM 1994]. Les points communs viennent du fait que les deux méthodes proposent en sortie une classification des données sur une grille 2D en utilisant des probabilités locales. Cependant, les heuristiques locales sont différentes. Une comparaison avec ce type d'algorithme est prévue ainsi que d'étendre notre méthode pour la fouille visuelle des données. En particulier, nous envisageons d'utiliser des techniques de zoom pour visualiser la grille et considérer que chaque donnée peut être une image.

5 Remerciement

Nous tenons à remercier David Ratsimba pour son aide dans l'implémentation de cet algorithme.

6 Références

[AZZ 2004] AZZAG H., PICAROUGNE F., GUINOT C., VENTURINI G., *Un survol des algorithmes biomimétiques pour la classification*. Classification Et Fouille de Donnée, pages 13-24, RNTI-C-1, Cépaduès. 2004.

[BLA 1998] BLAKE C.L., MERZ C.J., *UCI Repository of machine learning databases*. http://www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, CA: University of California, Department of Information and Computer Science. 1998.

[GAN 2003] GANGULY N., SIKDAR B. K., DEUTSCH A., CANRIGHT G., CHAUDHURI P. P., *A Survey on Cellular Automata*. Technical Report Centre for High Performance Computing, Dresden University of Technology, December 2003.

[GAR 1970] GARDNER M., *Mathematical Games: The fantastic combinations of John Conway's new solitaire game 'life'*. Scientific American, pages. 120-123, Octobre 1970.

[JAI 1999] JAIN A. K., MURTY M. N., FLYNN P. J., *Data clustering: a review*, ACM Computing Surveys, 31(3), pages. 264-323, 1999.

[LUM 94] LUMER E., FAIETA B., *Diversity and adaption in populations of clustering ants.* In Proceedings of the Third International Conference on Simulation of Adaptive Behaviour: From Animals to Animats 3, pages 501-508. MIT Press, Cambridge, MA, 1994.

[NEU 1966] VON NEUMANN J., *Theory of Self Reproducing Automata.*, University of Illinois Press, Urbana Champaign, Illinois, 1966.

Classification non supervisée hiérarchique incrémentale basée sur le calcul de dissimilarités

Eugen Barbu – Pierre Héroux – Eric Trupin

Laboratoire Perception Systèmes Information
CNRS FRE 2645 – Université de Rouen
UFR des Sciences et Techniques
Place E. Blondel
76 821 Mont-Saint-Aignan cedex - France

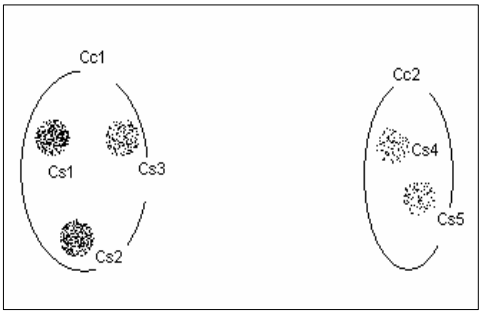
RÉSUMÉ. Cet article présente un algorithme incrémental de classification non supervisée. Cet algorithme se base uniquement sur la connaissance des dissimilarités entre les objets pris deux à deux. Une structure hiérarchique est construite puis mise à jour pour découvrir la structure cachée des données. Cette approche s'inspire des systèmes de classification conceptuelle non supervisée comme COBWEB ou HIERARCH... La recherche de plus proches voisins est optimisée en utilisant une structure de type M-tree.

MOTS-CLES : Classification non supervisée incrémentale, Dissimilarités.

1 Introduction

Etant donné un ensemble d'objets décrits par un nombre fixe d'attributs, l'objectif d'une tâche de classification non supervisée [KAU 90], [GOR 99] consiste à proposer une partition des objets en *k* sous-ensembles où le paramètre *k* est le nombre de regroupements attendus par l'utilisateur. Une variation de cette tâche est de ne pas utiliser le nombre attendu de regroupements comme une donnée du problème. Dans ce cas, l'algorithme construit plusieurs partitions candidates et choisit la meilleure. La meilleure partition est celle qui optimise un critère de qualité des partitions [MIL 85], [ROU 87] (statistiques sur les distances entre regroupements, indice de Calinsky-Harabasz, C-index, Silhouette Index, indice de Duda-Hart...). Plusieurs stratégies permettant la recherche des regroupements dans l'espace de toutes les partitions. On distingue les méthodes procédant par partitionnement, les méthodes hiérachiques (ascendantes ou procédant par divisions successives), les méthodes basées sur les densités et les méthodes de quantification. Dans certaines applications, les données à regrouper ne sont pas représentées par des vecteurs d'attributs. Elles peuvent par exemple être représentées par des graphes, des listes d'attributs de longueur variable, des mots ou sac de mots, des images... La seule information accessible est alors une mesure de dissimilarité entre objets. Plusieurs algorithmes de classification non supervisée [KAU 90]peuvent être appliqués à partir de la matrice contenant les mesures de dissimilarité entre les objets pris deux à deux. Les données peuvent être assignées à des regroupements, ces derniers pouvant alors être associés jusqu'à obtenir une hiérarchie de partitions (Fig. 1). La présentation hiérarchique des données se révèle souvent utile car la relation de catégorie à sous-catégorie existe dans bon nombre d'applications. Par exemple, dans une application d'analyse d'image de document, chaque caractère représente un regroupement de pixels, mais ces caractères peuvent eux-mêmes être regroupés en mots, lignes et paragraphes. Par ailleurs, certains regroupements de pixels correspondent non pas des données textuelles mais à des parties graphiques (schéma, images). Cette idée illustre les faiblesses des méthodes ne

proposant pas une hiérarchie des données mais un partitionnement à un seul niveau. Cependant, le problème majeur des algorithmes de classification hiérarchique est la complexité spatiale et temporelle (quadratique).
Dans les applications recevant en entrée un flux de données, comme c'est le cas pour les systèmes d'analyse de documents, log analysis ou les market-basket analysis systems, le caractère incrémental des algorithmes est primordial. Les algorithmes hiérarchiques de classification non supervisée ont été étudiés dans le domaine de l'apprentissage automatique et plus précisément dans le cadre de la formation de concepts dont les systèmes COBWEB, Classit et HIERARCH sont des mises en oeuvre. Ces systèmes utilisent en entrée des descriptions à base de vecteurs d'attributs, ce qui n'est pas sans répercussion sur la fonction objectif utilisée (par exemple le critère d'utilité d'une partition) pour construire la hiérarchie.
Dans la suite de cet article, nous décrivons une méthode de mise à jour incrémentale d'une hiérarchie (taxonomie) d'objets basée sur une mesure de dissimilarité entre objets. Lorsque la seule information disponible concernant objets à regrouper est leur dissimilarité réciproque, il n'est pas possible de déterminer le centre d'un regroupement. La mesure de qualité du partitionnement doit en conséquence se dispenser du calcul du centre du regroupement.



Comme beaucoup d'autres, notre méthode utilise une recherche de plus proche voisin. Nous l'optimisons en utilisant une méthode d'indexation basée sur la structure M-tree [CIA 97].
La section 2 décrit les opérations appliquées à la hiérarchie à l'arrivée de nouvelles données. L'algorithme proposé est donné en section 3. Cette section présente également la fonction

Figure 1 Catégories et sous-catégories de données bidimensionnelles

objectif utilisée et les transformations appliquées aux données d'entrée pour permettre l'utilisation de la méthode d'indexation M-tree. Un exemple d'application de notre algorithme est donné en section 4. La section 5 conclue cet article en pointant les faiblesses de notre approche et en proposant un certain nombre de perspectives.

2 Gestion de la hiérarchie

Nous distinguons les regroupements dits de base (C_{s1} à C_{s5} de la figure 1) des regroupements composites (C_{c1} et C_{c2}). Les regroupements composites contiennent au minimum deux regroupements (simples ou composites). Lorsqu'un nouvel élément est présenté deux évènements peuvent survenir : l'élément est assigné à un regroupement existant ou un nouveau regroupement est créé. Dans chaque cas de figure la hiérarchie est mise à jour. La figure 2 présente les modifications pouvant affecter la hiérarchie de partitions lors de l'ajout d'un nouvel élément. En plus de l'assignation de l'élément à un regroupement existant et de la création d'un nouveau regroupement, nous envisageons deux autres scénarios à l'arrivée d'un élément : deux regroupements peuvent être fusionnés ou, au

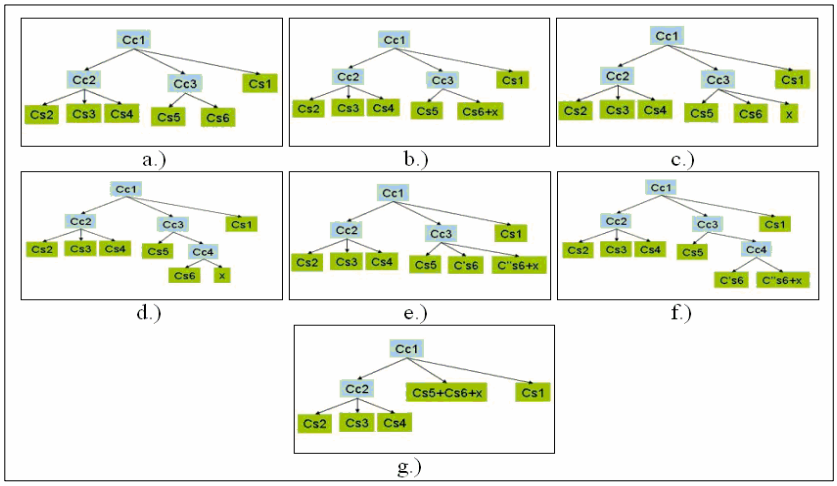


Figure 2 Mise à jour de la hiérarchie à l'ajout d'une nouvelle donnée (objet x)

contraire, un regroupement existant peut être divisé.

3 Algorithme

Add(Object obj)

```
PartitionHierarchy p; // initial partition
int hierarchyLevel;
Cluster cs1 = findNearestSimpleCluster(p, obj, &hierarchyLevel);
Cluster cs2 = findSecondNearestSimpleClusterAtHierarchyLevel(p, obj, hierarchyLevel);
if(isAddableTo(cs1))
{
        cs1'=cs1+obj;
  if(isSplitable(cs1'))
  {
            (cs1a,cs1b)=split(cs1');

        if(isBetterToCreateNewCategory(p,cs1a,cs1b))
            {
                cc=
                createNew
                ComplexCluster(cs1a,cs1b);
                add(p,cc,cs1.father);
                delete(p,cs1);
                //Fig. 2.f
            }
            else
            {
                add(p,cs1a,cs1.father);
                add(p,cs1b,cs1.father);
                delete(p,cs1);
                //Fig. 2.e
            }
  }
  else
  {
            if(isBetterToUnify(p,cs1',cs2))
            {
                cs = unify(p,cs1',cs2);
                delete(p,cs1);
                //Fig. 2.g
            }
            else
            {
                add(p,cs1',cs1.father);
                delete(p,cs1);
                //Fig. 2.b
            }
    }
}
else
{
    Cluster cs3 = new Cluster(obj);
    if(isBetterToCreateNewCategory(p,cs1,cs3))
    {
        cc = createNewComplexCluster(cs1,cs3);
        add(p,cc,cs1.father);
        delete(p,cs1);
        //Fig. 2.d
    }
    else
    {
        add(p,cs3,cs1.father);
        //Fig. 2.c
    }
}
```

La méthode M-tree est utilisée afin d'optimiser la recherche de plus proche voisin. Cette méthode, basée sur une structure dynamique (incrémentale), partitionne et organise l'espace de recherche. La mesure de dissimilarité entre éléments doit être une métrique. En effet, l'inégalité triangulaire y est employée pour réduire le nombre d'étapes lors de la résolution d'une requête donnée. Nous appliquons une transformations aux mesures de dissimilarité entre objets de respectant pas cette contrainte. Si *d(a,b)* est une mesure de dissimilarité alors la quantité $D(a,b)=\frac{d(a,b)}{1+d(a,b)}+1$ respecte la même relation d'ordre que *d*, ainsi que l'inégalité triangulaire.

$$D(a,b)+D(a,c)\geq D(b,c)\Leftrightarrow \frac{d(a,b)}{1+d(a,b)}+1+\frac{d(a,c)}{1+d(a,c)}+1\geq \frac{d(b,c)}{1+d(b,b)}+1\Leftrightarrow \frac{d(a,b)}{1+d(a,b)}+\frac{d(a,c)}{1+d(a,c)}+1\geq \frac{d(b,c)}{1+d(b,c)}$$

mais $\frac{d(a,b)}{1+d(a,b)}+\frac{d(a,c)}{1+d(a,c)}+1\geq 1>\frac{d(b,c)}{1+d(b,c)}$ .Aussi $D(a,b)\geq D(a,c)\Leftrightarrow d(a,b)\geq d(a,c)$ utilisant le propriétés de la fonction $\frac{x}{1+x}$. La propriété *d(a,a)=0* n'est plus respectée par *D*, mais cela n'empêche pas l'utilisation de la méthode M-tree. La transformation *D* permet donc d'obtenir une nouvelle base de dissimilarité dans laquelle est effectuée la recherche de plus proche voisin. Notre implémentation des fonctions permettant de décider de l'action à effectuer (création d'un nouveau regroupement, fusion de deux regroupements, division d'un regroupement) est actuellement basée sur la maximisation de la valeur moyenne du Silhouette-index [ROU 87]. Pour chaque objet u d'un regroupement A, nous définissons les valeurs suivantes : $a(u)=\frac{1}{|A|-1}\sum_{v\in A, v\neq u} d(u,v)$, $d(u,C)=\frac{1}{|C|}\sum_{v\in C} d(u,v)$, $b(u)=\min_{C\neq A} d(u,C)$ Le silhouette-

index est alors défini par : $s(u)=\frac{b(u)-a(u)}{\max\{a(u),b(u)\}}$. Si s*(u)* est proche de *1.0*, alors le rattachement de *u* au regroupement *A* est justifié. Si *s(u)* est proche de *0*, alors *u* se situe entre deux regroupement. Enfin, si s(u) est proche de –1.0, le rattachement de u au regroupement A n'est pas justifié, il devrait être rapproché d'un autre regroupement..

4 Exemple didactique

Nous employons la F-mesure globale afin d'évaluer la qualité du partitionnement effectué par notre algorithme de classification non supervisée. Soient *D* l'ensemble des objets et *C = {C1, . . . , Ck}* une partition de *D*. Soit par ailleurs, *C' = {C'1 , . . . , C'l }* la partition de référence. Le rappel, la précision et la F-mesure du regroupement *j* par rapport à la classe *i* sont respectivement définis par $rec(i,j)=\frac{|C_j\cap C'_i|}{C'_i}$, $prec(i,j)=\frac{|C_j\cap C'_i|}{C_j}$ et $F_{ij}=\frac{2prec(i,j)rec(i,j)}{prec(i,j)+rec(i,j)}$. La F-mesure globale d'une partition définie par rapport à une partition de référence est définie par $F=\sum_{i=1}^{l}\frac{|C'_i|}{|D|}*\max\{F_{ij}\}_{j=1\ldots k}$. Cette mesure vaut *1.0* lorsqu'il y a correspondance parfaite entre *C* et *C'*.

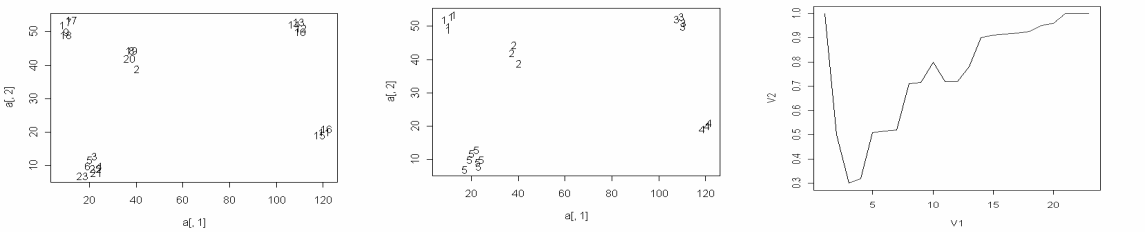


Figure 3. a) Ordre d'introduction des données b) Vérité-terrain c) Evolution de la F-Mesure lors de l'introduction des données

5 Conclusion et perspectives

Dans cet article, nous avons proposé une construction incrémentale d'une classification hiérarchique n'utilisant que des dissimilarités entre les éléments. Nous envisageons de tester l'approche décrite dans cet article sur des données synthétiques et réelles. D'autres fonctions objectif devront également être implémentées et testées. L'étude de la distance entre la sortie de l'algorithme de classification non supervisée et la partition de référence devra être approfondie.

6 Bibliographie

[CIA 97] CIACCIA C., PATELLA M., ZEZULA P., " M-tree an Efficient Access Method for Similarity Search in Metric Spaces ", In Proc. of the 23th Conference on Very Large Databases, p. 426-435, 1997.

[FIS 87] FISCHER D., " Knowledge acquisition via Incremental Conceptual Clustering ", ML, n° 2, 1987

[GOR 99] GORDON A. D., *Classification 2nd Edition*, Chapman & Hall, 1999.

[KAU 90] KAUFMAN L., ROUSSEEUW P. J., *Finding Groups in Data*, John Willey & Sons, 1990.

[MIL 85] MILLIGAN G. W., COOPER M. C., " An Examination of Procedures for Determining the Number of Clusters in a Data Set ", *Psychometrika*, vol. 58, n° 2, 1985, p. 159-179.

[NEV 95] NEVINS A. J., " A Branch and Bound Incremental Conceptual Clusterer ", ML , n° 18, 1995.

[ROU 87] ROUSSEEUW P. J., " Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis ", *J. Comput. Applied. Math.*, n° 20, 1987, p. 53-65.

STABILITE DES METHODES DE CLASSIFICATION HIERARCHIQUES : APPROCHES QUALITATIVES.

Jean-Pierre Barthélemy

Département Logique des Usages, Sciences Sociales et de l'Information,
Ecole Nationale Supérieure des Télécommunications de Bretagne,
Technopôle de Brest-Iroise, CS 83818
29238 Brest cedex France
et
Centre d'Analyse et de Mathématiques Sociales
U.M.R. CNRS 8557 (EHESS)
54, Boulevard Raspail, 75270 Paris cedex 06, France.

RÉSUMÉ. Cette contribution traite de la stabilité/instabilité des méthodes de classification hiérarchique, dans un cadre qualitatif (les données de départ ne sont pas des dissimilarités). Elle est fondée sur deux notions de restrictions d'une hiérarchie à un sous-ensemble d'objets. La première partie discute des propriétés de ces notions de restriction. La seconde présente dix axiomes d'Icelles issus et met en place quelques résultats de possibilités/impossibilité.

MOTS-CLÉS : Classifications hiérarchiques, Méthodes de classification hiérarchique, stabilité par restriction et par enlèvement, théorèmes de possibilité ou d'impossibilité.

1 Introduction

Il y a bien longtemps, Simon Règnier constatait que, dans le cas de la recherche d'une partition sur un ensemble S d'objets décrits par des caractéristiques (étude republiée ensuite dans Mathématiques, Informatique et Sciences Humaines [REG, 1983]) aucune méthode n'était stable. Le but de cette contribution est de généraliser les intuitions de Règnier aux classifications hiérarchiques, dans une optique qualitative. Par « qualitatif », nous entendons que les données sont un ensemble de sous-ensembles de S (c'est le notamment lorsque les éléments de S sont décrits par des caractéristiques), vérifiant des condititions de normalisation qui leur font mériter le nom de *système de classe*. Le modèle est, quant à lui, une hiérarchie de parties sur S. Une première partie est consacrée à quelques notions de base. On y étudie, en particulier deux variations sur le thème de la restriction d'un système de classes à un sous ensemble d'objets. Dans la seconde partie, on décrit dix axiomes portant sur la stabilité. Finalement, on énonce un théorème d'impossibilité.

2 Prolégomènes.

2.1 Systèmes de classes

Nous notons par S_n l'ensemble{1, 2, …, n}. Un n-*système de classes* (ou n-CS, ou, plus simplement CS) est un ensemble **K** de sous-ensembles non vides de S_n, qui contient S_n ainsi que tous les singletons {i}, ($1 \leq i \leq n$). Les éléments d'un CS sont appelés ses *classes*, les classes S_n et les singletons sont dites *triviales*.

Une n-*hiérarchie* est un n-CS **H** tel que si A et B sont des classes non disjointes, A est inclus dans B ou B est inclus dans A. Les classes triviales d'un n-SC constituent une n-hiérarchie notée $\mathbf{H}_{\emptyset}^{n}$ (ou simplement $\mathbf{H}_{\emptyset}$).

On note $\mathcal{K}_n$ l'ensemble de tous les n-CS et par On note $\mathcal{H}_n$ l'ensemble de toutes les n-hiérarchies.

2.2 Descriptions locales d'un Système de classes.

Soit **K** un n-CS et X un sous-ensemble de S_n, la *restriction* de **K** à X est le CS $\mathbf{K}_{|X} = \{X \cap C: C \text{ est une classe non triviale de } \mathbf{K} \text{ et } C \cap X \neq \emptyset\} \cup \mathbf{H}_{\emptyset}$.

Pour X avec $1 < |X| < n$, l'*enlèvement* de X de **K** est le CS CS $\mathbf{K}/X = \mathbf{K}_{|X}/\{X\}$.

Remarquons que $\mathbf{K}/X = \mathbf{K}_{|X}$ si et seulement si X n'est inclus dans aucune classe non triviale de **K**. Remarquons aussi qu' une restriction ou un enlèvement d'une hiérarchie reste une hiérarchie.

Les notions de restriction et d'enlèvement peuvent être utilisées pour exprimer que deux CS sont « localement identiques ». Une conséquence de la définition de $\mathbf{K}_{|X}$, est que X peut, ou non, en être une classe. C'est ainsi que pour $p < n$, l'ensemble $\mathcal{K}_p$ n'est pas l'exacte copie de la collection des $\mathbf{K}_{|X}$, lorsque **K** parcourt $\mathbf{K}_n$ et X est un sous-ensemble donné, à p éléments, de S_n. La notion d'enlèvement apporte un remède : X n'est jamais une classe. Une possibilité alternative aurait été d'imposer que X soit toujours une classe. On serait alors amené à considérer une notion de *restriction forte* : $\mathbf{K}_{||X} = \mathbf{K}_{|X} \cup \{X\}$. Le lecteur vérifiera facilement que tout résultat mobilisant la notion d'enlèvement peut être traduit en terme de restriction forte.

Signalons, pour en finir avec ces définitions, que la notion d'enlèvement a été introduite et utilisée en théorie du consensus ([BAR 92], [BAR, 95], [DWY, 99], [POW, 00]) .

Lemme 1. Soient deux CS **K** et **K'**.

(i) Si $\mathbf{K}_{|X} \subseteq \mathbf{K'}_{|X}$, pour tout $X \subset S_n$, on a $\mathbf{K} \subseteq \mathbf{K'}$.

(ii) Si $\mathbf{K}/X \subseteq \mathbf{K'}/X$, pour tout $X \subsetneq S_n$, et si A est une classe non triviale de **K**, avec $|A| < n-1$, alors A est une classe de **K'**.

(iii) Supposons $n > 3$. Si $\mathbf{K}/X \subseteq \mathbf{K'}/X$, pour tout $X \subsetneq S_n$, et si **K'** est une hiérarchie, **K** est également une hiérarchie et $\mathbf{K} \subseteq \mathbf{K'}$.

(iv) Si $\mathbf{K}_{|X} = \mathbf{K'}_{|X}$, pour tout $X \subset S_n$, alors $\mathbf{K} = \mathbf{K'}$.

Notons que pour n = 3, on $\mathbf{K}/X = \mathbf{H}_{\emptyset}^{3}$, pour tout $X \subsetneq S_3$. Donc : $\mathbf{K}/X = \mathbf{K'}/X$.

Colonius and Schulze [COL, 81] ont montré qu'une hiérarchie sur S_n est caractérisée par une relation ternaire issue de ses restrictions aux sous-ensembles à trois éléments. En particulier, deux hiérarchies sont identiques si et seulement si leurs restrictions à tous les sous-ensembles à trois éléments le sont. Nous reprenons ci-dessous ce résultat et le complètons en utilisant la notion d'enlèvement .

Proposition 1. Supposons $n > 3$. Soient **H** et **H'** deux n-hiérarchies. Les trois conditions ci-dessous sont équivalentes :

(i) $\mathbf{H} = \mathbf{H'}$.

(ii) Pour tout $X \subsetneq S_n$, avec $|X| = 3$, $\mathbf{H}/X = \mathbf{H'}/X$.

(iii) Pour tout $X \subsetneq S_n$, avec $|X| = 3$, $\mathbf{H}_{|X} = \mathbf{H'}_{|X}$.

3 Sur la stabilité des méthodes de classification hiérarchiques.

3.1 Méthodes de classification hiérarchique.

Une n-*Méthode de Classification Hiérarchique* (n-MCH, ou plus simplement MCH) est une application c de $\mathcal{K}_n$ dans $\mathcal{H}_n$ telle que pour toute hiérarchie **H**, on ait : c(**H**) = **H**. Nous disons que c(**K**) est une c-*solution* de **K**.

3.2 Stabilités.

De manière informelle, on dit qu'une MCH c est stable si lorsque deux SC sont localement identiques, leur c-solutions le sont aussi. Ceci se laisse énoncer de diverses manières peu ou prou inspirées de [BAR 92]. Les conditions S1, S2, S3, S4, S5, S6, S8 et S9 ci-dessous combinent les restrictions et les enlèvements. Les conditions S7 et S10 mobilisent les classes. Les conditions S1, S2, S3 et S4 n'impliquent qu'un seul SC.

S1: $c(\mathbf{K}_{|X}) = c(\mathbf{K})_{|X}$.
S2: $c(\mathbf{K}_{|X}) = c(\mathbf{K})/X$.
S3: $c(\mathbf{K}/X) = c(\mathbf{K})/X$.
S4: $c(\mathbf{K}/X) = c(\mathbf{K})_{|X}$.

S5: $\mathbf{K}_{|X} = \mathbf{K'}_{|X}$ implique $c(\mathbf{K})_{|X} = c(\mathbf{K'})_{|X}$.
S6: $\mathbf{K}_{|X} = \mathbf{K'}_{|X}$ implique $c(\mathbf{K})/X = c(\mathbf{K'})/X$.
S7: $\mathbf{K}_{|X} = \mathbf{K'}_{|X}$ implique $X \in c(\mathbf{K})$ si et seulement si $X \in c(\mathbf{K'})$.

S8: $\mathbf{K}/X = \mathbf{K'}/X$ implique $c(\mathbf{K})_{|X} = c(\mathbf{K'})_{|X}$.
S9: $\mathbf{K}/X = \mathbf{K'}/X$ implique $c(\mathbf{K})/X = c(\mathbf{K'})/X$.
S10: $\mathbf{K}/X = \mathbf{K'}/X$ implique $X \in c(\mathbf{K})$ si et seulement si $X \in c(\mathbf{K'})$.

3.3 Le résultat.

Proposition 2.
(i) Chaque 2-MCH vérifie les conditions Si, $1 \leq i \leq 10$.
(ii) Pour n >2, il n'existe pas de n-MCH vérifiant S1, S2, S4, S5, S7, S8, S10.
(iii) Toute 3-MCH vérifie S3, S6 et S9.
(iv) Il existe (au moins) une 4-MCH vérifiant S3, S6 et S9, mais ce n'est le cas de toutes.
(v) Pour $n \geq 5$, aucune n-MCH ne vérifie S3, S6, ou S9.

Bibliographie

[BAR 92] BARTHELEMY J.-P., McMORRIS F.R., POWERS R.C. "Dictatorial consensus functions on n-trees", *Mathematical Social Sciences,* vol 25, p. 59-64, 1992.

[BAR 95] BARTHELEMY J.-P., McMORRIS F.R., POWERS R.C. "Stability conditions for consensus functions defined on n-trees", *Mathematical and Computer Modelling*, vol 22, n°1, p. 79-87, 1995.

[COL 81] COLONIUS H., SCHULZE H.H., « Tree structures for proximity data », *British Journal of Mathematical and Statistical Psychology* 34, 167-180, 1981.

[DWY 99] DWYER M., McMORRIS R.R., POWERS R.C., “Removal independence and multi-consensus functions”, *Mathématiques, Informatique et Sciences* Humaines n°148, p. 31-40, 1999.

[POW 00] POWERS R.C., “ Consensus n-trees and removal independence”, *Journal of the Korean Mathematical* Society, vol 37, n°3, p. 473-490, 2000.

[REG 83] REGNIER S., “Stabilité d’un opérateur de classification”, *Mathématiques, Informatique et Sciences* Humaines, 82, 1983.

Identification des rôles sémantiques par la classification

Luc Bélanger et Guy Lapalme

Laboratoire RALI
Département d'informatique et de recherche opérationnelle
Université de Montréal
C.P. 6128, succursale Centre-ville
Montréal (Québec) Canada H3C 3J7
{*belanglu,lapalme*}*@iro.umontreal.ca*

RÉSUMÉ. *Dans cet article, nous montrons qu'il est possible de développer un analyseur sémantique à partir d'un corpus annoté de rôles sémantiques. L'identification d'attributs sur les noeuds d'un arbre de dérivation syntaxique nous permet de considérer le problème comme un problème de classification. L'utilisation d'algorithmes de classification à base de vecteur de support permet d'obtenir un analyseur performant qui peut être utilisé dans plusieurs applications du traitement des langues naturelles.*

MOTS-CLÉS : *structures prédicat-arguments, machine à vecteur de support, rôle sémantique, analyse sémantique*

1. Introduction

L'étiquetage des rôles sémantiques est une tâche qui suscite un intérêt croissant dans le domaine du traitement automatisé des langues naturelles. La disponibilité de corpus annotés sémantiquement tels que le PropBank [KIN 02] et FrameNet [BAK 98] a largement contribué au développement d'analyseurs sémantiques. Nous nous sommes intéressé à ce type d'analyseur sémantique parce qu'ils peuvent être précis tout en ayant une couverture indépendante du domaine à traiter.

Dans cet article, nous allons montrer comment nous avons développé un analyseur sémantique à partir d'un corpus annoté sémantiquement. Le développement de notre identificateur de rôles sémantiques s'inspire grandement des travaux de Gildea et Jurafsky [GIL 02a] et de Pradhan et al. [PRA 04]. L'algorithme de classification utilisé pour l'analyseur sémantique étant une machine à vecteur de support (SVM), nous mettrons l'accent sur la modélisation du problème en un problème de classification.

2. Description de la tâche

Le problème qui nous intéresse est l'identification des structures prédicat-arguments présentes dans une phrase. Une structure prédicat-arguments est un prédicat avec ses arguments étiquetés selon le rôle qu'ils jouent dans la réalisation du prédicat. Ces structures permettent de donner une interprétation sémantique des phrases en identifiant qui a fait quoi à qui, où, quand, comment et pourquoi.

Les données que nous utilisons proviennent du corpus PropBank et du Penn TreeBank, le PropBank étant une couche sémantique ajoutée au Penn TreeBank. Le corpus est composé de deux parties, l'ensemble des frames donnant un sens aux arguments des prédicats et l'annotation des rôles sémantiques sur les arbres de dérivation syntaxique du Penn TreeBank. Chaque frame est composé d'un ensemble de rôles déterminés par les sens que peut avoir le prédicat. Le tableau 1 contient deux exemples de frame.

L'annotation d'une structure prédicat-argument se fait par l'assignation d'une étiquette préfixée par ARG, suivie d'un chiffre entre 0 et 5 ou de la lettre M suivie d'un suffixe dénotant un argument d'adjonction (12 étiquettes secondaires possibles). De façon générale l'argument ARG0 a le rôle du sujet et l'argument ARG1 celui d'objet direct. Le rôle des autres arguments variant d'un verbe à l'autre, il est impossible d'en spécifier le rôle sans utiliser le frame du prédicat. L'interprétation des rôles doit toujours être réalisée par rapport au sens que le prédicat a dans son frame. Le tableau 1 donne la correspondance entre les arguments des prédicats *purchase* et *issue* et leur rôle sémantique tel que défini dans les frames du PropBank.

Rôles	purchase	issue
ARG0	purchaser	issued
ARG1	thing purchased	thing issued
ARG2	seller	issued to
ARG3	price paid	attribute, issued as or at
ARG4	benefactive	-

TAB. 1. *Sémantique des arguments des prédicats* purchase *et* issue*, extraient du PropBank*

L'identification des arguments peut être réalisée de deux manières : en utilisant l'arbre de dérivation syntaxique de la phrase ou en analysant les caractéristiques de surface de la phrase sans utiliser la dérivation syntaxique [CAR 04]. Nous utilisons la dérivation syntaxique car ceci facilite l'extraction des attributs que nous utiliserons pour la classification et donne de meilleurs résultats, en autant que la dérivation syntaxique soit exacte [GIL 02b].

L'identification des arguments à partir de l'arbre de dérivation syntaxique débute par l'identification des prédicats. Nous identifions les prédicats à partir de l'arbre de dérivation syntaxique, un noeud dont l'étiquette est un verbe est identifié comme un prédicat. Pour chaque prédicat nous prenons ensuite tous les noeuds de l'arbre de dérivation syntaxique et nous affectons une étiquette ARG[0-5] ou ARGM aux noeuds qui sont des arguments au prédicat considéré à ce moment.

Nous considérons la détermination des étiquettes comme un problème de classification multi-classes des noeuds de l'arbre d'une dérivation. Puisque plus de 80% des noeuds de la dérivation syntaxique d'une phrase ne sont pas des arguments, l'entraînement d'un seul classificateur multi-classes sur tous les noeuds est inefficace pour deux raisons : la disproportion entre les classes amène un problème de surentraînement et les noeuds non-arguments n'ont pas à être classifiés dans les classes d'arguments.

Le problème d'efficacité lié aux noeuds non-arguments peut être contourné en éliminant ces noeuds par des heuristiques et par classificateur binaire séparant les arguments des non-arguments. Cette approche nous permet donc de diviser le problème en trois étapes :

Étape 1 : Rejet des candidats qui ne sont assurément pas des arguments et extraction des attributs pour chaque candidat restant ;

Étape 2 : Classification des candidats selon qu'ils sont ou non des arguments ;

Étape 3 : Classification multi-classes des arguments selon leur rôle.

3. Modélisation du problème

Pour identifier les arguments, il faut définir un candidat par rapport aux autres, de sorte qu'il puisse être reconnu comme appartenant à une catégorie. Un candidat se définit par un couple $\langle p, a \rangle \in \mathcal{P} \times \mathcal{A}$ où $\mathcal{P}$ est l'ensemble des prédicats et $\mathcal{A}$ est l'ensemble des noeuds d'une dérivation syntaxique. Pour chaque couple $\langle p, a \rangle$, on extrait une représentation $F_{p,a}$ sous la forme d'une liste d'attributs.

Les attributs présentés dans le tableau ci-dessous sont ceux du candidat $\langle purchase, PP \rangle$, couvrant le texte *at 7.0%*, extraits de l'arbre de dérivation syntaxique de la phrase *How can I purchase Bell Canada bonds issued at 7.0% ?*. Ces attributs sont ceux proposés dans la littérature [GIL 02a, PRA 04].

Attributs	$F_{\langle purchase,PP\rangle}$
Prédicat	*purchase*
Type de la phrase	PP
Chemin dans l'arbre du noeud au prédicat	$PP \uparrow VP \uparrow NP \uparrow VP \downarrow VB$
Position du noeud relativement au verbe (avant ou après)	*après*
Voix, le verbe est-il actif ou passif	*actif*
Mot de tête (Head Word) avec la case et la morphologie	*at*
Catégorie gouvernante, seulement si le noeud est un NP	*null*

L'algorithme de classification utilisé pour réaliser ces expériences est C-SVC, implémenté dans les logiciels LIBSVM [CHA 01] et SVM^{light} [JOA 99]. Nous avons utilisé des fonctions de noyau de type RBF et polynomial. Les données utilisées pour l'entraînement des classificateurs proviennent toutes du corpus PropBank. Le corpus est composé de 112 917 annotations couvrant 3 323 verbes et il se divise en 25 parties (wsj-00 à wsj-24). L'extraction des attributs pour l'ensemble du corpus produit 754 000 attributs en réalisant un encodage binaire de ceux-ci.

4. Résultats

4.1. *Évaluation de classificateurs pour l'identification des arguments*

Les expériences réalisées dans le but d'optimiser la première étape de classification des candidats en arguments *vs* non-arguments nécessitent énormément de temps de calcul. Nous avons fait une vingtaine d'expériences d'entraînement de classificateurs avec plusieurs logiciels et paramètres différents, tout en variant les sections du corpus utilisées.

L'entraînement d'un classificateur sur un ensemble restreint de 4 sections du PropBank, wsj-[01-04] (91 122 annotations produisant 7 255 767 candidats) à nécessité 6 jours de calcul. Selon les caractéristiques souhaitées du classificateur nous sommes capables d'obtenir des résultats similaires en n'entraînant que sur une seule section du corpus.

Puisque nous voulons utiliser la première étape de classification comme un filtre, nous voulons un taux de rappel élevé sur les arguments. Une façon d'obtenir un classificateur avec un taux de rappel élevé est de donner un poids 5 fois supérieur aux erreurs commises sur les exemples positifs avec une fonction de noyau polynomiale de degré 2. L'entraînement d'un tel classificateur, réalisé sur le corpus wsj-01 donne avec les corpus wsj-[01-04], wsj-05 et wsj-06, une moyenne de précision de 60%, de rappel de 92% et d'*accuracy* de $96,4\%$ ($|$candidats bien classifiés$|/|$candidats$|$) Pour notre problème ces résultats sont acceptables car nous récupérons plus de 90% des arguments, les 2 candidats sur 5 qui ne sont pas des arguments pourront être éliminés lors de l'étape suivante.

Nos expériences nous ont permis de constater que le problème nécessite un très grand temps de calcul pour entraîner les classificateurs, principalement à cause de la disproportion entre les candidats arguments et non-arguments. Les méthodes de classification à base de fonction de noyau ont de la difficulté à traiter les ensembles de données débalancés.

4.2. *Évaluation de classificateurs pour la classification des arguments*

Lors de la deuxième étape, nous avons entraîné 5 classificateurs binaires sur le corpus d'entraînement wsj-[02-21] ne contenant que des arguments, de cette façon nous partons de l'hypothèse que l'étape précédente classifie les arguments et non-arguments parfaitement. Les classificateurs sont entraînés pour séparer les classes d'arguments les unes des autres, ainsi le premier classificateur entraîné sépare la classe des arguments ARG0 de tous les autres arguments. Les classificateurs ont ensuite été utilisés dans une configuration un contre tous (OVA). Il n'y a pas de classificateur pour ARG4 et ARG5, leur nombre d'occurences étant trop petit. La combinaison de ces classificateurs

appliquée au corpus wsj-23 donne les résultats rapportés dans le tableau ci-dessous. Chaque ligne du tableau donne la distribution de la classification des arguments réalisés par les classificateurs. Par exemple la ligne de ARG0 est le nombre d'arguments de type ARG0 qui ont été classifiés dans chacune des catégories. La lecture horizontale du tableau donne le taux de rappel et la lecture verticale la précision du classificateur pour chacune des catégories.

	ARG0	ARG1	ARG2	ARG3	ARG4	ARG5	ARGM	null	Total	Rappel (%)
ARG0	6653	98	6	0	0	0	19	36	6812	97.67
ARG1	221	5285	36	7	0	0	60	149	5758	91.79
ARG2	18	149	1060	0	0	0	144	218	1589	66.71
ARG3	0	15	9	108	0	0	33	69	234	46.15
ARG4	0	1	14	0	0	0	11	126	152	-
ARG5	0	0	0	0	0	0	4	13	17	-
ARGM	17	44	68	3	0	0	7640	186	7958	96.00
Total	6909	5592	1193	118	0	0	7911	797	22520	
Précision (%)	96.29	94.51	88.85	91.53	-	-	96.57	-		92.12

Les résultats de cette classification sont satisfaisants dans la mesure où la classification des arguments représentant le sujet ARG0 et l'objet direct ARG1 contient peu d'erreurs.

5. Conclusion

Dans cet article, nous avons montré qu'il est possible de développer un identificateur de rôles sémantiques à partir d'un corpus annoté et d'un algorithme de classification. L'ajout de nouveaux attributs et une meilleure sélection de ceux-ci permettraient d'améliorer nos résultats. L'utilisation d'algorithmes de classification dont le temps d'entraînement est moins prohibitif est aussi à considérer pour développer ces nouveaux attributs. La méthode décrite dans cet article permet de créer un analyseur sémantique pour analyser des courriels. L'analyseur servira à extraire les arguments liés aux prédicats contenus dans le courriel dans le but d'identifier la requête pour traiter automatiquement les courriels.

6. Bibliographie

[BAK 98] BAKER C. F., FILLMORE C. J., LOWE J. B., The Berkeley FrameNet project, *Proceedings of the COLING-ACL*, Montreal, Canada, 1998.

[CAR 04] CARRERAS X., MÀRQUEZ L., Introduction to the CoNLL-2004 Shared Task : Semantic Role Labeling., *Proceedings of CoNLL 2004*, 2004.

[CHA 01] CHANG C.-C., LIN C.-J., LIBSVM : a library for support vector machines, 2001, Software available at `http ://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[GIL 02a] GILDEA D., JURAFSKY D., Automatic labeling of semantic roles, *Computational Linguistics*, vol. 28, n^o 3, 2002, p. 245–288.

[GIL 02b] GILDEA D., PALMER M., The necessity of syntactic parsing for predicate argument recognition, *Proceedings of the 40th Annual Conference of the Association for Computational Linguistics (ACL-02)*, Philadelphia, PA, 2002.

[JOA 99] JOACHIMS T., *Advances in Kernel Methods - Support Vector Learning*, Chapitre 11 Making Large-Scale SVM Learning Practical, p. 41-56, MIT-Press, 1999.

[KIN 02] KINGSBURY P., PALMER M., From Treebank to PropBank, *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC2002)*, Las Palmas, Spain, 2002.

[PRA 04] PRADHAN S., WARD W., HACIOGLU K., MARTIN J. H., JURAFSKY D., Shallow Semantic parsing using support vector machines, *Proceedings of the Human Language Technology Conference/North American chapter of the Association for Computational Linguistic annual meeting*, Boston, MA, May 2004, Association for Computational Linguistics.

Analyse spatiale de la communauté végétale de la portion nord du désert du Chihuahua

Guillaume Blanchet, Pierre Legendre

Département de sciences biologiques,
Université de Montréal,
Case postale 6128, succursale Centre-ville,
Montréal (Québec) Canada, H3C 3J8

RÉSUMÉ. Cet article présente les résultats d'analyses statistiques réalisées sur la communauté végétale de la portion nord du désert du Chihuahua afin d'élucider sa structure spatiale. Ces analyses avaient pour but d'approfondir les résultats de Muldavin [MUL 02] en répondant aux questions suivantes : 1) à quelle échelle spatiale la communauté est-elle significativement structurée ? 2) Jusqu'à quel point peut-on expliquer la diversité bêta par les variables environnementales ? 3) Quelles sont les associations significatives d'espèces ?

MOTS-CLÉS : Concordance de Kendall, coordonnées principales d'une matrice de voisinage (CPMV), désert du Chihuahua, diversité bêta, espèces indicatrices, partition de la variation, krigeage, partitionnement par la méthode des K centroïdes (K-means).

1 Introduction

Le désert du Chihuahua couvre quelque 630000 km^2 et s'étend du Nouveau-Mexique (USA) jusqu'à l'état de San Luis Potos (Mexique). La flore de ce désert est d'une grande richesse. Nos analyses porteront sur un échantillon de 1510 sites situés dans la portion septentrionale de ce désert au Nouveau-Mexique et au Texas. 529 espèces de plantes furent répertoriées par Muldavin. Ce spécialiste de la végétation du désert chihuahuaien a développé un indice d'affinité floristique, fondé sur l'abondance de 3 espèces indicatrices, qui lui a permis d'établir la limite nord du désert du Chihuahua au Nouveau-Mexique [MUL 02].

Dans les écosystèmes, un très grand nombre de variables influencent la distribution des espèces vivant dans un milieu donné. Pour structurer l'échantillonnage, il importe de se poser des questions précises qui permettront de déterminer quelles variables doivent être observées ou mesurées. Les analyses présentées dans cet article ont pour but de répondre à trois questions à propos du désert du Chihuahua : 1) à quelle échelle spatiale la communauté végétale est-elle significativement structurée ? 2) Jusqu'à quel point peut-on expliquer la variation de composition entre les sites (diversité bêta) par la variation en altitude des sites d'échantillonnage ? 3) Quelles sont les associations significatives d'espèces ?

Différentes méthodes seront utilisées afin de répondre à ces questions. Pour la première question, les coordonnées principales d'une matrice de voisinages (CPMV) seront utilisées [BOR 02, LEG 05b], alors que la deuxième question trouvera réponse avec le partitionnement de la variation [BOR 92]. Une analyse de concordance de Kendall [LEG 05a] sera utilisée pour répondre à la troisième question. Avant ces dernières analyses, le tableau des abondances d'espèces a subi une transformation de Hellinger qui consiste à calculer la racine carrée des données préalablement transformées en abondances relatives par site [LEG 01].

2 Structure spatiale de la communauté végétale

L'analyse en coordonnées principales d'une matrice de voisinage (CPMV) réalise une décomposition spectrale des relations spatiales entre les sites. Cette décomposition permet de modéliser la structure spatiale multi-échelle du tableau sites × espèces en utilisant uniquement les variables CPMV dérivées des positions géographiques des sites. Les CPMV peuvent décrire des structures spatiales contrôlées par des processus à échelle très large, comme la présence de chaînes de montagnes de part et d'autre de la région d'échantillonnage, mais aussi des phénomènes locaux comme la présence d'une espèce dans une zone très spécifique de la région échantillonnée. 71 CPMV significatives, sur 722 au départ, ont été retenues par sélection progressive pas à pas pour décrire la structure spatiale de la végétation du désert du Chihuahua. Elles représentent des influences spatiales à échelle large (taches d'environ 100 km : Fig. 1, gauche) à moyenne (taches d'environ 10 km). Une CPMV se distingue : elle explique une portion significative de variation spatiale dans une zone particulière ; la tache mesure environ 10 km de diamètre (Fig. 1, droite).

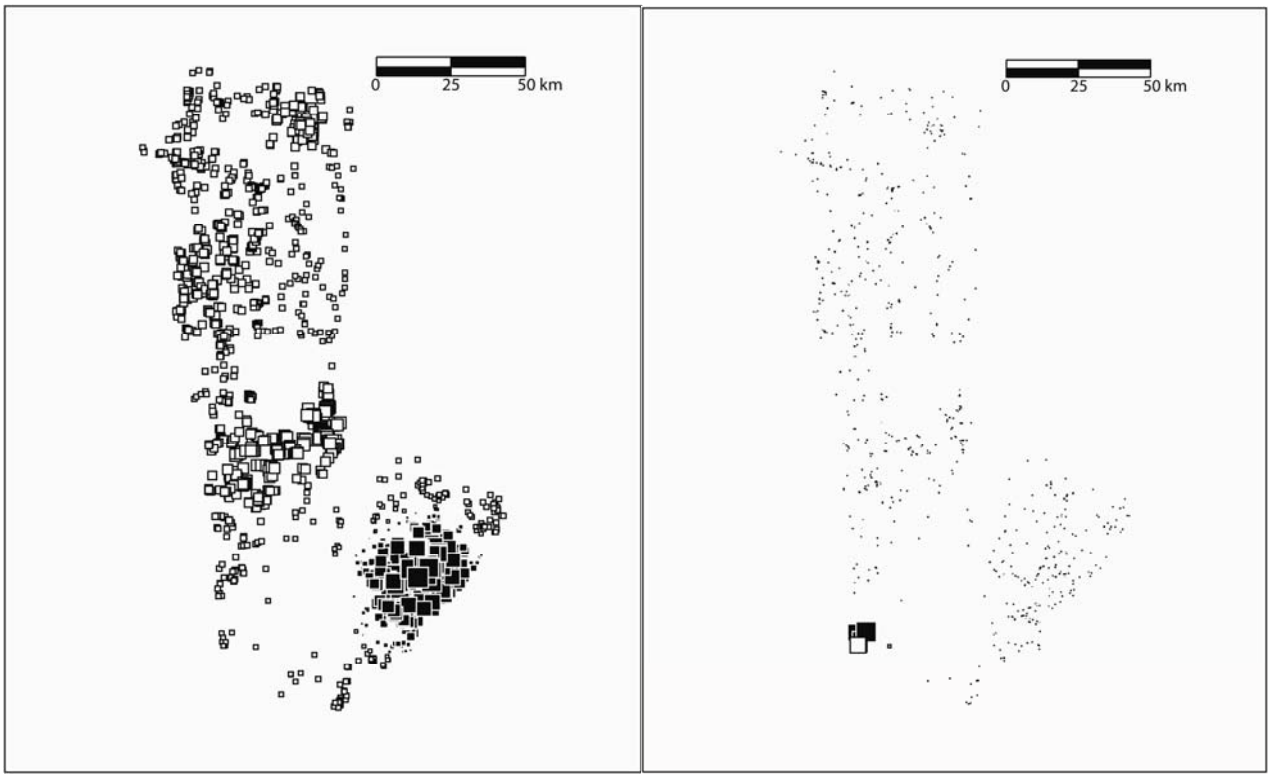


Figure 1. Carte de deux des variables CPMV expliquant une portion significative de la variation du tableau sites × espèces. La taille des carrés est proportionnelle à la valeur de la variable CPMV à chaque site ; carrés blancs : signe négatif ; carrés noirs : signe positif. La carte de gauche représente la première CPMV qui explique un phénomène à grande échelle ; celle de droite correspond à un phénomène local situé dans la portion sud-ouest de la carte.

3 Partition de la diversité bêta : influences environnementale et spatiale

Nous ajouterons maintenant une variable environnementale importante, l'altitude, à l'analyse et tenterons de déterminer jusqu'à quel point nous pouvons expliquer la variation de composition entre les sites (diversité bêta) par la variation en altitude des sites d'échantillonnage.

Pour vérifier si l'altitude avait de l'importance pour expliquer la variation en composition végétale entre les sites, nous avons partitionné les sites par la méthode des *K* centroïdes (*K*-means), puis testé l'hypothèse que les sites se regroupent en fonction de l'altitude. La méthode des *K* centroïdes regroupe les

sites selon leurs proximités ; l'algorithme a préservé la distance de Hellinger entre les sites puisque les données avaient subi une transformation de Hellinger avant le partitionnement [LEG 01]. Nous avons retenu la solution minimisant la variation intragroupe, pour $K = 2$ à $K = 10$ groupes, après 100 attributions aléatoires des objets aux groupes initiaux. La statistique de Calinski-Harabasz fut utilisée pour déterminer le nombre optimal de groupes au sens des moindres carrés. Cette statistique indiqua que la division des sites en 5 groupes était optimale. Une analyse de variance portant sur la variable altitude supporta l'hypothèse d'une forte relation entre ces 5 groupes et l'altitude des sites ($F = 185.80$, $P < 0.0001$). La cartographie de l'altitude par krigeage avait été nécessaire pour obtenir des estimations de l'altitude aux 1510 sites d'échantillonnage alors que des données d'altitude n'étaient disponibles que pour 1471 sites.

L'examen visuel de la distribution des 5 groupes de sites superposés à la topographie sur la carte permet d'apprécier leur relation avec l'altitude. Les deux groupes de sites les plus importants sont situés à basse altitude, dans une vallée. Un groupe se trouve au flanc des deux chaînes de montagne alors que les deux autres groupes recouvrent les régions de plus hautes altitudes.

La partition de la variation spatiale du tableau sites × espèces (diversité bêta), par analyse canonique de redondance (ACR ou RDA), a permis de déterminer quelle fraction de la variation des données est influencée par quelles variables. La communauté végétale du désert du Chihuahua est structurée à 19.4 % (R^2 ajusté, $P = 0.001$; $R^2 = 57.9$ %) par les 722 variables CPMV et à 5.5 % (R^2 ajusté, $P = 0.001$; $R^2 = 5.6$ %) par l'altitude. 71 des 722 CPMV ont été retenues par sélection progressive des variables explicatives en ACR. Cette méthode permet aussi de calculer, par soustraction, la fraction [b] qui correspond à la variation expliquée conjointement par les deux types de variables (CPMV et altitude); cette portion compte pour 4.4 % (R^2 ajusté) de la variation du tableau sites × espèces (Fig. 2).

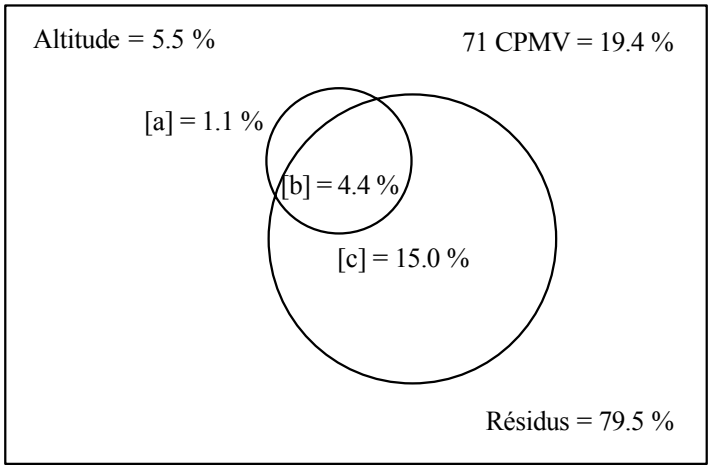


Figure 2. Diagramme de Venn partitionnant la variation spatiale (R^2 ajustés) de la communauté végétale du désert du Chihuahua. [a + b] représente la portion de la variation expliquée par l'altitude alors que [b + c] est la portion représentée par les variables CPMV. Les pourcentages sont relatifs à la variation totale du tableau sites × espèces après transformation de Hellinger ($n = 1510$, $p = 529$). Les portions de variation expliquées par les cercles sont approximatives.

Nous avons cherché à savoir si certaines espèces étaient particulièrement associées à chacun des 5 groupes de sites. Pour ce faire, nous avons utilisé une méthode de recherche des espèces indicatrices des groupes d'une partition [DUF 97]. Au niveau de signification $\alpha = 0.01$, nous avons identifié 170 espèces indicatrices de l'un ou l'autre des cinq groupes de sites.

4 Les associations significatives d'espèces

Les associations d'espèces sont intéressantes car elles permettent de synthétiser certaines caractéristiques environnementales au niveau des associations plutôt qu'à celui des espèces individuelles. Par ailleurs, les

associations d'espèces sont moins sujettes à l'erreur d'échantillonnage que les espèces individuelles ; on peut donc les utiliser pour prédire certaines caractéristiques du milieu physique [LEG 05a]. Les espèces furent partitionnées en groupes en appliquant la méthode des *K* centroïdes appliquée aux vecteurs propres d'une analyse en composante principale (ACP) de la matrice de corrélations entre les espèces. Le coefficient de Calinski-Harabasz présentait un maximum pour la division des espèces en 9 groupes.

Le coefficient de concordance de Kendall (*W*) a permis de tester la concordance des espèces réunies en groupes préliminaires et d'identifier celles qui sont significativement concordantes avec d'autres espèces de leur groupe [LEG 05a]. Le nombre d'espèces significativement concordantes par groupe va de 47 pour chacun des deux plus grands groupes à 28, 19, 18, 16, 15, 15, et enfin 9 pour le plus petit groupe.

5 Discussion

La plupart des 71 variables CPMV significatives représentent la variation spatiale de la communauté végétale à échelle large dans le nord du désert du Chihuahua. L'influence que pourraient avoir les deux chaînes de montagnes sur la répartition spatiale des espèces en est un bon exemple. Il faut cependant noter la présence possible d'une ou de quelques espèces endémiques à une zone particulière du désert, la dernière variable CPMV significative ayant détecté une structure spatiale très locale.

La partition de la variation a montré que 22.4 % (R^2 ajusté) de la diversité bêta était explicable par des variables CPMV alors que l'altitude n'expliquait que 21% de cette variation spatiale. Puisque les variables CPMV significatives représentent surtout de la variation à échelle large, cela suggère que la variation qu'elles modélisent a pour origine des variables physiques du milieu. Il nous faudra ajouter des variables physiques ou pédologiques au tableau des variables environnementales. Considérant le fait qu'il y a une vallée dans la région d'échantillonnages, il est cependant possible que la dispersion de certaines espèces se soit arrêtée parce qu'elles ne pouvaient pas passer la barrière naturelle créée par la chaîne de montagnes. Il se peut également qu'un microclimat dans la vallée favorise la reproduction de certaines espèces.

Muldavin [MUL 02], a utilisé trois espèces de plantes, qu'il a considérées comme indicatrices, pour établir la limite nord du désert du Chihuahua au Nouveau-Mexique. Les résultats de notre recherche d'espèces indicatrices fournissent de nouvelles espèces qui pourraient être utilisées pour construire de nouveaux indices pour délimiter le désert du Chihuahua.

6 Bibliographie

[BOR 92] BORCARD D., LEGENDRE P., DRAPEAU P., "Partialling out the spatial component of ecological variation", *Ecology*, vol. 73, n° 3, 1992, p. 1045-1055.

[BOR 02] BORCARD D., LEGENDRE P., "All-scale analysis of ecological data by means of principal coordinates of neighbour matrices", *Ecological Modelling*, vol. 153, 2002, p. 51-68.

[DUF 97] DUFRÊNE M., LEGENDRE P., "Species assemblages and indicator species: the need for a flexible asymmetrical approach", *Ecological Monographs*, vol. 67, n°3, 1997, p. 345-366.

[LEG 05a] LEGENDRE P., "Species association: the Kendall coefficient of concordance revisited", *Journal of Agricultural, Biological and Environmental Statistics*, 2005 (*sous presse*).

[LEG 05b] LEGENDRE P., BORCARD D., "Quelles sont les échelles spatiales importantes dans un écosystème ?", *in: Analyse statistique de données spatiales*, Droesbeke J.-J., Lejeune M., Saporta G., éditeurs, Éditions TECHNIP, Paris, 2005 (*sous presse*).

[LEG 01] LEGENDRE P., GALLAGHER E.D., "Ecologically meaningful transformations for ordination of species data", *Oecologia*, vol. 129, 2001, p. 271-280.

[MUL 02] MULDAVIN E.H., "Some floristic characteristics of the northern Chihuahuan Desert: a search for its northern boundary", *Taxon,* vol. 51, 2002, p. 453-462.

Apprentissage des délais dans les réseaux de neurones récurrents. Application à la prévision de séries temporelles.

Romuald Boné, Hubert Cardot

Université François-Rabelais de Tours,
Laboratoire d'Informatique (EA 2101),
64 avenue Jean Portalis
37200 Tours, FRANCE
{romuald.bone, hubert.cardot}@univ-tours.fr

RÉSUMÉ. Les réseaux de neurones récurrents sont par nature des outils bien adaptés à la prévision des séries temporelles. L'utilisation de connexions à délais judicieusement choisies permet d'améliorer la prise en compte des dépendances à long terme des algorithmes basés sur le gradient. Nous démontrons que le principe de l'apprentissage des délais eux-mêmes par le gradient, efficace dans le cas des réseaux à propagation avant et théoriquement extensible aux réseaux récurrents, se révèle d'une utilisation délicate pour ces derniers. Nous évaluons les performances de l'algorithme ainsi obtenu sur deux séries temporelles de référence.

MOTS-CLÉS : Prévision, séries temporelles, réseaux de neurones récurrents, connexions à délais, algorithme d'apprentissage.

1 Introduction

Les réseaux de neurones artificiels récurrents (RNR), qui possèdent une mémoire interne grâce aux cycles dans leur graphe d'interconnexion, ont des capacités d'approximation universelle pour les problèmes temporels (voir par exemple [JIN 95]) comparables à celles des réseaux à propagation avant (RPA) pour les problèmes statiques. Cependant, en pratique les réseaux récurrents demeurent peu répandus. En effet, ils sont le plus souvent associés à des algorithmes d'apprentissage basés sur le calcul du gradient, comme la rétropropagation à travers le temps (Back Propagation Through Time, BPTT), plus consommateurs de temps de calcul que leurs équivalents pour les RPA, à nombre de paramètres égal. Par ailleurs, ces algorithmes éprouvent des difficultés dans la prise en compte des dépendances à long terme [BEN 94][LIN 96]. Une alternative est d'utiliser des architectures récurrentes conservant une structure à couches qui partagent la caractéristique d'avoir été initialement élaborées pour pouvoir utiliser la rétropropagation du gradient de l'erreur issue des RPA (voir [CAM 99] pour des versions adaptées). Ainsi les RPA localement récurrents [TSO 94] introduisent des neurones particuliers, avec bouclages locaux. Dans la forme la plus générale, ces neurones peuvent comporter des retards en entrée ou dans les bouclages. Ces architectures à bouclage local restent toutes limitées : les neurones cachés sont mutuellement indépendants et ne peuvent donc pas capter certaines dynamiques complexes. Pour limiter ce problème, un certain nombre d'architectures à couches récurrentes ont été proposées (voir [LIN 96] [CAM 99] pour une présentation) dont les connexions récurrentes peuvent être « gelées » (poids fixes).
Cependant la connectivité restreinte de ces modèles n'offre potentiellement pas la même puissance que les réseaux globalement récurrents. Il a été démontré qu'en pratique l'emploi de connexions à retards dans ces réseaux permettait de réduire les temps d'apprentissage [GUI 94] et d'améliorer la prise en compte des

dépendances à long terme [LIN 96] [BON 02]. On parle alors de réseaux récurrents à délais ou à retards. Dans ce cas, à moins d'appliquer des algorithmes d'ajout sélectif de connexions à délais [BON 02] qui améliorent les performances en prévision mais au prix d'un surcroît de calculs, les réseaux finalement retenus sont souvent surdimensionnés, utilisant des méta-connexions avec connexions à retards consécutifs (encore nommées connexions FIR ou, si elles contiennent des bouclages, IIR [TSO 94]). L'algorithme d'apprentissage doit traiter les paramètres supplémentaires associés à ces connexions. La solution peut alors résider dans l'apprentissage des délais des connexions. [DUR 99] a proposé, pour un RPA associant un délai à chaque connexion, un algorithme basé sur le gradient qui ajuste simultanément les poids et les délais. Nous proposons d'étudier l'adaptation de cette technique aux RNR.

2 Apprentissage des délais

Considérons un RNR où à chaque connexion d'un neurone j à un neurone i sont associées deux valeurs : une pondération classique w_{ij} du signal et un retard τ_{ij} de valeur non entière indiquant le temps nécessaire à ce signal pour traverser la connexion. La sortie d'un neurone $s_i(t)$ est donnée par :

$$s_i(t) = f_i(net_i(t-1)) \text{ et } net_i(t-1) = \sum_{j \in Pred(i)} w_{ij} s_j(t - \tau_{ij} - 1)$$

Les valeurs $s_j(t - \tau_{ij} - 1)$ sont obtenues en effectuant une interpolation linéaire correspondant aux deux valeurs entières les plus proches du retard τ_{ij} [DUR 99]. L'ensemble $Pred(i)$ contient pour un neurone i les indices des neurones entrants $Pred(i) = \{ j \in N \mid \exists (w_{ij}, \tau_{ij}) \}$. De la même manière, nous définissons les neurones successeurs d'un neurone i : $Succ(i) = \{ j \in N \mid \exists (w_{ji}, \tau_{ji}) \}$.

Nous adaptons l'algorithme BPTT à cette architecture en réalisant un apprentissage simultané des poids et des délais des connexions, inspiré de [DUR 99]. L'idée centrale de l'algorithme BPTT est de déplier dans le temps le réseau récurrent d'origine pour obtenir un réseau à l couches à propagation avant, qui permette d'appliquer l'apprentissage par rétropropagation du gradient de l'erreur.

La variation d'un retard τ_{ij} se calcule comme la somme des variations des copies de ce paramètre sur chaque élément de la séquence correspondant aux instants t_1 à t_l. On ajoute ensuite à toutes les copies cette somme. Nous ne donnons ici que la démonstration de l'apprentissage des retards, l'apprentissage des poids s'en déduisant aisément. Nous appliquons une remontée du gradient sur l'erreur quadratique moyenne $E(t_1, t_l)$ définie comme la somme des erreurs instantanées $e(t)$ entre t_1 à t_l :

$$E(t_1, t_l) = \sum_{t=t_1}^{t_l} e(t) = \sum_{t=t_1}^{t_l} \frac{1}{2} \sum_{p \in T(t)} (d_p(t) - s_p(t))^2$$

$$\Delta\tau_{ij}(t_1, t_l - 1) = -\lambda \frac{\partial E(t_1, t_l)}{\partial \tau_{ij}} = \sum_{\tau = t_1 + \lceil \tau_{ij} \rceil}^{t_l - 1} \Delta\tau_{ij}(\tau) = -\lambda \sum_{\tau = t_1 + \lceil \tau_{ij} \rceil}^{t_l - 1} \frac{\partial E(t_1, t_l)}{\partial \tau_{ij}(\tau)}$$

où $T(t)$ est l'ensemble des indices des neurones ayant une sortie désirée à l'instant t, $d_p(t)$ est la sortie désirée du neurone p à cet instant et $\Delta\tau_{ij}(\tau)$ est la copie de τ_{ij} pour $t = \tau$ dans le réseau déplié dans le temps que BPTT construit virtuellement [RUM 86]. Nous notons $\lceil . \rceil$ l'opérateur d'arrondi par excès. On peut écrire

$$\frac{\partial E(t_1, t_l)}{\partial \tau_{ij}(\tau)} = \frac{\partial E(t_1, t_l)}{\partial net_i(\tau)} \frac{\partial net_i(\tau)}{\partial \tau_{ij}(\tau)}$$

Par approximation du premier ordre, $\partial net_i(\tau) / \partial \tau_{ij}(\tau) \approx w_{ij} (s_j(\tau - \tau_{ij} - 1) - s_j(\tau - \tau_{ij}))$. Calculons $\partial E(t_1, t_l) / \partial net_i(\tau)$:

$$\frac{\partial E(t_1,t_l)}{\partial net_i(\tau)} = \frac{\partial E(t_1,t_l)}{\partial s_i(\tau+1)} \frac{\partial s_i(\tau+1)}{\partial net_i(\tau)} \text{ avec } \frac{\partial s_i(\tau+1)}{\partial net_i(\tau)} = f'(net_i(\tau)).$$

Si le neurone i appartient à la dernière couche ($\tau = t_l - 1$) :

$$\frac{\partial E(t_1,t_l)}{\partial s_i(\tau+1)} = \frac{\partial e(t_l)}{\partial s_i(t_l)} = \underset{i\in T(\tau+1)}{\delta}(s_i(t_l) - d_i(t_l))$$

où $\underset{i\in T(\tau+1)}{\delta} = 1$ si $i \in T(\tau+1)$ et 0 sinon. Si le neurone i fait partie des couches précédentes :

$$\frac{\partial E(t_1,t_l)}{\partial s_i(\tau+1)} = \frac{\partial e(\tau+1)}{\partial s_i(\tau+1)} + \sum_{j\in Succ(i)} \left(\frac{\partial E(t_1,t_l)}{\partial net_j(\tau+\tau_{ji}+1)} \frac{\partial net_j(\tau+\tau_{ji}+1)}{\partial s_i(\tau+1)} \right)$$

or $\partial net_j(\tau+\tau_{ji}+1) / \partial s_i(\tau+1) = w_{ji}(\tau+1)$. Pour finir, les relations qui permettent d'apprendre le délai associé à chaque connexion sont :

$$\Delta\tau_{ij}(t_1,t_l-1) = -\lambda \sum_{\tau=t_1+k}^{t_l-1} \frac{\partial E(t_1,t_l)}{\partial net_i(\tau)} w_{ij}(s_j(\tau-\tau_{ij}-1) - s_j(\tau-\tau_{ij})) \text{ avec}$$

$$\frac{\partial E(t_1,t_l)}{\partial net_i(\tau)} = \underset{i\in T(\tau+1)}{\delta}(s_i(t_l) - d_i(t_l)) f_i'(net_i(\tau)) \text{ pour } \tau = t_l - 1 \text{ et}$$

$$\frac{\partial E(t_1,t_l)}{\partial net_i(\tau)} = \left[\begin{array}{l} \underset{i\in T(\tau+1)}{\delta}(s_i(\tau+1) - d_i(\tau+1)) \\ + \sum_{j\in Succ(i)} \frac{\partial E(t_1,t_l)}{\partial net_j(\tau+\tau_{ji}+1)} w_{ji}(\tau+1) \end{array} \right] f_i'(net_i(\tau)) \text{ pour } t_1 \le \tau < t_l - 1.$$

3 Expérimentations

Nous avons appliqué nos algorithmes à des réseaux récurrents avec un neurone d'entrée, un neurone (linéaire) de sortie, un neurone de biais et une couche cachée totalement récurrente composée de neurones à fonction de transfert tangente hyperbolique. Nous avons abordé la prévision à un pas de temps de deux séries temporelles de référence, dont on trouvera une présentation détaillée dans [BON 02]. Les taches solaires sont des taches sombres liées à l'activité du champ magnétique du soleil. La série donne le nombre moyen annuel de taches apparues à la surface du soleil de 1700 à 1979. L'ensemble d'apprentissage contient les valeurs de la période 1700-1920, les performances étant évaluées sur deux ensembles, 1921-1955 (test1) et 1956-1979 (test2, considéré plus difficile car à variance plus élevée). Les séries de Mackey-Glass sont des séries de référence pour l'évaluation de nombreux systèmes de prévision. Ces séries sont générées par l'équation différentielle non linéaire $dx(t)/dt = -0{,}1x(t) + 0{,}2x(t-\tau)/(1+x^{10}(t-\tau))$. Pour $\tau > 16{,}8$, la dynamique du modèle correspond au chaos déterministe. Le choix de $\tau = 17$, valeur habituellement retenue, permet de générer la série MG 17.
Les tableaux 1 et 2 ci-après comparent les meilleurs résultats obtenus (erreur quadratique moyenne normalisée, EQMN) pour respectivement 3 et 7 neurones cachés avec ceux de la littérature. On pourra trouver une description détaillée des modèles cités dans les tableaux dans [BON 00] et [BON 02]. Les premiers résultats montrent un comportement parfois instable de l'algorithme, certains apprentissages se bloquant rapidement à des valeurs d'erreur élevées. L'état interne du réseau (l'ensemble des sorties des neurones de la couche cachée) se révèle très sensible aux variations de délais. Le choix des deux pas d'apprentissage, pour les poids ou les délais des connexions, nécessitent un réglage très délicat. Il est intéressant de noter que nos résultats sont légèrement inférieurs à ceux de l'algorithme EBPTT [BON 00] qui ajoute des connexions à un RNR sans utiliser directement une méthode liée au gradient de l'erreur.

Modèle	Test 1	Test 2
Copie carbone	0,427	0,966
Autorégressif à seuil	0,097	0,280
RPA	0,086	0,350
RNR à connexions FIR 1	0,091	0,273
RNR à connexions FIR 2	0,093	0,246
RNR avec BPTT	0,084	0,300
RNR avec EBPTT	0,078	0,227
Notre algorithme	0,081	0,261

Tableau 1 : EQMN sur la série des taches solaires

Modèle	EQMN
Réseau RBF	$10,7.10^{-3}$
RPA	10.10^{-3}
RPA à connexions FIR	$4,9.10^{-3}$
TDNN	$0,8.10^{-3}$
RNR à connexions FIR	$4,7.10^{-3}$
RNR avec BPTT	$0,23.10^{-3}$
RNR avec EBPTT	$0,13.10^{-3}$
Notre algorithme	$0,15.10^{-3}$

Tableau 2 : EQMN sur la série MG 17

4 Conclusion

Un nouveau type de neurone avec connexions à délais non entiers a été adapté aux réseaux récurrents. Un nouvel algorithme d'apprentissage dédié a été présenté. Les meilleurs résultats obtenus sur deux problèmes de prévision sont encourageants mais démontrent un paramétrage délicat et un comportement instable de l'algorithme. L'architecture semble apporter un surcroît de puissance mais une alternative à l'utilisation du calcul de gradient reste à étudier pour l'apprentissage des délais. Nous travaillons actuellement sur une version stochastique.

5 Bibliographie

[BEN 94] BENGIO Y., SIMARD P., FRASCONI P., Learning Long-Term Dependencies with Gradient Descent is Difficult, *IEEE Transactions on Neural Networks*, vol. 5, n° 2, p. 157-166, 1994.

[BON 00] BONÉ R., CRUCIANU M., VERLEY G., ASSELIN DE BEAUVILLE J.-P., A Bounded Exploration Approach to Constructive Algorithms for Recurrent Neural Networks, *International Joint Conference on Neural Networks*, Como, Italy, 2000.

[BON 02] BONÉ R., CRUCIANU M., ASSELIN DE BEAUVILLE J.-P., Learning Long-Term Dependencies by the Selective Addition of Time-Delayed Connections to Recurrent Neural Networks, *NeuroComputing*, vol. 48, n° 1-4, p. 251-266, 2002.

[CAM 99] CAMPOLUCCI P., UNCINI A., PIAZZA F., RAO B. D., On-Line Learning Algorithms for Locally Recurrent Neural Networks, *IEEE Trans. on Neural Networks*, vol. 10, n° 2, p. 253-271, 1999.

[DUR 99] DURO R. J., SANTOS REYES J., Discrete-Time Backpropagation for Training Synaptic Delay-Based Artificial Neural Networks, *IEEE Transactions on Neural Networks*, vol. 10, n° 4, p. 779-789, 1999.

[GUI 94] GUIGNOT J., GALLINARI P., Recurrent Neural Networks with Delays, *International Conference on Artificial Neural Networks*, p. 389-392, Sorrento, Italy, 1994.

[JIN 95] JIN L., NIKIFORUK N., GUPTA M. M., Uniform Approximation of Nonlinear Dynamic Systems Using Dynamic Neural Networks, *International Conference on Artificial Neural Networks*, p. 191-196, Paris, France, 1995.

[LIN 96] LIN T., HORNE B. G., TINO P., GILES C. L., Learning Long-Term Dependencies in NARX Recurrent Neural Networks, *IEEE Transactions on Neural Networks*, vol. 7, n° 6, p. 13-29, 1996.

[RUM 86] RUMELHART D. E., HINTON G. E., WILLIAMS R. J., Learning Internal Representations by Error Propagation, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, D. E. Rumelhart and J. McClelland eds, Cambridge, MA, MIT Press, p. 318-362 , 1986.

[TSO 94] TSOI A. C., BACK A. D., Locally Recurrent Globally Feedforward Networks: A Critical Review of Architectures, *IEEE Transactions on Neural Networks*, vol. 5, n° 2, p. 229-239, 1994.

Utilisation du recuit simulé pour la recherche d'une ultramétrique optimale.

M. BOUBOU, A. BOUNEKKAR, D. TOUNISSOUX, M. LAMURE

Université Claude Bernard Lyon I
LASS - UMR5823 - bat Braconnier
43, boulevard du 11/11/1918
{boubou,bounekkar,tounissoux,lamure}@univ-lyon1.fr

Résumé : Dans ce papier, nous proposons une méthode de construction d'une ultramétrique parmi les plus proches d'une dissimilarité donnée, définie sur un ensemble d'individus, au sens des moindres carrés. La méthode du recuit simulé est utilisée pour résoudre le problème d'optimisation. La résolution est décomposée en étapes afin d'arriver à une hiérarchie ne présentant pas d'inversion.

MOTS-CLES : classification, hiérarchie, ultramétrique, recuit simulé, dissimilarité, moindres carrés.

1 Introduction

Dans une classification hiérarchique, la hiérarchie indicée permet de construire une ultramétrique sur un ensemble d'individus. Cette classification ne vise généralement pas à optimiser un critère global portant sur la hiérarchie indicée ou sur l'ultramétrique. Dans ce papier, nous proposons une méthode de recherche d'une ultramétrique parmi les plus proches des dissimilarités définies sur l'ensemble des individus. Pour minimiser l'écart entre la dissimilarité adoptée et l'ultramétrique, nous avons utilisé l'algorithme du recuit simulé. Après quelques rappels de définitions concernant les hiérarchies valuées et indicées, nous présentons le problème à résoudre et nous justifions la décomposition de ce problème en sous-problèmes intermédiaires. Lors de la résolution du problème, nous proposons d'utiliser des méthodes spécifiques afin de réduire la complexité des calculs.

2 Définitions

Considérons un ensemble fini I d'individus, de cardinal n, muni d'une dissimilarité ρ. Nous rappelons qu'un ensemble de parties $\mathcal{A}$ non vides de I, ordonné par la relation d'inclusion est appelé hiérarchie totale binaire sur I, s'il satisfait aux trois propositions suivantes :

(i) $\forall A,B \in \mathcal{A}$ soit $A \cap B = \phi$ soit $(A \subseteq B$ ou $B \subseteq A)$

(ii) $I \in \mathcal{A}$ et $\forall i \in I, \{i\} \in \mathcal{A}$

(iii) $\forall A \in \mathcal{A}$, $Card(A)>1 \Rightarrow \exists A_1, A_2 \in \mathcal{A}, A_1 \cap A_2 = \varnothing$ tel que $A = A_1 \cup A_2$

En pratique, la gestion d'une hiérarchie $\mathcal{A}$ est assurée au moyen de deux fonctions a et b :

$\forall A \in \mathcal{A}$ tel que $A = A_1 \cup A_2$ avec $A_1 \in \mathcal{A}$ *et* $A_2 \in \mathcal{A}$, $a(A) = A_1$ et $b(A) = A_2$.
On appelle *a(A)* le fils aîné de *A* et *b(A)* le benjamin de *A*. Une permutation entre *a(A)* et *b(A)* ne modifie pas la hiérarchie.
Dans la suite, pour simplifier, nous parlerons de hiérarchie pour désigner une hiérarchie totale binaire.
On dit que la hiérarchie $\mathcal{A}$ est valuée par la fonction $f: A \rightarrow R^+$ *si* $f(\{i\}) = 0 \ \forall i \in I$. Nous noterons $\mathcal{H}_0$ l'ensemble des hiérarchies valuées sur *I* et $H = (\mathcal{A}, f)$ un élément de $\mathcal{H}_0$.
Une hiérarchie valuée $(\mathcal{A}, f)$ est dite indicée si : $\forall A,B \in \mathcal{A}, A \subseteq B \Rightarrow f(A) \leq f(B)$.
Nous désignerons par $\mathcal{H}$ l'ensemble des hiérarchies indicées sur *I*. On sait qu'une hiérarchie indicée permet de construire une distance ultramétrique δ sur *I* de la façon suivante :
Etant donné $H = (\mathcal{A}, f)$, $\forall (i,j) \in I \times I$ on considère le plus petit élément *M(i,j)* (au sens de l'inclusion) de $\mathcal{A}$ qui contient à la fois *i* et *j*; $\delta(i,j)$ est alors défini par :

$$\delta(i,j) = f(M(i,j)) \qquad (1)$$

Réciproquement, la donnée d'une distance ultramétrique permet de construire une hiérarchie indicée.
Notons que si $(\mathcal{A}, f)$ est une hiérarchie valuée (1) permet seulement de construire une dissimilarité δ sur *I*.
Les algorithmes ascendants de construction de hiérarchies généralement ne visent pas à optimiser un critère global. L'algorithme de *Lance* et *Williams* basé sur le critère du saut minimal est une exception à cette affirmation.
Plusieurs auteurs ont posé le problème de trouver l'ultramétrique δ^* qui minimise la quantité :

$$\Phi(\delta) = \sum_{(i,j) \in I \times I} [\rho(i,j) - \delta(i,j)]^2$$

Devant la complexité de ce problème, des algorithmes de recherche d'un optimal local ont été proposés. *Chandon et al.* dans [CHA 80] ont proposé un algorithme (*Branch and bound*) fournissant un optimal global, mais inutilisable pour des ensembles comportant plus d'une dizaine d'éléments.
Nous désignons par (*P*) le problème suivant : Trouver une hiérarchie indicée $H^* = (\mathcal{A}^*, f^*)$ telle que :

$$\Phi(\delta^*) = Min[\Phi(\delta) \mid H \in \mathcal{H}] \qquad (P)$$

où δ et δ^* désignent respectivement les ultramétrique associées à *H* et *H**.
C'est le problème (*P*) que nous proposons de résoudre par la méthode du recuit simulé.
Pour des raisons que nous exposerons plus loin, on ne peut pas résoudre le problème (*P*). Nous commencerons par proposer une méthode permettant de résoudre le problème (P_0) suivant :
Trouver la hiérarchie valuée $H_0^* = (A_0^*, f_0^*)$ telle que $\Phi(\delta_0^*) = Min[\Phi(\delta) \mid H \in \mathcal{H}_0] \qquad (P_0)$
où δ et δ_0^* désignent resp. Les dissimilarités associées à *H* et H_0^*.

3 Utilisation de l'algorithme du recuit simulé pour la résolution du problème (P_0)

3.1 Principe de l'algorithme.

Le principe de la méthode est basé sur l'algorithme du recuit simulé [KIR 83] qui consiste, à partir d'une hiérarchie initiale à effectuer des modifications élémentaires :
Chaque modification est acceptée si :

- Elle conduit à une amélioration du critère.
- Elle conduit à une dégradation du critère avec une probabilité qui tend vers zéro.

La propriété suivante [GOR 99] caractérisant la valuation de l'optimal permet de résoudre le problème (P_0)

Propriété 1 :

Soit $H_0^* = (A_0^*, f_0^*)$ une solution du problème (P_0). Pour tout sommet *s* de H_0^*, on a nécessairement :

$$f_0^*(s) = \sum_{\substack{i \in a(s) \\ j \in b(s)}} \rho(i,j) \frac{1}{|a(s)||b(s)|} \qquad (2)$$

Remarque :

On peut noter que la solution H^* du problème (*P*) vérifie la même propriété.
Nous imposerons donc à la valuation *f* de la hiérarchie valuée *H* de l'algorithme de toujours vérifier la propriété (2), plus précisément :

- la hiérarchie H_0^* résultante de l'algorithme sera construite de façon à vérifier (2),
- après chaque modification élémentaire, la fonction *f* sera également modifiée, de façon que (2) soit toujours vérifiée.

3.2 Etude d'une modification élémentaire de la hiérarchie valuée H = (𝒜,f)

La modification élémentaire envisagée procède en deux temps :

- modification de la hiérarchie $\mathcal{A}$.
- mise à jour de la valuation de façon que (2) soit vérifiée.

Modification de 𝒜 :

Pour définir la modification de la hiérarchie $\mathcal{A}$, nous faisons le choix de deux éléments *P* et *Q* de $\mathcal{A}$ vérifiant : *Q = b(P) et Card(Q)>1.*
La modification envisagée dépend de *P*, et sera donc notée *τ(P).* La hiérarchie obtenue après transformation de $\mathcal{A}$ par *τ(P)* sera noté $\mathcal{A}'$, nous notons a' et b' les fonctions *"aîné"* et *"benjamin"* attachées à la hiérarchie $\mathcal{A}'$ Voir figure (1).
Nous notons *A = a(P) , B = a(Q) et C = b(Q). τ(P)* est la transformation qui, à la hiérarchie $\mathcal{A}$ associe la hiérarchie $\mathcal{A}'$ définie par $\mathcal{A}' = \mathcal{A} - Q + Q'$ *où* $Q' = A \cup B$. On vérifiera facilement que si $\mathcal{A}$ est une hiérarchie totale binaire sur *I*, il en est de même de $\mathcal{A}'$. Les fonctions a' et b' sont alors définies comme suit : $\forall X \in \mathcal{A}$ *tel que Card(X) >1 et* $X \neq P$, $X \neq Q$

$$a'(X) = a(X) \qquad b'(X) = b(X)$$
$$a'(P) = C \qquad b'(P) = Q'$$
$$a'(Q') = B \qquad b'(Q') = A$$

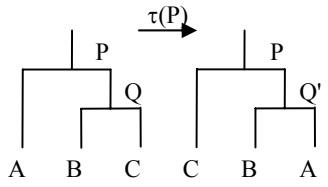


figure (1)

Mise à jour de la valuation :

Nous notons $H' = (\mathcal{A}', f')$ la hiérarchie valuée transformée de *H*. Il est clair que, compte tenu de (2) :
$(f'(X) = f(X)$ *si* $X \neq P$ *et* $X \neq Q')$

$$f'(P) = \sum_{i \in C,\ j \in A \cup B} \rho(i,j) \frac{1}{|C||A \cup B|} \qquad (3)$$

$$f'(Q') = \sum_{i \in A,\ j \in B} \rho(i,j) \frac{1}{|A||B|} \qquad (4)$$

Sur le plan algorithmique, il suffira de calculer la quantité $f'(Q')$ par la formule (4) ci-dessus. La quantité $f'(P)$ pourra s'en déduire en opérant le calcul suivant :

$$f'(P) = \frac{|A|}{|A|+|B|}\beta + \frac{|B|}{|A|+|B|} f(Q) \qquad avec : \beta = \frac{|B|+|C|}{|C|} f(P) - \frac{|B|}{|B|+|C|} f'(Q') \qquad (5)$$

Détermination de la variation du critère Φ

Soit Δ_0 la quantité $\Phi(\delta') - \Phi(\delta)$, compte tenu de la définition de δ et δ' (dissimilarité associée à la hiérarchie H'), et des remarques précédentes, on a :

$$\Delta_0 = \sum_{\substack{(i,j)\in AxB \\ (i,j)\in AxC \\ (i,j)\in BxC}} \left[(\rho(i,j) - \delta'(i,j))^2 - (\rho(i,j) - \delta(i,j))^2 \right]$$

Les relations suivantes permettent un calcul simple de Δ_0 :

$$\sum_{i\in A, j\in B} \left[(\rho(i,j) - \delta'(i,j))^2 - (\rho(i,j) - \delta(i,j))^2 \right] = -(f(P) - f'(Q'))^2 |A||B|$$

$$\sum_{i\in B, j\in C} \left[(\rho(i,j) - \delta'(i,j))^2 - (\rho(i,j) - \delta(i,j))^2 \right] = (f(Q) - f'(P))^2 |B||C|$$

$$\sum_{i\in A, j\in C} \left[(\rho(i,j) - \delta'(i,j))^2 - (\rho(i,j) - \delta(i,j))^2 \right] = (f(P) - f'(P))(2\beta - f(P) - f'(P)) |A||C|$$

Avec β défini dans (5). Seul le calcul de $f'(Q')$ a une complexité au pire en $O(n^2)$, les autres calculs ayant une complexité constante.

4 Résolution du problème *(P)*

Le problème de la recherche d'une hiérarchie indicée, solution de problème *(P)*, est un problème d'optimisation avec contraintes.

Pour tenir compte de ces contraintes, nous utilisons une méthode de pénalisation consistant à modifier à chaque pas la valeur de Δ_0 en introduisant une pénalité k suivant le cas, très schématiquement :

a) Si $f(Q) \le f(P)$ et $f'(Q') > f'(P)$ (Création d'une inversion). Δ_0 est remplacé par $\Delta = \Delta_0 + k$, $(k>0)$

b) Si $f(Q) > f(P)$ et $f'(Q') \le f'(P)$ (Suppression d'une inversion). Δ_0 est remplacé par $\Delta = \Delta_0 - k, (k>0)$

c) Dans tous les autres cas, on prend $\Delta = \Delta_0$

de cette façon les hiérarchies qui ne présentent pas d'inversion sont favorisées.

5 Conclusion

Nous savons que la recherche d'une ultramétrique optimale est un problème NP- complet.

Sachant que le recuit simulé est adapté à la résolution de tels problèmes, nous avons montré comment cet algorithme peut être utilisé dans ce contexte particulier.

6 Bibliographie :

[CHA 80] CHANDON J.L., LEMAIRE, J., POUGET J., "Construction de l'ultramétrique la plus proche d'une dissimilarité au sens des moindres carrés", *Revue d'Automatique, d'Informatique et de Recherche Opérationnelle, RAIRO*, vol. 14, n° 2, 1980, p. 157 - 170.

[GOR 99] Gordon A.D., *Classification*, Chapman, 1999.

[HAR 75] HARTIGAN J. A. *Clustering algorithms*, Wiley. 1975.

[JUA 82] JUAN J. " Programme HIVOR de classification ascendante hiérarchique selon les voisins réciproques et le critère de la variance ", *Les Cahiers de l'Analyse des Données*, vol. 7, n° 2, 1982, p. 173-184.

[KIR 83] Kirkpatrick, S., C. D. Gelatt Jr., M. P. Vecchi, " Optimization by Simulated Annealing ", *Science*, vol. 220, n° 4598, 1983, p. 671-680.

[ROU 85] Roux M., *Algorithms de classification*, Masson. 1985.

Classification vs recherche d'information : vers une caractérisation des bases d'images

Alain Boucher[1,2], Thanh-Ha Dang[1,3], Thi-Lan Le[2]

1 - Institut de la Francophonie pour l'Informatique, Bât. D, ruelle 42, rue Ta Quang Buu, Hanoi, Vietnam
2 - Centre de recherche MICA, Bât. C10, Institut Polytechnique de Hanoi, 1 Dai Co Viet, Hanoi, Vietnam
3 - Pôle IA, LIP6, Université Pierre et Marie Curie, 8 rue du Capitaine Scott, 75015 Paris, France

RÉSUMÉ. Dans la littérrature en traitement d'images, et plus particulièrement dans le domaine de la recherche d'images par le contenu, nous retrouvons fréquemment des travaux présentant des idées intéressantes mais difficiles à comparer parce que testées sur des bases d'images non-disponibles ou à accès restreint. Cet article présente quelques réflexions et idées afin de caractériser les bases d'images à des fins de comparaisons de résultats scientifiques. Pour cela, nous caractérisons les bases d'images à l'aide d'un protocole défini et reproductible, à base d'arbres de décision et de descripteurs simples comme la couleur RGB et la texture par matrices de co-occurrence. Enfin, la dernière partie de cet article compare les domaines de la classification et de l'indexation et recherche d'images par le contenu, en établissant certains parallèles entre les deux domaines.

MOTS-CLÉS : Classification, Indexation et recherche d'images par le contenu, Arbres de décision, Bases d'images

1 Introduction

La dernière décennie a vu une explosion du nombre d'articles publiés dans le domaine de l'indexation et de la recherche d'images par le contenu. Mais comme dans beaucoup de domaines, le problème de la validation et de la comparaison des résultats publiés par les différentes équipes de recherche demeure crucial [JER 02]. Alors que plusieurs voix s'élèvent pour demander l'introduction de bases d'images de référence pour comparer les approches, on voit plutôt l'effet inverse, c'est-à-dire que chaque équipe utilise souvent sa propre base d'images, soit par intérêt particulier, ou tout simplement parce que les travaux ont une finalité vers une application spécifique, donc une base d'images spécifique. Les résultats d'une approche de manipulation (classification, recherche d'information,…) d'une base d'images dépendent forcément de ses propriétés. Il est évident que si une base est « facile » alors on obtiendra de bons résultats et inversement. Cela cause des problèmes de comparaison des différentes approches proposées. Dans cet article, nous souhaitons aborder le problème sous un autre angle. Plutôt que de contraindre les chercheurs à utiliser les mêmes bases d'images, est-ce qu'il serait possible de leur donner des outils permettant de comparer les différentes bases d'images entre elles ? C'est ce que nous tentons de faire en utilisant un outil classique dans le domaine de la classification que sont les arbres de décision. Nous suggérons une méthode pour caractériser les bases d'images et ainsi permettre la comparaison des travaux de recherche. Cela permet d'estimer la difficulté des bases d'images pour donner une base de comparaison des méthodes de recherche d'images par le contenu. Bien que certains problèmes restent à résoudre, cette approche permet néanmoins de poser le problème de la validation sous un nouvel angle. Nous poussons notre réflexion un peu plus loin en établissant quelques parallèles entre les domaines de la classification et de la recherche d'information.

2 Protocole d'étude

2.1 Bases d'images

Les quatre bases d'images qui ont servi pour cette étude sont disponibles sur Internet librement sauf une et possèdent déjà des classes définies où chaque image n'appartient qu'à une seule classe :

- la base de 1000 images naturelles en couleurs (10 classes x 100 images/classe) de J.Z. Wang de l'Université de Pennsylvanie : http://wang.ist.psu.edu/docs/related/ ;
- la base de 7200 images d'objets en couleurs (100 objets x 72 images/objet) de l'Université Columbia (COIL-100) : http://www1.cs.columbia.edu/CAVE/research/softlib/coil-100.html ;
- la base de plus de 14000 images de textures (70 textures x ~200 images/texture) des Universités Columbia et d'Utrecht (CUReT) : http://www1.cs.columbia.edu/CAVE/curet ;
- la base de 347 images de grains de pollen (31 classes au total) de l'INRIA Sophia-Antipolis et provenant du projet Européen ASTHMA[1] [BON 02].

Sur toutes les images, nous avons calculé les principales caractéristiques utilisées en recherche d'images par le contenu, soit la couleur et la texture. Nous avons calculé pour la couleur l'histogramme RGB de l'image discrétisé en 24 *bins* et pour la texture un vecteur contenant les quatre caractéristiques les plus appropriés extraites des matrices de co-occurrences : énergie, entropie, contraste et moment inverse de différence. Un ensemble de 28 valeurs par image a servi de base pour les expérimentations de cet article.

2.2 Méthode de classification par arbres de décision

L'arbre de décision est une méthode très utilisée pour des raisons d'efficacité, de simplicité et d'interprétabilité par rapport aux autres méthodes existantes. Il existe plusieurs approches de construction des arbres de décision qui se distinguent principalement par le choix de la mesure de discrimination [DAN 04]. Pour cette étude, nous utilisons une version étendue de l'algorithme ID3 classique qui tient compte des attributs numériques en ajoutant une étape supplémentaire de discrétisation [MAR 98] et la coupure entre les intervalles est déterminée en utilisant l'entropie de Shannon. Les expérimentations ont été effectuées en utilisant le système DTGen (Decision Tree Generation) de l'équipe LOFTI du LIP6 et une variante de la validation croisée. La base des exemples est décomposée aléatoirement en deux parties : une partie de 80% d'exemples de chaque classe pour la base d'apprentissage et le reste forme la base de test. Cette validation est répétée un certain nombre de fois. En utilisation normale, le système donne en sortie la classe d'appartenance pour une image requête, mais nous utilisons aussi un certain nombre d'informations de la phase d'apprentissage et de test pour caractériser les différentes bases d'images (voir section 3).

2.3 Méthode de recherche d'images par le contenu

Un système d'indexation et recherche d'images par le contenu est un système qui permet de rechercher des images similaires à une requête dans une base d'images en se basant sur les caractéristiques propres aux images comme les couleurs, les textures, les formes, etc. Nous utilisons pour cette expérimentation un système classique basé sur les caractéristiques les plus importantes pour ce domaine, soient la couleur et la texture. Ce système est un sous-ensemble d'un système plus avancé [LE 04] que nous avons choisi pour établir les premières comparaisons de cet article. La méthode d'intersection d'histogrammes est utilisée pour la couleur et la distance des vecteurs de caractéristiques est calculée pour la texture. Les distances en couleurs et en textures sont normalisées indépendamment pour rétablir les ordres de grandeur des grandeurs calculées. Enfin, les distances normalisées permettent de calculer la distance entre deux images, image requête et image de la base d'images. Le système donne en sortie non pas une classe d'appartenance, mais un certain nombre d'images jugées pertinentes et similaires à l'image requête proposée. Ce nombre d'images voulues dans la réponse finale est un paramètre permettant de calculer des mesures d'évaluation de l'algorithme comme le rappel et la précision.

[1] Les auteurs remercient le projet Orion de l'INRIA pour avoir permis l'utilisation ici de cette base d'images.

3 Caractérisation des bases d'images

Nous proposons de calculer un index caractéristique pour une base d'images donnée comme suit :

Index = f(Base d'images, Attributs image, Méthode de composition)

Pour une base d'images, nous devons calculer des attributs sur chaque image et utiliser une méthode de composition de ces attributs pour produire un index valable. Ensuite, si nous utilisons la même procédure pour une nouvelle base d'images, nous espérons obtenir un nouvel index qui permettra de comparer les deux bases d'images, et aussi ensuite les résultats d'algorithmes travaillant sur ces deux bases d'images. Pour cela, les attributs utilisés et la méthode de composition doivent être fixés et ne pas changer, car autrement la comparaison des index devient impossible. Comme nous nous intéressons au domaine de la recherche d'images similaires, nous avons choisi les attributs les plus fréquents pour ce domaine que sont la couleur et la texture [SME 00]. Nous avons ensuite choisi comme méthode de composition de l'index le calcul par arbres de décision, qui constitue une méthode simple, reproductible, indépendante, mais présentant des similarités avec le domaine ciblé. Nous avons expérimenté la méthode à base d'arbres de décision décrite à la section 2.2 sur les bases d'images de la section 2.1 et calculé différents index possibles (tableau 1). Parmi ces mesures, l'entropie est une mesure d'incertitude, qu'on peut aussi décrire comme la quantité d'information de la base de test. La mesure du gain exprime la quantité d'information que l'arbre a acquise sur la base de test après la base d'apprentissage, tandis que le taux de gain d'information exprime le gain corrigé par l'entropie. Toutes ces mesures permettent de caractériser une base d'images, principalement le **taux de bonne classification** (en %) et le **taux de gain d'information** (en %). Ces deux mesures sont des index adéquats pour caractériser les différentes bases d'images. Plus le taux est élevé pour une base d'images et plus cette base est considérée facile (uniquement pour le problème de la recherche d'images) et inversement.

Mesure	Calcul	Wang	Coil-100	CUReT	INRIA
Taux de bonne classification (%)	$\frac{\text{Nombre d'exemples bien classifiés}}{\text{Nombre d'exemples}}$	64.5	93.46	52.25	67.07
Entropie (bit)	Entropie de la base de test	3.32	6.64	6.13	4.89
Gain (bit)	Entropie de la base de test – Entropie cond. par l'arbre de la base de test	1.61	6.19	3.14	3.22
Taux de gain d'information (%)	$\frac{\text{Gain}}{\text{Entropie}} \times 100\%$	48.45	93.24	51.30	65.81

Tableau 1. Différentes mesures caractéristiques sur les quatre bases d'images différentes.

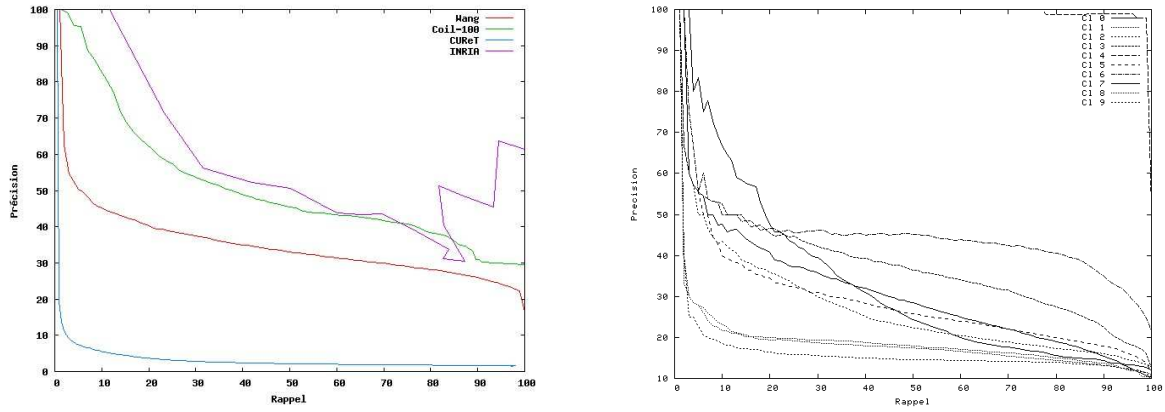
Mesure	Cl 0	Cl 1	Cl 2	Cl 3	Cl 4	Cl 5	Cl 6	Cl 7	Cl 8	Cl 9
Taux de bonne classification des exemples de classe i (%)	50.63	43.75	33.75	53.75	100.0	60.00	91.88	80.63	36.25	56.88
Entropie (bit)	0.469	0.469	0.469	0.469	0.469	0.469	0.469	0.469	0.469	0.469
Gain (bit)	0.100	0.064	0.029	0.118	0.469	0.143	0.345	0.253	0.044	0.122
Taux de gain d'information (%)	21.42	13.62	6.16	25.09	100.0	30.55	73.59	54.04	9.49	26.03

Tableau 2. Différentes mesures caractéristiques sur les classes d'images de la base d'images de Wang.

Dans la même veine, alors qu'on ne s'intéresse très souvent qu'au résultat d'un algorithme sur une base d'images, la structure, ou l'équilibre, de cette base d'images, et de ses classes, est intéressante à analyser parce qu'elle peut nous renseigner sur les résultats que l'on peut en tirer (tableau 2). Dans l'exemple donné dans le tableau 2, les classes 4 et 6 (contenant des dessins de dinosaures et des fleurs en gros plans – visuellement très proches) sont de toute évidence plus faciles, tandis que les classes 2 et 8 (contenant des monuments urbains et des paysages de montagnes – visuellement très différents) posent plus de difficultés. Cependant, l'interprétation de ces observations n'est pas universelle et dépend des images, de la structure définie ou non dans les classes, de l'application et des objectifs visés.

La figure 1 montre les résultats de rappel vs précision de l'algorithme en recherche d'images par le contenu (section 2.3) sur les bases d'images. Plus la courbe est élevée et plus le résultat est bon. L'ordre des courbes est à comparer avec les résultats de la méthode de classification pour déterminer l'ordre des bases d'images, de la plus « facile » (courbe élevée) à la plus « difficile ». La figure 1a comparée au

tableau 1 permet de juger de la qualité des index proposées, tandis que la figure 1b comparée au tableau 2 permet de faire de même pour les classes individuelles d'une même base d'images.



(a) **(b)**

Figure 1. Résultats de rappel vs précision en recherche d'images par le contenu. (a) Courbes globales pour les quatre bases d'images (Note : pour la base INRIA, les valeurs erronées pour des hautes valeurs de rappel sont dues à la faible taille de certaines classes). (b) Courbes pour chaque classe de la base d'images de Wang.

4 Conclusion

Nous avons présenté dans cet article une nouvelle approche pour aider l'évaluation et la comparaison des systèmes d'indexation et recherche d'images par le contenu, en utilisant une technique de classification à base d'arbres de décision pour classifier les bases d'images existantes et les ordonner. En publiant les résultats de leurs travaux sur une base d'images particulière, les auteurs pourraient ainsi donner en même temps les différents index associés à leur base afin de permettre une meilleure comparaison des résultats. Cette approche est selon nous possible dans le domaine de la recherche d'images par le contenu parce que les travaux dans ce domaine se font très souvent sur des applications générales et des bases d'images non-spécifiques, ce qui justifie le choix d'attributs d'images généraux comme la couleur et la texture. Il reste à étudier le cas où plusieurs classes (ou mots-clés) sont associées à chaque image.

5 Bibliographie

[BON 02] BONTON P., BOUCHER A., THONNAT M., TOMCZAK R., HIDALGO P.J., BELMONTE J., GALAN C., « Colour image in 2D and 3D microscopy for the automation of pollen rate measurement », *Image Analysis and Stereology*, vol. 21, n° 1, march 2002, pp. 25-30.

[DAN 04] DANG T.H., BOUCHON-MEUNIER B., MARSALA C., « Measures of information for inductive learning », *Information Processing and Management of Uncertainty in Knownledge-Based Systems (IPMU)*, july 2004, pp. 1495-1502, Perugia (Italia).

[JER 02] JERMYN I.H., SHAFFREY C.W., KINGSBURY N.G., « Evaluation Methodologies for Image Retrieval Results », *Proc. of Advanced Concepts for Intelligent Vision Systems (ACIVS)*, 2002.

[LE 04] LE T.L., BOUCHER A., « An interactive image retrieval system: from symbolic to semantic », *Int. Conf. on Electronics, Informations and Communications (ICEIC)*, august 2004, Hanoi (Vietnam).

[MAR 98] MARSALA C., « Apprentissage inductif en présence de données imprécises : Construction et utilisation d'arbres de décision flous », Thèse de doctorat de l'Université Paris VI, 1998.

[SME 00] SMEULDERS A.W.M., WORRING M., SANTINI S., GUPTA A., JAIN R., « Content-Based Image Retrieval at the End of the Early Years », *IEEE Trans. on Patt. Anal. and Machine Intell.*, vol. 22, n° 12, 2000, pp. 1349-1380.

Analyse en composantes principales et analyse discriminante de densités de probabilité dans l'environnement R

R. Boumaza[*], P. Guillermin[*], P. Revollon[**], L. Durandet[**]

(*) UMR SAGAH
(**) Département Economie, Traitement de l'Information et Communication
Institut National d'Horticulture
2 rue Le Nôtre
Angers, France, 49045

RÉSUMÉ. Après un rappel de l'analyse en composantes principales et de l'analyse discriminante de densités de probabilité, méthodes traitant des tableaux de données à 3 voies, deux fonctions FPCAD et FDAD dans l'environnement R seront présentées et illustrées sur la base d'exemples.

MOTS-CLÉS : données ternaires, ACP, AD, estimation de densités.

1 Introduction

On considère des données ternaires " *individus* x *variables* x *occasions* " (Tableau 1) où à chaque occasion t (t=1,…,T), on observe les p mêmes variables quantitatives sur un lot de n_t individus. Ces n_t observations sont considérées comme un n_t – échantillon d'un vecteur aléatoire à valeurs dans ∇^p, de densité de probabilité f_t (par rapport à la mesure de Lebesgue), qui permet d'estimer cette densité.
L'objectif de l'analyse en composantes principales (ACP) fonctionnelle de densités de probabilité est de visualiser ces occasions, via les densités associées, sur un sous-espace de dimension réduite afin d'apprécier les différences et ressemblances entre lots d'individus. Si t fait référence au temps, cette ACP permettra d'apprécier qualitativement l'évolution des lots d'individus.
Aux données précédentes, on ajoute une variable qualitative G à Q modalités (Tableau 2), définie sur l'ensemble des T occasions. Un nouveau lot T+1 de n individus sur lesquels on a observé les p variables quantitatives, se présente ; on cherche à prédire la valeur de G pour ce nouveau lot. C'est l'objectif de l'analyse discriminante (AD) de densités ([BOU 04]).

2 La fonction FPCAD

L'ACP des densités f_t (t = 1,…,T) dans l'espace $L^2(\nabla^p)$ ([BOU 98, KNE 01]) permet d'obtenir la décomposition des f_t suivant un système orthonormé (h_k) :

$$f_t = \sum_k \alpha_{kt} h_k .$$

Cette décomposition permet ainsi la visualisation de ces densités respectant au mieux les distances entre ces densités ; ces distances sont calculées en s'appuyant sur la norme L^2 induite par le produit scalaire classique de $L^2(\nabla^p)$: $\langle f_1,f_2\rangle = \int f_1(x)\, f_2(x)\, dx$.
Dans le cas de densités gaussiennes, on peut trouver en [BOU 98] des relations entre cette ACP de densités et la première étape, ou étape de l'interstructure, de la méthode STATIS duale.

La fonction FPCAD de l'environnement R qui réalise cette ACP de fonctions de densités, calcule les coordonnées, les aides à l'interprétation classiques (contributions et qualités) suivant chaque axe principal, et réalise les représentations graphiques.
Les options classiques de l'ACP : centrage ou réduction, peuvent être sélectionnées. Les densités peuvent être considérées soit gaussiennes et estimées paramétriquement, soit quelconques et estimées par la méthode du noyau. Pour la méthode du noyau on utilise le noyau gaussien avec la fenêtre AMISE :

$$w=(4 / (n(p+2)))^{1/(p+4)}$$ ([SIL 86])

qui minimise une approximation de l'erreur quadratique moyenne intégrée (MISE).
Les résultats de l'ACP de densités, avec ces différentes options, seront illustrés sur la base d'un jeu de lots de pommes dont on étudie la texture.

		X_1 X_2 … X_p
Lot 1	Individu (1,1) • • • Individu (1, n_1)	
• • •		
Lot t	Individu (t,1) • • • Individu (t, n_t)	
• • •		
Lot T	Individu (T,1) • • • Individu (T, n_T)	

Tableau 1. Données " *Individus* x *Variables* " où les individus sont divisés en lots (occasions). Les variables sont quantitatives.

3 La fonction FDAD

L'AD de densités de probabilité permet d'affecter le lot T+1 à l'une des modalités de la variable G. Dans [BOU 04], il a été proposé des règles géométriques et des règles probabilistes. Comme en ACP de densités, à chaque lot t est associée une densité f_t ; de plus à chaque modalité q de G est associée une densité g_q qu'on estime à partir des lots qui prennent cette modalité. Les règles géométriques affectent le lot T+1 de densité f à la modalité la plus proche au sens d'une mesure de dissimilarité. Les règles probabilistes quant à elles, supposent que les densités f_t et g_q sont gaussiennes et calculent une probabilité d'affectation du lot à chaque modalité, l'affectation pouvant alors se faire à la modalité pour laquelle la probabilité est maximum.
La fonction FDAD de l'environnement R qui réalise les calculs des mesures de dissimilarité ou probabilités comporte plusieurs options. Elle est paramétrée selon le type de données disponibles et la règle d'affectation à utiliser.

- Le type de données :
  - Le nombre T_q de lots par modalité q : $T_q>1$ ($\forall q$) ou $T_q=1$ ($\forall q$).

- Les densités $(f_t)_{t=1,T}$ *et* $(g_q)_{q=1,Q}$ respectivement associées aux T lots et aux Q modalités sont considérées gaussiennes ou non.

- Le choix de la règle d'affectation tient compte du type de données et doit préciser :
  - La méthode d'estimation des densités : méthode paramétrique ou méthode du noyau (non paramétrique), et pour cette dernière méthode le type de fenêtre de lissage : fenêtre AMISE par densité ou fenêtre commune à toutes les densités.
  - La mesure de dissimilarité entre densités : distance L^2, distance L^2 après normalisation des densités, distance de Matusita ou divergence de Kullback-Leibler.
  - Le critère utilisé : géométrique ou probabiliste.

Certains choix parmi les options décrites sont bien entendu incompatibles. Le tableau 3 présente les choix possibles.

Dans le cas où on dispose de plusieurs lots par modalité ($T_q>1$, $\forall q=1,\ldots,Q$), les différents règles d'affectation peuvent être comparées sur la base des taux de bon classement obtenus par validation croisée. Ce qui permet de discuter le choix d'une " meilleure " option pour l'affectation du lot $T+1$.

L'AD de densités sera illustrée sur des données simulées.

	X_1 X_2 . . . X_p	G
Lot 1		**1**
• • •		• • •
Lot T_1		**1**
	• • •	
Lot $T_1+\ldots+T_{q-1}+1$		**q**
• • •		• • •
Lot $T_1+\ldots+T_{q-1}+T_q$		**q**
	• • •	
Lot $T_1+\ldots+T_{Q-1}+1$		**Q**
• • •		• • •
Lot T $(= T_1 +\ldots + T_Q)$		**Q**
Lot $T+1$	**CONNU**	**?**

Tableau 2. La variable qualitative G, définie sur l'ensemble des T lots, est à Q modalités. Chaque lot est composé de plusieurs individus.

<table>
<tr><td colspan="2">Données</td><td colspan="3">Règle d'affectation</td></tr>
<tr><td>Nombre de lots par modalité</td><td>Hypothèse de normalité</td><td colspan="2">Méthode d'estimation des densités</td><td>Mesure de dissimilarité</td><td>Critère</td></tr>
<tr><td rowspan="3">$T_q > 1$ $(\forall q)$</td><td rowspan="2">Non</td><td rowspan="2">Méthode du noyau</td><td>Fenêtre AMISE par densité</td><td rowspan="3">L^2
L^2 normalisée</td><td rowspan="3">• Géométrique
• g_g est la moyenne des densités f_t qui prennent la modalité q</td></tr>
<tr><td>Fenêtre commune</td></tr>
<tr><td>Oui</td><td colspan="2">Paramétrique</td></tr>
<tr><td rowspan="4">$T_q = 1$ $(\forall q)$</td><td>Non</td><td>Méthode du noyau</td><td>Fenêtre AMISE</td><td>Géométrique</td></tr>
<tr><td rowspan="3">Oui</td><td colspan="2" rowspan="3">Paramétrique</td><td>Kullback-Leibler
Matusita
L^2 normalisée</td><td>Géométrique</td></tr>
<tr><td>L^2</td><td>Géométrique ou probabiliste</td></tr>
<tr><td colspan="2">Règle quadratique (généralisée)</td></tr>
</table>

Tableau 3. Compatibilité des différents choix possibles en analyse discriminante de densités.

4 Bibliographie

[BOU 98] BOUMAZA R., "Analyse en composantes principales de distributions gaussiennes multidimensionnelles", *Revue de Statistique Appliquée,* vol. XLVI, n° 2, 1998, p. 5-20.

[BOU 04] BOUMAZA R., "Discriminant analysis with independently repeated multivariate measurements: an L^2 approach", *Computational Statistics & Data Analysis*, vol. 47, 2004, p. 823-843.

[KNE 01] KNEIP A., UTIKAL K.J., "Inference for density families using functional principal component analysis", *Journal of the American Statistical Society*, vol. 96, n° 954, 2001, p. 519-542.

[R 04] R DEVELOPMENT CORE TEAM, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, 2004. URL: http://www.R-Project.org.

[SIL 86] SILVERMAN B.W., *Density estimation for statistics and data analysis*, Chapman & Hall, 1986.

Arbre de régression multivariable : application à une communauté de poissons littoraux d'un lac du Bouclier canadien

Anik Brind'Amour

Département de sciences biologiques,
Université de Montréal, Pavillon Marie-Victorin,
Case postale 6128, succursale Centre-ville,
Montréal (Québec) Canada, H3C 3J7

RÉSUMÉ. La combinaison de deux approches statistiques, une analyse de partitionnement et la technique de l'arbre de régression mutivariable, a été utilisée afin de développer un modèle d'utilisation d'habitats pour une communauté de poissons d'un lac du Bouclier canadien. Les résultats soulignent l'importance des variables dites associées dans le modèle et l'importance de la taille de l'habitat dans la succession d'autres variables environnementales structurantes et dans la détermination de la composition spécifique de la communauté de poissons du lac à l'étude.

MOTS-CLÉS : Arbre de régression mutivariable, communauté piscicole, habitat, lac, partitionnement par la méthode des K centroïdes (K-means).

1 Introduction

La zone littorale représente l'environnement physique le plus hétérogène, diversifié et productif des lacs. Avec sa variété de structures physiques (p. ex. débris de bois, substrats, macrophytes émergentes et submergées) et de ressources alimentaires, la zone littorale se caractérise par une mosaïque de micro-habitats [BOI 01]. Leur configuration spatiale et leur diversité confèrent à la zone littorale une place de choix car il s'y produit de nombreuses interactions intra et interspécifiques [WER 77]. Les communautés de poissons sont donc exposées à un environnement complexe structuré à plusieurs échelles spatiales variant du millimètre (p. ex. interstices dans les substrats rocheux) à une centaine de mètres (p. ex. distance entre deux tributaires ou entre lits de macrophytes). Par conséquent, les interactions entre les espèces de poissons et l'environnement ont le potentiel d'être fortement spatialisées. Toutefois, les modèles d'habitat développés en zone littorale incorporent rarement des données se référant à la dimension spatiale de cette dernière. De plus, ces modèles portent généralement sur des descripteurs généraux de la communauté (p. ex. abondance totale) qui ne se soucient guère de l'identité (espèce) des membres de la communauté. Dans ce contexte, le présent travail propose de développer un modèle d'utilisation d'habitats portant sur l'ensemble des espèces d'une communauté de poissons littoraux d'un lac du Bouclier canadien. Ce modèle est développé à l'aide de l' arbre de régression mutivariable [DEA 00] à l'échelle des habitats dans le lac à l'étude.

2 Méthodes

2.1 Lac à l'étude

Les données ayant servies aux analyses de ce travail ont été recueillies au lac Paré (46°08' N; 73°54'W), un lac mésotrophe de 31 ha situé dans la région de Lanaudière (Québec). Le lac Paré est caractérisé par

une zone littorale diversifiée (roche, plantes aquatiques, troncs d'arbres morts, etc.) et a une profondeur maximale de 9 m. La zone littorale est peuplée par une communauté de poissons comportant huit espèces représentant cinq familles. La zone littorale du lac Paré a été divisée en 60 sites d'échantillonnage (~ 200 m^2) dans lesquels des données biologiques (abondance de poissons) et des variables environnementales ont été estimées.

2.2 Analyse de partitionnement

Un partitionnement des sites d'échantillonnage par la méthode des *K* centroïdes (*K-means*) a été réalisé afin d'identifier les différents habitats présents dans le lac à l'étude. Le nombre de groupes optimal fut déterminé à l'aide de la statistique de Calinski-Harabasz utilisant une solution minimisant la variation intragroupe, pour K = 2 à K = 10 groupes, après 100 attributions aléatoires des objets aux groupes initiaux. La confirmation visuelle des groupes de sites similaires a été obtenue à l'aide d'une analyse en composantes principales (ACP) calculée sur la matrice des corrélations entre les variables environnementales. À la suite de cette procédure, les sites contigus appartenant à un même habitat ont été groupés. L'abondance de chaque espèce de poissons dans les sites regroupés a été sommée, puis divisée par la superficie du groupe de sites afin de contrer l'effet additif engendré par la procédure de groupement des sites. La moyenne simple des valeurs des variables environnementales a également été utilisée dans les analyses statistiques. L'analyse de partitionnement et l'ACP ont été réalisées à l'aide du Progiciel R [CAS 05].

2.3 Arbre de régression mutivariable

Un modèle d'utilisation d'habitats de poissons a été développé à l'aide de l'approche des arbres de régression mutivariable. Cette méthode permet notamment de modéliser les données écologiques caractérisées par i) des relations non linéaires entre les variables réponses (abondance d'espèces) et les variables explicatives (caractéristiques environnementales), ii) une distribution non unimodale des variables réponses et iii) des devis expérimentaux non balancés (c.-à-d., groupes de différentes tailles). Les analyses ont été effectuées à l'aide de la fonction fournie par Ouellette et coll. [OUE en prép.], une version modifiée de la fonction de De'Ath [DEA 00], exécutée dans le langage R [R 04]. La composition relative des espèces à chaque partition de l'arbre a été estimée à l'aide d'une méthode de recherche des espèces indicatrices [DUF 97] qui s'est avérée complémentaire aux arbres de régression. La recherche des espèces indicatrices a été réalisée à chaque noeud de l'arbre de régression élaboré précédemment.

3 Résultats

3.1 Habitats littoraux du lac Paré

Le partitionnement a permis de grouper en quatre types d'habitats (groupes non contigus sur le pourtour du lac) les 60 sites d'échantillonnage du Lac Paré (Fig. 1). Les groupes de sites se distinguent principalement par le degré d'amplitude de la pente, le degré d'exposition au vent dominant (fetch), la densité de macrophytes submergées et émergentes, le type d'utilisation riveraine (chalet ou forêt) et la présence de sable comme substrat. Les groupes de sites sont distribués en taches de différentes tailles variant entre 120 m^2 et 1755 m^2, représentant ainsi une mosaïque d'habitats le long du littoral du lac Paré. La superficie totale des quatre habitats se répartit comme suit : Habitat O (56,9 %), Habitat Δ (26,8 %), Habitat □ (10,4 %), Habitat ◊ (5,9 %). Le partitionnement des sites fut confirmé par l'ACP et les trois premières dimensions de l'ordination représentaient 48 % de la variation totale des variables environnementales centrées réduites aux 60 sites.

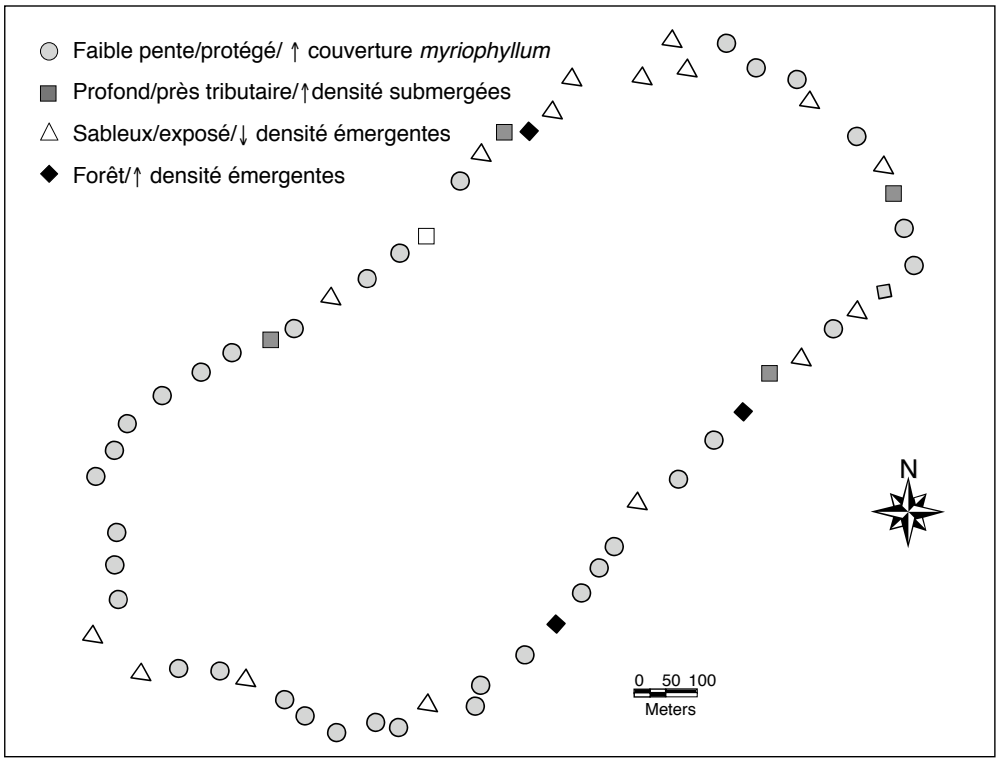


Figure 1: Carte du Lac Paré présentant le partitionnement des sites d'échantillonnage en quatre types d'habitat. La légende fournit les caractéristiques environnementales dominantes pour chaque type d'habitat.

3.2 Modèle multivariable d'utilisation des habitats par les poissons

L'arbre de régression multivariable développé à l'aide des densités de poissons de la communauté littorale du lac Paré est composé de six nœuds et explique 48.3 % de la variation totale de la composition en espèces de la communauté du lac (Fig. 2). Dans cet arbre, les deux premiers nœuds sont associés à des variables environnementales dites « associées », c'est-à-dire estimées à l'extérieur de l'unité d'échantillonnage et qui se réfèrent à la spatialité des habitats. La distance à un tributaire et l'exposition au vent dominant (fetch) sont les deux variables qui contrôlent les premiers nœuds de l'arbre. La taille d'un habitat semble être également un facteur déterminant du modèle puisque cette variable contrôle le troisième nœud qui permet de séparer en deux groupes 85 % (29/34) des taches d'habitat de la zone littorale du lac. L'arbre indique que lorsque la taille d'un habitat est supérieure à 196 m^2, la présence d'une forêt riveraine et la distance qui sépare cet habitat d'un habitat similaire deviennent alors des facteurs déterminants dans la structure de la communauté autour du lac. La composition relative (%) en espèces des trois partitions subséquentes au nœud associé à la taille de l'habitat est similaire ; ces trois nœuds ne sont pas conservés par validation croisée. La communauté de ces 5 groupes est représentée majoritairement par les cyprinidés (Seat_S, Copl_S, Copl_L) et les perchaudes de grande taille (Pefl_L).

Lorsque la taille de l'habitat est inférieure à 196 m^2, la complexité de l'habitat, ici associée à la densité de macrophytes émergentes, devient un facteur déterminant de la structure de la communauté de poissons. La communauté est alors majoritairement composée de cyprinidés et des poissons de petite taille. Toutefois, les poissons de petite taille des espèces benthiques adultes (Amne_S et Caco_S) montrent des abondances relatives de 10 % supérieures dans les habitats arborant des densités de macrophytes émergentes inférieures à 6.5 tiges$\cdot m^{-2}$ par rapport aux habitats exposant des densités de macrophytes émergentes supérieures à 6.5 tiges$\cdot m^{-2}$.

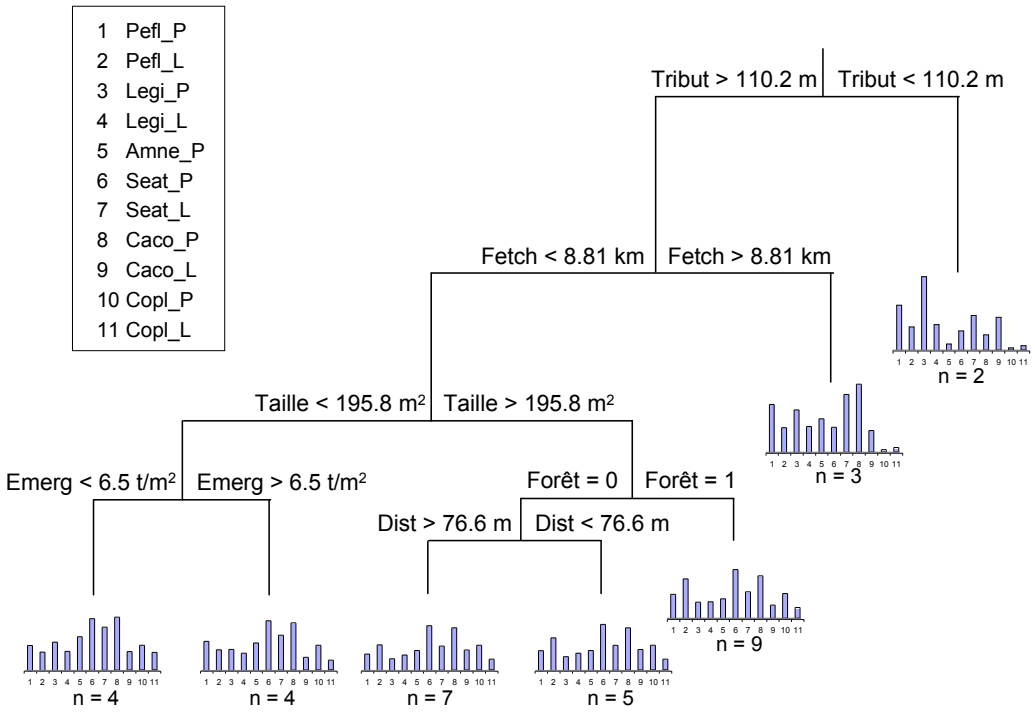


Figure 2 : Arbre de régression multivariable développé pour la communauté de poissons littoraux du lac Paré. P: petite taille; L: grande taille; Pefl: *Perca flavescens*; Legi: *Lepomis gibbosus*; Amne: *Amiurius nebulosus*; Seat: *Semotilus atromaculatus*; Caco: *Catostomus commersoni*; Copl: *Couesius plumbeus.*

4 Discussion

Les approches statistiques basées sur l'étude des communautés d'espèces (c.-à-d., approches mutivariables) représentent une avancée de taille dans le développement de modèles d'utilisation des habitats par les poissons. Dans le présent travail, la technique des arbres de régression multivariable a permis de considérer l'identité spécifique des membres d'une communauté dans l'élaboration du modèle d'utilisation d'habitat d'une communauté piscicole et, par conséquent, de définir implicitement les relations fonctionnelles des espèces face aux caractéristiques environnementales des habitats [OLD 04]. Les résultats ont montré l'importance notamment de la taille d'un fragment d'habitat comme facteur déterminant dans la structure des communautés littorales de poissons lacustres.

Parallèlement à l'intérêt des études multivariables des communautés piscicoles, ce travail souligne l'importance d'incorporer des informations portant sur la spatialité des habitats de poissons dans les lacs. Quatre des six nœuds de l'arbre de régression multivariable impliquaient des variables associées. Des conclusions similaires ont été obtenues récemment par Brind'Amour [BRI soumis]. Ces derniers ont mis en évidence l'importance des variables dites associées dans la performance des modèles d'utilisation d'habitat.

Alors que la zone littorale des lacs est de plus en plus perçue comme un paysage composé de multiples habitats de taille et de qualité variables, ce travail montre que la combinaison de méthodes statistiques telles les méthodes d'analyses multivariables offre un cadre spatialisé très intéressant pour améliorer l'étude des relations poisson-habitat dans la zone littorale des lacs.

5 Remerciements

Je remercie P. Legendre pour la révision de ce travail et M.-H. Ouellet pour m'avoir permis d'utiliser sa fonction en langage R. Merci à M. Coinçon et J. Guimmond-Cataford pour leur aide lors de l'échantillonnage au lac Paré. Le support financier a été fourni par D. Boisclair et par les bourses d'études octroyées à A. Brind'Amour par le CRSNG et le FCAR.

6 Bibliographie

[BOI 01] BOISCLAIR D., "Fish habitat modelling: from conceptual framework to functional tools". *Canadian Journal of Fisheries and Aquatic Sciences*, vol. 58, n° 1, 2001, p. 1-9, NRC Research press, Ottawa.

[BRI prep] BRIND'AMOUR A., BOISCLAIR D., "The effect of the spatial arrangement of habitats on the development of fish habitat models in the littoral zone of a Laurentian lake", (soumis à CJFAS).

[CAS 05] CASGRAIN P., LEGENDRE P., *The R Package for multivariate and spatial analysis, Version 4.0 (development release 10) – User's manual*, Département de sciences biologiques, Université de Montréal, available from <http://www.bio.umontreal.ca/legendre/>, 2005.

[DEA 00] DE'ATH G., FABRICIUS K.E., "Classification and regression trees: A powerful yet simple technique for ecological data analysis". *Ecology,* vol. 81, n° 11, 2000, p. 3178-3192, ESA, Ithaca.

[DUF 97] DUFRÊNE M., LEGENDRE P., "Species assemblages and indicator species: The need for a flexible asymetrical approach". *Ecological Monographs,* vol. 67, n° 3, 1997, p. 345-366, ESA, Ithaca.

[OLD 03] OLDEN J.D., "A species-specific approach to modeling biological communities and its potential for conservation". *Conservation Biology,* vol. 17, n° 3, 2003, p. 854-863, Blackwell Publishing, Gainesville.

[OUE prep] OUELLETTE M.-H., DESGRANGES J.-L., LEGENDRE P., BORCARD D. "Multivariate regression tree analysis: bird assemblages from the Saint Lawrence corridor", (en préparation).

[R 04] R DEVELOPMENT CORE TEAM, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, 2004.

[WER 77] WERNER E.E., HALL D. J.D., LAUGHLIN R., WAGNER D.J., WILSMANN L. A., FUNK, F.C. "Habitat partitioning in a freshwater fish community". *Journal of Fisheries and Research Board of Canada,* vol. 34, 1977, p. 360-370, NRC Research press, Ottawa.

Modèles VL en Classification Non-Hiérarchique

Paula Brito* - Fernanda Sousa** - Sara Tavares Pinto**

**Faculdade de Economia /LIACC, Universidade do Porto*
Rua Dr. Roberto Frias
4200-464 Porto, Portugal
mpbrito@fep.up.pt

***Faculdade de Engenharia /CEC, Universidade do Porto*
Rua Dr. Roberto Frias
4200-465 Porto, Portugal
fcsousa@fe.up.pt ; saratavares@net.sapo.pt

RÉSUMÉ. Dans ce papier on s'intéresse à l'application de la méthodologie de la Vraisemblance du Lien (VL) en classification non-hiérarchique. On propose des méthodes de classification non-hiérarchique qui utilisent la distribution de statistiques de tendance centrale telles que la médiane, le centre ou la moyenne géométrique. Dans chaque cas, la détermination des centres initiaux ainsi que la mise à jour des centres à chaque itération font appel aux statistiques correspondantes. Des études de simulation permettent d'évaluer et comparer le comportement des différentes méthodes.

MOTS-CLÉS : Classification, Partition, Vraisemblance du Lien

1 Introduction

Les méthodes de classification non-hiérarchique visent la détermination de partitions en classes homogènes et bien-séparées, usuellement basées sur une mesure de similarité ou de dissimilarité entre paires d'éléments. Lerman [LER 70] a proposé, dans le cadre de la classification hiérarchique, l'utilisation d'une mesure de probabilité pour l'évaluation de la liaison entre éléments ou classes - approche Vraisemblance du Lien. Cela a conduit à la méthode AVL, qui utilise comme fonction de comparaison entre classes la fonction de répartition du maximum des similarités entre éléments. Bacelar-Nicolau [BAC 72] a étudiée la performance de cette méthode et Nicolau [NIC 80] a proposé la méthode AVM qui utilise la fonction de répartition de la moyenne arithmétique des valeurs de comparaison. Plus récemment des méthodes de classification hiérarchique basées sur la médiane, le centre et la moyenne géométrique ont été proposées [SOU 00].

Dans le cadre de la classification non-hiérarchique, Nicolau et Brito [NIC 85, NIC 89] ont considéré comme critère d'affectation des éléments aux classes la fonction de répartition de la moyenne arithmétique des valeurs de comparaison entre l'élément à affecter et les membres de la classe.

Dans ce papier, on propose de nouvelles méthodes de classification non-hiérarchique, qui utilisent comme critère d'affectation des éléments aux classes la fonction de répartition de statistiques de tendance centrale, à savoir, la médiane, le centre et la moyenne géométrique, des valeurs de comparaison entre l'élément à affecter et les membres de la classe.

La Section 2 détaille la méthodologie VL. Dans la Section 3, on présente les méthodes proposées. La Section 4 détaille la méthode utilisée pour la détermination des centres initiaux, dénommée méthode *wise-choice*. Dans la Section 5 on présente une étude de simulation, permettant d'évaluer la performance des différentes méthodes sur des données avec des configurations variables. Enfin, dans la Section 6 on conclut, en présentant des perspectives de développements futurs.

2 La Méthodologie VL

La méthodologie VL peut être résumée comme suit. Suivant les caractéristiques de l'ensemble E à classifier, on utilise un indice de similarité ou dissimilarité entre paires d'éléments, soit $c : E \times E \to \Re_0^+$. Les $m(m-1)/2$ valeurs $c(i,j)$ sont considérées des réalisations d'une variable aléatoire Z, et la fonction de comparaison entre éléments est donnée par: $\gamma(i,j) = P(Z \le c(i,j)) = F_z(c(i,j))$, si c est une mesure de similarité ou $\gamma(i,j) = P(Z \ge c(i,j)) = 1 - F_z(c(i,j))$, si c est une mesure de dissimilarité. Si l'on admet l'hypothèse d'indépendance entre les éléments de E, alors l'ensemble des valeurs $\gamma(i,j)$ constitue un échantillon de $m(m-1)/2$ variables aléatoires indépendantes et identiquement distribuées d'une loi uniforme dans l'intervalle $[0,1]$.

La fonction de comparaison entre sous-ensembles de E, $\Gamma : P(E) \times P(E) \to \Re_0^+$ est définie de la façon suivante. Étant données deux classes A et B de E, de cardinaux respectivement n_A et n_B, on choisit une statistique U des valeurs $\{\gamma(a,b) \mid (a,b) \in A \times B\}$; la fonction Γ, dans le cas où $\gamma(a,b)$ est une similarité, est alors définie par $\Gamma(A,B) = P(U \le u_{A,B})$. La conversion est simple pour le cas de dissimilarités.

3 Méthodes VL en Classification Non-Hiérarchique

Les méthodes de classification non-hiérarchique déterminent une partition de l'ensemble E en un nombre k de classes, souvent fixé au préalable. Les méthodes de type nuées dynamiques, partant d'un ensemble de k centres initiaux, procèdent par l'application itérative d'une fonction d'affectation suivie d'une fonction de représentation jusqu'à convergence (ou que le nombre maximum d'itérations soit atteint). La fonction d'affectation détermine une partition en k classes étant donnés k centres; la fonction de représentation détermine les k représentants des classes, étant donnée une partition en k classes.

L'application de la méthodologie VL dans ce contexte consiste à considérer comme fonction d'affectation d'un élément x à une classe A la fonction $\Gamma(\{x\}, A) = P(U \le u_{\{x\},A})$, pour une statistique U des valeurs de similarité entre x et les éléments de la classe A. La fonction de représentation détermine comme centre c_A d'une classe A l'élément de A tel que $U(\{c_A\}, A \setminus \{c_A\}) = Max\{U(\{x\}, A \setminus \{x\}), x \in A\}$.

Le cas où U est la moyenne arithmétique des valeurs de similarité entre x et les éléments de chaque classe A a donné lieu à la définition de la méthode NHMEAN [NIC 85, NIC 89]. Dans ce travail, on étudie les méthodes qui résultent de considérer, pour la statistique U, la médiane (méthode NHVMED), le centre (méthode NHVC) et la moyenne géométrique (méthode NHVGM). Dans chaque cas, les fonctions $\Gamma(\{x\}, A) = P(U \le u_{\{x\},A})$ qui définissent les fonctions d'affectation sont, respectivement, les suivantes :

– Moyenne arithmétique: $\Gamma(\{x\},A) = P(U \le u_{\{x\},A}) = \frac{1}{n_A!} \sum_{j=0}^{n_A u_{\{x\},A}} (-1)^j \binom{n_A}{j} (n_A u_{\{x\},A} - j)^{n_A}$

– Moyenne géométrique: $\Gamma(\{x\},A) = 1 - F(-2n_A \ln u_{\{x\},A})$, $F \sim \chi^2(2n_A)$.

– Médiane: si n_A est impair, $\Gamma(\{x\},A) = 1 - F(u_{\{x\},A})$, $F \sim Beta\left(\frac{n_A+1}{2}, \frac{n_A+1}{2}\right)$, sinon

$$\Gamma(\{x\},A) = \begin{cases} 2\sum_{i=0}^{k-1} (-1)^i I_{u_{\{x\},A}}(k+i+1, k-i) & si \quad u_{\{x\},A} \in [0,\ 0.5] \\ 2\sum_{i=0}^{k-1} (-1)^i I_{u_{\{x\},A}}(k+i+1, k-i) + (-1)^k (2u_{\{x\},A} - 1) & si \quad u_{\{x\},A} \in]0.5,\ 1] \end{cases}$$

où $I_x(m,n)$ dénote la fonction Beta incomplète ;

– Centre: $\Gamma(\{x\},A)=\begin{cases}2^{n_A-1}\left(u_{\{x\},A}\right)^{n_A}, & si\ \ u_{\{x\},A}\in[0,\ 0.5]\\ 1-2^{n_A-1}\left(1-u_{\{x\},A}\right)^{n_A}, & si\ \ u_{\{x\},A}\in\ ]0.5,\ 1]\end{cases}$

L'algorithme procède comme suit. On commence par calculer les similarités VL entre les éléments de l'ensemble à classifier. Étant donné le nombre de classes *k*, les *k* centres initiaux, éléments de *E*, sont déterminés par la méthode *wise-choice* (voir Section 4) en utilisant, dans chaque cas, la statistique correspondante des similarités pour évaluer la densité des voisinages. Ensuite, on affecte les autres éléments de *E* à la classe P_j ($j=1,...,k$) pour laquelle la fonction $\Gamma\left(\{x\},P_j\right)=P\left(U\le u_{\{x\},P_j}\right)$, pour la statistique *U* considérée, est maximale. On obtient ainsi une partition de *E* en *k* classes. On calcule alors un nouveau centre pour chaque classe; le centre c_j de la classe P_j sera l'élément de P_j tel que $U\left(\{c_j\},P_j\setminus\{c_j\}\right)=Max\left\{U\left(\{x\},P_j\setminus\{x\}\right),x\in P_j\right\}$, $j=1,...,\ k$. La méthode est itérée, en appliquant la fonction d'affectation suivie de la fonction de représentation jusqu'à stabilisation (ou que le nombre maximum d'itérations soit atteint). Soulignons que la convergence théorique de la méthode n'a pas été prouvée, ce qui ne semble pas immédiat dans le contexte probabiliste.

4 La Procédure d'Initialisation : Méthode *Wise-Choice*

Nicolau [NIC 85] a proposé une méthode d'initialisation - méthode *wise-choice* - qui conduit à prendre comme centres initiaux des éléments de l'ensemble à classifier bien séparées et centrés dans des voisinages denses. Pour chaque $x\in E$, le *L*-voisinage de *x* se définie comme l'ensemble *V(x)* des *L* éléments de *E* les plus proches de *x*. Soit *U* la statistique des valeurs de similarité considérée par la méthode de classification. Le premier centre c_1 sera alors l'élément de *E* qui maximise la valeur de *U* calculée dans son voisinage, $U\left(\{x\},V(x)\setminus\{x\}\right)$. Chacun des centres suivants, c_i, $i=2,...,k$, sera l'élément *x* de *E*, non encore pris comme centre, qui maximise $\dfrac{U\left(\{x\},V(x)\setminus\{x\}\right)}{\min\limits_{1\le j\le i-1}\gamma(x,c_j)}$. On cherche ainsi à obtenir des centres initiaux localisés dans des régions denses, et bien éloignés les uns des autres.

5 Applications

À fin d'évaluer le comportement des différentes méthodes, on a simulé des ensembles de données, en dimension deux, avec des configurations en trois classes de 50 éléments chacune, et différents degrés de séparation. Les valeurs sont simulées selon des distributions gaussiennes, avec des paramètres fixés (Tableau 1); 100 répétitions ont été effectuées dans chaque cas.

Pour chacun des 100 ensembles simulés de chaque configuration, on a appliqué les quatre méthodes considérées, plus la méthode *K-Means*. Le Tableau 2 présente les valeurs des moyennes *m* et écarts-type *s* de l'indice de Rand corrigé [HUB 85] qui compare la partition obtenue par chaque méthode avec la partition de référence, associée à l'ensemble particulier qui est classifié.

	μ_1			μ_2			σ_1^2			σ_2^2			σ_{12}		
	Cl_1	Cl_2	Cl_3	Cl_1	Cl_2	Cl_3	Cl_1	Cl_2	Cl_3	Cl_1	Cl_2	Cl_3	Cl_1	Cl_2	Cl_3
Exemple 1	10	10	-24	-20	10	-24	0.3	0.5	0.3	0.2	0.5	0.2	0	0	0
Exemple 2	0	2	-4	0	5	2	0.3	0.25	0.6	0.5	0.25	0.3	0.3	0	-0.4
Exemple 3	0	3	-5	0	6	3	0.7	0.5	0.9	1.5	0.5	0.8	1	0	-0.7
Exemple 4	0	2	-4	0	5	2	0.7	0.5	0.9	1.5	0.5	0.8	1	0	-0.7
Exemple 5	0	2	-3	0	5	2	1.2	0.9	1	2	0.9	1.2	1.5	0	-1
Exemple 6	0	10	-5	0	10	5	0.3	0.5	0.3	0.2	0.5	0.2	0	0	0

Tableau 1 : Paramètres des distributions associées aux classes simulées

Les résultats ont montré que les différentes méthodes retrouvent bien, en général, les partitions de référence, en un nombre réduit d'itérations. On observe néanmoins un effet de massification pour les méthodes NHMEAN, NHVGM et NHVC, qui favorise la formation de grandes classes, d'où les résultats moins bons pour ces méthodes dans le cas de l'Exemple 6. Ceci s'explique par l'effet des cardinaux des classes dans les fonctions d'affectation $\Gamma\left(\{x\}, P_j\right)$ respectives.

	NHMEAN		NHVGM		NHVC		NHVMED		K-MEANS	
	m	s	m	s	m	s	m	s	m	s
Exemple 1	1	0	1	0	1	0	1	0	1	0
Exemple 2	0.974	0.025	0.946	0.049	0.941	0.131	0.971	0.026	0.990	0.019
Exemple 3	0.968	0.031	0.945	0.064	0.958	0.097	0.956	0.035	0.967	0.033
Exemple 4	0.926	0.047	0.913	0.071	0.947	0.079	0.886	0.074	0.909	0.062
Exemple 5	0.793	0.099	0.766	0.129	0.852	0.15	0.726	0.105	0.783	0.085
Exemple 6	0.636	0.142	0.566	0	0.566	0	0.956	0.04	1	0

Tableau 2 : Moyennes et écarts-type des valeurs de l'indice de Rand corrigé

6 Conclusion et Perspectives

Dans ce papier on a présenté de nouvelles méthodes de classification non-hiérarchique, dans le cadre de la méthodologie VL, qui utilisent comme critère d'affectation des éléments aux classes la fonction de répartition de statistiques de tendance centrale des similarités. L'application de ces méthodes à des données simulées a produit des résultats satisfaisants, trouvant des partitions proches des partitions de référence en un nombre réduit d'itérations. La suite de l'étude concerne le contrôle de l'effet de massification, et l'application de la méthodologie proposée à d'autres statistiques.

7 Bibliographie

[BAC 72] BACELAR-NICOLAU H., *Analyse d'un Algorithme de Classification Automatique*, Thèse de Doctorat de 3ème Cycle, ISUP, Paris VI, 1972.

[LER 70] LERMAN I. C., "Sur l'Analyse des Données préalable à une Classification Automatique", *Mathématiques et Sciences Humaines*, vol. 32, 1970, p. 5-15.

[HUB 85] HUBERT L., ARABIE P., "Comparing Partitions", *Journal of Classification*, vol. 2, 1985, p. 193-218.

[NIC 80] NICOLAU F. C., *Critérios de Análise Classificatória Hierárquica Baseados na Função Distribuição,* Thèse de Doctorat, Faculté de Sciences, Université de Lisbonne, 1980.

[NIC 85] NICOLAU F. C., "Analysis of a Non-Hierarchical Clustering Method Based on VL-Similarity", Rapport de Recherche, Centre de Statistique et Applications / INIC, Lisbonne, nº18, 1985.

[NIC 89] NICOLAU F. C., BRITO P., "Improvements in NHMEAN Method", in *Data Analysis, Learning Symbolic and Numerical Knowledge*, Diday E. (Ed.), Nova Science Publishers Inc., New York, 1989, p. 109-116.

[SOU 00] SOUSA F., *Novas Metodologias em Classificação Hierárquica Ascendente,* Thèse de Doctorat, Faculté de Sciences et Technologie, Université Nouvelle de Lisbonne, 2000.

Inférieures-maximales faiblement hiérarchiques

François Brucker

Département LUSSI,
GET-ENST Bretagne,
Technopôle Brest Iroise
CS 83818
29238 Brest Cedex, France

RÉSUMÉ. Nous présentons dans cette communication une construction algorithmique d'une dissimilarité inférieure-maximale faiblement ultramétrique à partir d'une dissimilarité donnée. La complexité de cet algorithme est de plus polynomiale et est bornée en $O(|X|^4)$ où X est l'ensemble des données.

MOTS-CLÉS : hiérarchies faibles, dissimilarité, inférieure-maximale

1 Introduction

Pour rendre compte de phénomènes comme l'hybridation, les systèmes de classification non-empiétants (partitions et hiérarchies) ne semblent pas adaptés. En effet, l'hybride partageant des caractères des deux parents, il a sa place et dans la classe relative à son « père » et dans la classe relative à sa « mère ». Pour pallier ce problème de nombreux modèles de classification admettant l'empiétance (deux classes A et B sont dites empiétantes si $A \cap B \in \{\emptyset, A, B\}$) ont été développés, en particulier les hiérarchies faibles [BAN 89].

Si l'on appelle *système de classe*s d'un ensemble de données X un sous-ensemble K de 2^X tel que :

- $\emptyset \notin K$,
- $X \in K$,
- Pour tout $x \in X$, $\{x\} \in K$,

Une hiérarchie faible est alors un système de classes K sur X tel que pour toutes classes A, B, C de K : $A \cap B \cap C \in \{A \cap B, A \cap C, B \cap C\}$. Une hiérarchie faible fermée par intersection (si A et B sont des classes, $A \cap B$ l'est également) est appelée quasi-ultramétrique [DIA 98].

Lorsque les données sont décrites par une dissimilarité *d* on a coutume, dans la lignée de Jardine et Sibson [JAR 71] et de Bertrand [BER 00], de lui associer un système de classes K_d contenant les cliques maximales de ses graphes seuils (un graphe seuil de d pour le seuil h étant le graphe contenant toutes les arêtes xy telles que $d(x,y) \leq h$). Ce système de classes pouvant admettre un nombre exponentiel de classes, l'étude directe de K_d est malaisée. On approxime donc habituellement d par une dissimilarité dont on connaît les systèmes de classes (une ultramétrique ou une quasi-ultramétrique par exemple).

On peut montrer [BRU 04] que les dissimilarités d sur X admettant un système de classes faiblement hiérarchiques sont exactement celles telles que pour tous $x,y,z,t_x,t_y,t_z \in X$, les trois inégalités suivantes ne sont pas satisfaites simultanément :

- $d(t_x,x) > \max\{d(y,z), d(t_x,y), d(t_x,z)\}$
- $d(t_y,y) > \max\{d(x,z), d(t_y,x), d(t_y,z)\}$
- $d(t_z,z) > \max\{d(x,y), d(t_z,x), d(t_z,y)\}$

Ces dissimilarités sont appelées *ultramétriques faibles*. De plus, quelque soit la hiérarchie faible K sur X, il existe une dissimilarité d sur X telle que $K=K_d$.

2 Approximation par inférieures-maximales

Lorsque les données sont décrites par une dissimilarité d non faiblement hiérarchique, et que l'on souhaite approximer celle-ci par une dissimilarité w faiblement hiérarchique, plusieurs cas sont possibles :

- Effectuer une approximation optimale (*i.e.* trouver une dissimilarité faiblement hiérarchique minimisant l'écart point à point avec la dissimilarité d'origine), mais cette approche conduit à un problème NP-complet [BAR 01];
- Utiliser une heuristique (par exemple celle développée par Bandelt et Dress [BAN 89]) mais l'on ne peut caractériser le résultat ;
- Effectuer une approximation par inférieures-maximales, c'est-à-dire trouver une dissimilarité faiblement ultramétrique plus petite point à point avec la dissimilarité d'origine et telle qu'il n'existe pas de dissimilarité faiblement ultramétrique à la fois plus grande (point à point) avec la dissimilarité faiblement ultramétrique trouvée et plus petite (point à point) avec la dissimilarité d'origine.

Nous suivrons dans cette communication la troisième approche. On peut en effet montrer que l'ensemble des ultramétriques faibles admet, quelque soit la dissimilarité d'origine, au moins une dissimilarité inférieure-maximale à celle-ci.

Ainsi, la dissimilarité d (tableau 1) admet trois ultramétriques faibles inférieures-maximales (les dissimilarités d1, d2 et d3 du tableau 2. Les différences entre d et ces dissimilarités sont mises en gras)

x	0					
y	1	0				
z	1	1	0			
u	1	1	2	0		
v	2	2	1	2	0	
w	2	1	1	2	2	0
	x	y	z	u	v	w

Tableau 1 : la dissimilarité d

x	0					
y	1	0				
z	1	1	0			
u	1	1	**1**	0		
v	2	2	1	2	0	
w	2	1	1	2	2	0
	x	y	z	u	v	w

x	0					
y	1	0				
z	1	1	0			
u	1	1	2	0		
v	2	**1**	1	2	0	
w	2	1	1	2	2	0
	x	y	z	u	v	w

x	0					
y	1	0				
z	1	1	0			
u	1	1	2	0		
v	2	2	1	2	0	
w	**1**	1	1	2	2	0
	x	y	Z	u	v	w

Tableau 2 : les dissimilarités d1, d2 et d3

La partie suivante montre un algorithme permettant de calculer effectivement une ultramétrique faible inférieure-maximale à une dissimilarité donnée.

3 Algorithme

Soit d une dissimilarité sur X.
On associe à X un ordonnancement $x_1 < x_2 < \ldots < x_n$ tel que : $\min\{ d(x_{i+1},y) \mid y \in X_i\} \leq \min\{ d(u,v) \mid u \neq v \in X_{i-1}\}$, avec $X_i = \{x_1, \ldots x_i\}$. Un tel ordonnancement est toujours possible.

En notant $\partial_i[d](x,y) = \cap_{z \in Xi} \{t \mid d(z,t) \leq \max\{d(x,z), d(y,z), d(x,y)\}, t \in X_i\}$, on peut montrer que d est une ultramétrique faible si et seulement si pour tout i et tous x,y,z de X_i :

$$\partial_i[d](x,y) \cap \partial_i[d](x,z) \cap \partial_i[d](y,z) \cap \{x,y,z\} \neq \varnothing$$

L'algorithme procède alors comme suit :

- Initialisation :
  - $w \leftarrow d$
  - *w restreint à X_3 est une ultramétrique faible.*
- Étape i >3 :
  - *On suppose que w restreint à X_{i-1} est une ultramétrique faible*
  - Tant qu'il existe x,y,z dans X_i tel que : $\partial_i[w](x,y) \cap \partial_i[w](x,z) \cap \partial_i[w](y,z) \cap \{x,y,z\} = \varnothing$
    - Si $x,y,z \in X_{i-1}$ alors :
      - On peut supposer sans perte de généralité que $w(x_i,x) \leq w(x_i,y) < w(x_i,z)$
      - $w(x_i,z) \leftarrow w(x_i,y)$
    - Sinon : on peut supposer sans perte de généralité que $x=x_i$ et que $y,z \in X_{i-1}$
      - S'il existe t_1,t_2 de X_{i-1} tel que $w(t_1,y) > \max\{ w(x_i,z), w(x_i,t_1), w(z,t_1)\}$ et $w(t_2,z) > \max\{ w(x_i,y), w(x_i,t_2), w(y,t_2)\}$ alors :
        - Pour tout t de X_{i-1} tel que $w(t,x_i) > \max\{ w(y,z), w(y,t), w(z,t)\}$: $w(t,x_i) \leftarrow \max\{ w(y,z), w(y,t), w(z,t)\}$
      - Sinon, on peut supposer sans perte de généralité que $w(x_i,y) > w(x_i,z)$: $w(x_i,y) \leftarrow w(x_i,z)$
      - *Les affectations précédentes assurent le fait que maintenant : $\partial_i[w](x,y) \cap \partial_i[w](x,z) \cap \partial_i[w](y,z) \cap \{x,y,z\} \neq \varnothing$*
    -
  - $i \leftarrow i+1$

L'algorithme est ainsi un algorithme glouton qui construit sur les ensembles X_i une ultramétrique faible inférieure à la dissimilarité d'origine. On peut de plus prouver [BRU 05] que cette ultramétrique faible est inférieure-maximale. La complexité de cet algorithme est en $O(|X|^4)$ puisqu'il y a |X| étapes, qu'il faut à chaque étape examiner tous les triplets de X_i et que la mise à jour des ∂_i peut se faire en $O(|X|^3)$.

4 Bibliographie

[BAN 89] BANDELT H.-J., DRESS A. W. M.., " Weak Hierarchies Associated with Similarity Measures - an Additive Clustering Technique ", *Bulletin of Mathematical Biology*, 51,1989.

[BAR 01] BARTHÉLEMY, J-P., BRUCKER, F. "NP-hard Approximation Problems in Overlapping Clustering", *Journal of Classification*, 18, 159-183, 2001.

[BER 00] BERTRAND, P., "Set Systems and Dissimilarities", *European Journal of Combinatorics*, 21, 727-734, 2000.

[BRU 04] BRUCKER, F. "Les ultramétriques faibles", *Actes des rencontres de la SFC*, 123-126.

[BRU 05] BRUCKER, F. "Dissimilarités inférieures-maximales en taxonomie numérique", *Séminaire de Classification*, ENST-Bretagne, France.

[DIA 98] DIATTA, J., FICHET, B. "Quasi-ultrametrics and their 2-balls hyypergraphs", *Discrete Mathematics*, 192, 87-102, 1998.

[JAR 71] JARDINE, N., SIBSON, R. *Numerical Taxonomy*, London : Wiley.

Analyse factorielle multiple sur données mixtes : une application aux données de végétation

Sergio Camiz, Jérôme Pagès

Dipartimento di Matematica Guido Castelnuovo, Università di Roma La Sapienza
Piazzale Aldo Moro, 2 – I 00185 Roma Italie
sergio.camiz@uniroma1.it

Laboratoire de mathématiques appliquées, AGROCAMPUS
65 rue de Saint-Brieuc CS 84215 – F 35042 Rennes cedex France
pages@agrocampus-rennes.fr

RÉSUMÉ. On analyse un ensemble de données de végétation composé d'une part d'un tableau relevés-espèces codé en présence/absence et, d'autre part, d'un tableau de variables décrivant le milieu, ce dernier comportant à la fois des variables quantitatives relatives au sol et des variables qualitatives relatives à la position et l'usage du terrain. La difficulté méthodologique consiste à prendre en compte simultanément les deux groupes de variables (espèces et milieu), les types de variables étant hétérogènes, inter groupes et intra groupes. Cet exemple illustre comment l'Analyse Factorielle Multiple gère cette hétérogénéité et peut mettre en évidence des associations entre espèces et caractéristiques du milieu.

MOTS-CLÉS : Analyse Factorielle Multiple, Végétation, Données mixtes

1 Introduction

L'analyse de données de végétation a donné lieu a de nombreux travaux dans le cadre de la statistique exploratoire multidimensionnelle. La raison tient à leur double complexité.

Tout d'abord, le praticien analyse toujours la présence ou l'abondance d'une espèce en regard de caractéristiques du milieu ; il en résulte un tableau multiple dont :

- les lignes sont les relevés ;
- un premier groupe de colonnes est constitué par les espèces végétales ;
- un second groupe de colonnes rassemble les variables décrivant le milieu.

Pour aborder ce premier aspect de la complexité, c'est-à-dire tenir en compte simultanément ces deux groupes de colonnes, nous utilisons le cadre de l'Analyse Factorielle Multiple (*AFM* ; [ESC 84, 98]) en faisant jouer un rôle actif à ces deux groupes. Une telle analyse répond au double questionnement suivant :

- étudier les relations entre végétation et milieu ;
- mettre en évidence les principales dimensions de variabilité des relevés en tenant compte des deux groupes de variables de façon équilibrée.

Le second aspect de la complexité est technique : les variables mises en jeu peuvent être de différents types (quantitatif, qualitatif, fréquence) et leur analyse simultanée pose problème. En pratique, ces

différents types peuvent se superposer aux groupes de colonnes précédents : ainsi le groupe « végétation » est souvent constitué de variables qualitatives à deux modalités (présence/absence) et le groupe « milieu » rassemble des variables quantitatives. On dira que les groupes, définis par leur homogénéité de contenu, le sont aussi du point de vue de leur texture ([LEB 77]). Cette situation est prise en compte de façon adaptée par l'*AFM* qui permet l'introduction de groupes quantitatifs et/ou qualitatifs ([PAG 02]). Un niveau de complexité supplémentaire est atteint lorsque les groupes sont hétérogènes du point de vue de leur texture : ainsi, les variables de milieu peuvent être quantitatives ou qualitatives. On sait qu'il est possible, au sein d'une même analyse factorielle, de faire intervenir en tant qu'éléments actifs de variables des deux types : tel est l'objet de l'analyse factorielle de données mixtes (*AFDM* ; [ESC 79] ; [SAP 90] ; [PAG 04]). Le problème revient alors à introduire en *AFM* la possibilité de traiter des groupes de variables comme en *AFDM*. Telle est la caractéristique méthodologique de l'exemple que nous présentons.
Dans cette présentation, nous insisterons sur les contributions des différents types de variables pour montrer comment cette méthodologie permet d'équilibrer leur influence, intra groupe et inter-groupes.

2 Description sommaire des données

On a appliqué cette *AFM* sur des données mixtes provenant de [PIL 88 ; 92]. Il s'agit d'un tableau de végétation d'un pâturage typique du Brésil du Sud (Campos). L'échantillonnage a été fait sur 60 carrés de 0.5 x 0.5 m, situés le long de quatre transects correspondants à quatre gradients d'élévation et humidité, du haut de la colline jusqu'en bas (carrés 1-14, 15-29, 30-46, 47-60). L'élévation entre sommets et bases est de 30 mètres environ. Pour tout relevé on dispose de la présence absence de 60 espèces végétales, ainsi que des valeurs de 19 variables de milieu : 16 caractères quantitatifs de sol standard ([TED 95]), se réfèrent au contenu de nutriment et aux propriétés physiques et chimiques, et 3 indicateurs qualitatifs : un indicateur de position à 4 modalités (sommet, haut de pente, bas de pente, fond), un d'humidité (4 modalités) et un indicateur d'utilisation à 2 modalités (brouté ou pas).
Ces données ont déjà fait l'objet d'analyses. Bornant son attention aux classifications, [CAM 03] utilise la méthode proposée par [DEN 01] pour comparer deux classifications issues des analyses séparées sur la végétation et sur le sol. Avec la méthode qu'on propose on peut comparer plus directement les structures des différents tableaux se référant aux mêmes relevés.

3 Résultats

La comparaison des valeurs propres des analyses des deux groupes actifs (Tableau 1) montre la nécessité d'une pondération (gérée par l'*AFM*) : sans pondération le second groupe aurait une influence très forte.

Groupe	*Type d'analyse*	*axe 1*	*axe 2*	*axe 3*	*axe 4*	*axe 5*
1	*ACM*	0,1842	0,0928	0,0700	0,0516	0,0406
2	*AFDM (via ACP)*	6,6096	3,6132	2,3627	1,9767	1,5882

Tableau 1. Valeurs propres des analyse séparées des groupes

Remarquons aussi que le tableau de végétation a une inertie beaucoup plus dispersée sur les différents axes que le tableau des données de milieu. Dans le Tableau 2 on observe que le premier facteur du groupe 1 est corrélé à son homologue du groupe 2 ; deux autre coefficients de corrélations semblent élevés.

		Groupe végétation				
	Facteurs	*101*	*102*	*103*	*104*	*105*
Groupe milieu	*201*	0,85	0,11	0,13	-0,07	0,01
	202	0,16	-0,68	0,04	-0,08	0,03
	203	0,14	-0,32	0,12	0,20	-0,03
	204	-0,17	-0,06	0,63	-0,27	-0,05
	205	-0,06	0,03	0,15	0,25	-0,14

Tableau 2. Corrélations entre les facteurs de analyse séparées. Légende : 203 = 3^e facteur du groupe 2

Ce tableau suggère l'existence de 3 directions « communes » aux deux groupes. Les deux premiers facteurs sont prépondérants et nous nous y limiterons. Pour ces deux facteurs, les contributions des deux groupes sont fort équilibrées (Tableau 3) : la pondération inter-groupes classique de l'*AFM* (normalisation de l'inertie axiale maximum) fonctionne bien. Ces deux facteurs représentent donc une structure commune aux deux groupes.

		F1	*F2*	*F3*	*F4*	*F5*
Ensemble	*Deux groupes*	1,87	0,93	0,6	0,49	0,38
Groupe 1	*60 variables qualitatives à 2 modalités*	0,94	0,44	0,34	0,16	0,18
Groupe 2	*Ensemble : 19 variables mixtes*	0,93	0,49	0,26	0,33	0,2
	16 Variables quantitatives	0,67	0,35	0,11	0,32	0,10
	3 variables qualitatives (11 modalités)	0,27	0,14	0,16	0,02	0,10

Tableau 3. Inertie des facteurs de l'AFM et se décomposition par groupe

Au sein du groupe 2, les contributions des deux types de variables sont à peu près dans le rapport de leur dimensionalité (16 variables quantitatives d'un côté, 11 modalités auquel il faut retrancher 3 variables = 8 de l'autre) pour les deux premiers axes. Ce résultat est assez remarquable : dans ces données, les deux types de variables décrivent en gros avec la même inertie les deux principales structures communes.
Observant les tableaux des contributions, ainsi que le cercle des corrélations, on remarque que le premier axe oppose à gauche la *matière organique*, le *cuivre* et le *fer* opposés à l'*argile*, la *température*, le *magnésium* et la *potasse*. Le second axe oppose *aluminium* d'un coté et *zinc, manganèse, calcium* et *pH* de l'autre. Sur cette base on peut étudier la distribution des espèces sur le plan factoriel : on voit que la position fortement éloignée sur la gauche du premier axe des espèces *Eleocharis glauca, Relbunium hirtum, Rhyncospora tenuis, Paspalum pumilum, Centella biflora, Rhyncospora barrosiana* s'explique par l'abondance de *matière organique,* du *cuivre* et du *fer* et l'opposition à la *température*, au *magnésium* et au *potasse*. Par contre, la position éloignée en haut des espèces *Cuphea calophylla* et *Andropogon selloanus* s'explique par les oppositions entre les caractères opposés le long du second axe.

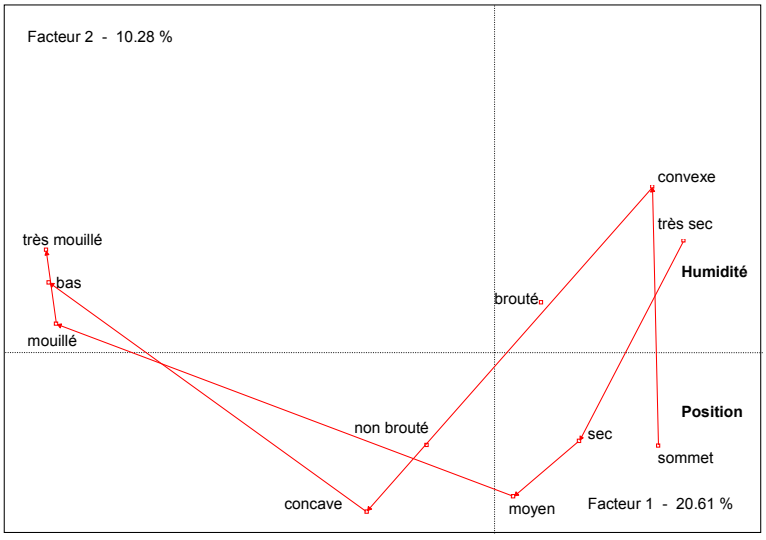


Figure 1. AFM : représentation des variables de milieu qualitatives

La représentation des relevés sur le plan factoriel présente une allure classique en forme de parabole, ce qui est en liaison avec le caractère dominant du premier axe. Un groupe de 10 points se démarque nettement sur la gauche. Néanmoins, c'est la projection des centres de gravité des modalités qualitatives sur le plan des premiers deux facteurs qui explique au mieux la distribution des plantes (Figure 1) : l'arc formé par les espèces correspond fort bien au gradient de l'*humidité* le long des transects passant de très humide et humide à moyen et ensuite à sec et très sec.
Dans le Tableau 4 le rapport de corrélation entre les variables qualitatives et facteurs s'avère très forte pour le premier et moyenne pour le second pour ce qui concerne humidité et position. Il faut remarquer que ce gradient ne correspond pas au transects, car la partie la plus sèche n'est pas au sommet de la pente mais sur la partie convexe, juste au dessous. C'est donc le gradient combiné de humidité et température du à la morphologie du terrain qui est représenté par la distribution en arc des espèces sur le plan factoriel.

	F1	*F2*	*F3*	*F4*	*F5*
Moisture (5 modalités)	83,2*	31,7*	20,3	7,1	37,5*
Elevation (4 modalités)	87,4*	49,3*	53,2*	0,4	9,8
Grazing level (2 modalités)	4,7	14,4*	30,0*	3,5	20,1*

*Tableau 4. Rapports de corrélation (au carré) entre les variables de milieu qualitatives et les facteurs de l'AFM. Légende : * = probabilité critique < 1%*

Cette distribution est confirmée par l'observation de la distribution des relevés sur le plan factoriel. Effectivement, on trouve à l'extrême gauche du premier axe les relevés les plus bas des transects et à l'extrême en haut du second les relevés sur la zone convexe des transects.

4 Conclusion

Cet exemple illustre bien la double pondération de l'*AFM* sur données mixtes. D'abord, comme en *AFDM*, une pondération (via en pratique un codage) permet d'équilibrer l'influence de variables quantitatives et qualitatives au sein d'un groupe. Ensuite, la pondération usuelle de l'*AFM* équilibre le rôle des groupes, même de types différents (ici un groupe qualitatif et un groupe mixte) dans une analyses globale. Dans cette analyse, l'interprétation de variables qualitatives se mène comme en *ACM* et celle des variables quantitatives comme en *ACP*.

Cette analyse, qui fait intervenir de façon active simultanément les données de végétation et de milieu, a permis de suggérer de façon simple et claire les dépendances fonctionnelles entre les plantes et les variables écologiques. De tels résultats ne peuvent être obtenus aussi directement par une *ACP* ou une analyse des correspondances, méthodes qui ne prennent pas en compte une éventuelle partition des colonnes (ici en deux groupes), ni même par une analyse canonique ou une de redondance, qui, elles, ne permettent pas l'introduction simultanée de variables quantitatives et qualitatives.

5 Bibliographie

[CAM 03] CAMIZ S., " Comparing hierarchical classifications node by node ". *Convegno Intermedio Analisi statistica multivariata per le scienze economico-sociali, le scienze naturali e la tecnologia*, Società Italiana di Statistica, 2003, Napoli, RCE Edizioni, sur CD-Rom.

[DEN 01] DENIMAL J.J., CAMIZ S. " Exact conditional tests for a reciprocal interpretation of hierarchical classifications built on a two way contingency table ", *Metron*, vol. 59, n° 3-4, 2001, pp.157-178.

[ESC 79] ESCOFIER B. " Traitement simultané de variables qualitatives et quantitatives en analyse factorielle ", *Les Cahiers de l'Analyse des Données,* vol. 4, 1979, pp.137-146.

[ESC 84] ESCOFIER B., PAGÈS J., " L'analyse factorielle multiple : une méthode de comparaison de groupes de variables ", dans Diday, E. (éd.), *Data Analysis and Informatics,* vol. 3, 1984, pp. 41-55, North-Holland. Elsevier.

[ESC 98] ESCOFIER B., PAGÈS J., *Analyses factorielles simples et multiples*. 1998, Paris, Dunod.

[LEB 77] LEBART L., MORINEAU A., TABARD N., *Techniques de la description statistique,* 1977, Paris, Dunod.

[PAG 02] PAGÈS J., " Analyse factorielle multiple appliquée aux variables qualitatives et aux données mixtes ". *Revue de Statistique Appliquée,* vol. 50, n° 4, 2002, pp. 5-37.

[PAG 04] PAGÈS J., " Analyse factorielle de données mixtes ". *Revue de Statistique Appliquée,* vol. 52, n* 4, 2004, pp. 93-111.

[PIL 88] PILLAR V.D., *Fatores de ambiente relacionados à variação da vegetação de um campo natural,* M.Sc. Dissertation, 1988, Porto Alegre (Brazil), Universidade Federal do Rio Grande do Sul.

[PIL 92] PILLAR V.D., JACQUES A.V.A., BOLDRINI I.i., " Fatores de ambiente relacionados à variação da vegetação de um campo natural ", *Pesquisa Agropecuária Brasilei*ra, vol. 27, 1992, pp.1089-1101.

[SAP 90] SAPORTA G., "Simultaneous analysis of qualitative and quantitative data ", *Atti della XXXV riunione scientifica,* Società Italiana di Statistica, 1990, pp. 63-72.

[TED 95] TEDESCO M.J., GIANELLO C., BISSANI C.A., BOHNEN H., VOLKWEISS S.J., *Análise de Solos, Plantas e Outros Materiais*, 2nd. ed., 1995, Porto Alegre, Faculdade de Agronomia, UFRGS.

Sur la normalisation pour la classification de données intervalles

Marie Chavent

Mathématiques Appliquées de Bordeaux, UMR 5466 CNRS
Université Bordeaux1 - 351, Cours de la liberation
33405 Talence Cedex
chavent@math.u-bordeaux.fr

RÉSUMÉ. *L'objectif de ce travail est de proposer plusieurs mesures de dispersion d'une variable décrite par des intervalles. On pourra en particulier utiliser ces mesures de dispersion pour normaliser le tableau de données intervalles ou encore de manière équivalente la distance utilisée dans l'algorithme de classification.*

MOTS-CLÉS : *Données symboliques intervalles, standardisation, distance normalisée, classification.*

1. Introduction

Dans un tableau de données $(x_i^j)_{n\times p}$ où n individus $\{1,...,i,...,n\}$ sont décrits par p variables quantitatives, lorsque toutes les variables ont des unités de mesures différentes, l'utilisation des variables telles quelles dans le calcul de la distance donnera de façon implicite plus de poids aux variables de plus forte dispersion, annihilant presque complètement l'effet des autres variables.
Une approche classique pour obtenir une classification ne privilégiant pas uniquement les variables de forte dispersion est de standardiser les données variable par variable. Or, une fois les variables centrées par la moyenne $\bar{x}^j$ (ou encore la médiane) et réduite par l'écart-type σ^j (ou encore l'étendue, l'intervalle interquartile...), la distance euclidienne entre i et i' s'écrit :

$$\sqrt{\sum_{j=1}^{p}(\frac{x_i^j-\bar{x}^j}{\sigma^j}-\frac{x_{i'}^j-\bar{x}^j}{\sigma^j})^2}=\sqrt{\sum_{j=1}^{p}\frac{1}{(\sigma^j)^2}(x_i^j-x_{i'}^j)^2} \qquad [1]$$

Les classifications obtenues sur les données centrées-réduites (ou même simplement réduites) avec la distance euclidienne simple et les classifications obtenues sur les données brutes avec la distance euclidienne normalisée par l'inverse de la variance sont équivalentes (cf. [1]). D'une manière plus générale, pour une distance basée sur des différences (écarts au carré, en valeur absolue...), la classification des données brutes ou centrées est la même. Nous ne nous intéressons donc pas ici au problème du centrage de variables intervalles. D'autre part, la classification obtenue avec les données brutes et la distance normalisée (i.e. distance "pondérée") est généralement la même que celle obtenue avec les données normalisées et la distance "simple" (non pondérée).
Dans ce travail, nous nous sommes intéressés au problème de la normalisation d'un tableau de données intervalles où chaque individu i est décrit sur chaque variable j par un intervalle

$$x_i^j=[a_i^j,b_i^j]\in I=\{[a,b]\mid a,b\in\Re\ ,\ a\leq b\}$$

Chaque individu i est ainsi décrit par un hyper-rectangle de $\mathbb{R}^p$:

$$x_i=\prod_{j=1}^{p}[a_i^j,b_i^j]$$

Il s'agit d'un cas particulier de données symboliques [DID 88], [BOC 00].
Le problème de la normalisation de données symboliques avait déjà été posé dans [CHA 97]. L'écriture de la variance comme une double somme pondérée des écarts était utilisée afin de définir la mesure de dispersion suivante :

$$\sigma^j = \frac{1}{2n} \sum_{i=1}^{n} \sum_{i'=1}^{n} d^2(x_i^j, x_{i'}^j) \quad [2]$$

où d est une fonction de comparaison entre deux descriptions symboliques quelconques. Cette mesure de dispersion y était utilisée pour définir une distance normalisée du type :

$$(\sum_{j=1}^{p} \frac{1}{(\sigma^j)^\alpha} d(x_i^j, x_{i'}^j)^\alpha)^{1/\alpha} \quad [3]$$

Le principal inconvénient de cette mesure de dispersion est la double somme qui peut rendre ce critère long à calculer pour de volumineuses bases de données.
On retrouve cette question de la normalisation de données intervalles dans [DEC 03] où les mesures de dispersion utilisées sont basées sur la dispersion des centres, des bornes supérieures ou inférieures des intervalles ou encore sur le maximum des bornes supérieures et le minimum des bornes inférieures.
Une approche parfois utilisée dans les algorithmes de classification de données intervalles est de considérer chaque intervalle comme un point de $\Re^2$, de comparer ces deux points avec la distance L_1 ou la distance L_2 en utilisant pour comparer deux hyper-rectangles :

$$\sum_{j=1}^{p} (d(x_i^j, x_{i'}^j)^\alpha)^{1/\alpha} \quad [4]$$

avec $\alpha = 1$ pour la distance L_1 et $\alpha = 2$ pour la distance L_2. Cette distance (avec $\alpha = 1$ et la distance L_1) est utilisée par exemple dans [DES 04] pour définir un algorithme de classification de type Nuées Dynamiques. Finalement, cela revient à créer dans le tableau de données initial deux colonnes indépendantes pour les bornes inférieures et les bornes supérieures des intervalles. La notion d'intervalle n'est donc pas vraiment prise en compte avec ce type de distance et cela revient à un recodage du tableau de données intervalles. On retrouve ainsi un tableau de données quantitatives classiques et des mesures de dispersion connues.
L'idée ici est donc double : utiliser une distance plus "spécifique" à la notion d'intervalle et s'affranchir de la double somme dans le calcul de la mesure de dispersion [2]. Afin de répondre à ce double objectif et dans la continuité des articles de [BOC 01], [CHA 02], [CHA 04], la distance de Hausdorff a été choisie pour définir des mesures de dispersion autour d'un centre optimal.

2. Mesures de dispersion autour d'un centre optimal

Pour une variable quantitative classique, la variance mesure la dispersion autour de la moyenne $\bar{x}^j$ qui est la solution optimale $\hat{y}$ du problème de minimisation suivant :

$$\min_{y \in \mathbb{R}} \sum_{i=1}^{n} (x_i^j - y)^2 \quad [5]$$

La variance s'écrit donc (à un coefficient près) :

$$f(\hat{y}) = \min_{y \in \mathbb{R}} \sum_{i=1}^{n} d^2(x_i^j, y) \quad [6]$$

De même, l'écart moyen à la médiane mesure la dispersion autour de la médiane x_M^j qui est la solution optimale $\hat{y}$ du problème de minimisation suivant :

$$\min_{y \in \mathbb{R}} \sum_{i=1}^{n} |x_i^j - y| \quad [7]$$

L'écart moyen à la médiane s'écrit donc (à un coefficient près) :

$$f(\hat{y}) = \min_{y \in \mathbb{R}} \sum_{i=1}^{n} d(x_i^j, y) \quad [8]$$

Si l'on se place maintenant dans le cas où $x_i^j = [a_i^j, b_i^j]$ et $y = [\alpha, \beta]$, plusieurs mesures de dispersion autour d'un centre optimal $\hat{y} = [\hat{\alpha}, \hat{\beta}]$ peuvent être définies selon la distance d choisie pour comparer deux intervalles et la fontion f choisie pour mesurer cette dispersion. Ici, la distance choisie pour comparer deux intervalles est la distance de Hausdorff. Cette distance d_H définie pour deux ensembles quelconques se simplifie dans le cas de deux intervalles [CHA 97] :

$$d_H([a_i^j, b_i^j], [a_{i'}^j, b_{i'}^j]) = \max(|a_i^j - a_{i'}^j|, |b_i^j - b_{i'}^j|) \quad [9]$$

La distance de Hausdorff est donc le maximum des écarts entre les bornes supérieures et les bornes inférieures des intervalles soit la distance L_∞ entre les vecteurs (a_i^j, b_i^j) et $(a_{i'}^j, b_{i'}^j)$.

2.1. *Mesure de "l'étoile"*

On se place dans le cas où la mesure de dispersion autour de $\hat{y}$ est :

$$f(\hat{y}) = \min_{y \in I} \sum_{i=1}^{n} d_H(x_i^j, y) \quad [10]$$

On retrouve ici une formulation proche de la mesure d'homogénéité d'une classe C dite mesure de "l'étoile" :

$$\min_{i \in C} \sum_{j \in C} d_{ij}$$

Dans [CHA 02] on démontre par un recodage des intervalles $[a_i^j, b_i^j]$ en fonction de leur milieu m_i^j et de leur demi-longueur l_i^j que l'intervalle $\hat{y}$ qui minimise $\sum_{i=1}^{n} d_H(x_i^j, y)$ a pour milieu $\hat{\mu}$ et pour demi-longueur $\hat{\lambda}$ avec :

$$\hat{\mu} = mediane\{m_i^j \mid i = 1, ..., n\} \quad [11]$$

$$\hat{\lambda} = mediane\{l_i^j \mid i = 1, ..., n\} \quad [12]$$

On définit alors la mesure de dispersion σ^j suivante :

$$\sigma^j = \sum_{i=1}^{n} \max(|a_i^j - \hat{\mu} + \hat{\lambda}|, |b_i^j - \hat{\mu} - \hat{\lambda}|) \quad [13]$$

2.2. *Mesure du "rayon"*

On se place dans le cas où la mesure de dispersion autour de $\hat{y}$ est :

$$f(\hat{y}) = \min_{y \in I} \max_{i=1...n} d_H(x_i^j, y) \quad [14]$$

On retrouve ici une formulation proche de la mesure d'homogénéité d'une classe C dite mesure du "rayon" :

$$\min_{i \in C} \max_{j \in C} d_{ij}$$

Dans [CHA 04], on démontre que l'intervalle $\hat{y}$ qui minimise $\max_{i=1\ldots n} d_H(x_i^j, y)$ a pour bornes inférieure et supérieure :

$$\hat{\alpha}^j = \frac{\max_{i=1\ldots n} a_i^j + \min_{i=1\ldots n} a_i^j}{2} \quad [15]$$

$$\hat{\beta}^j = \frac{\max_{i=1,\ldots,n} b_i^j - \min_{i=1,\ldots,n} b_i^j}{2} \quad [16]$$

On définit alors la mesure de dispersion σ^j suivante :

$$\sigma^j = \max_{i=1\ldots n} \max(|a_i^j - \hat{\alpha}^j|, |b_i^j - \hat{\beta}^j|) \quad [17]$$

3. Application en classification

Ces deux mesures de dispersion [13] et [17] peuvent être utilisées en classification. Par exemple dans les deux méthodes de type Nuées Dynamiques proposées dans [CHA 02] et [CHA 04] la classification obtenue avec le tableau des intervalles normalisés $z_i^j = [\frac{a_i^j}{\sigma^j}, \frac{b_i^j}{\sigma^j}]$ et la classification obtenue avec la distance normalisée (i.e. pondérée par l'inverse de σ^j) sont les mêmes. En effet,

$$d_H(z_i^j, z_{i'}^j) = \max(|\frac{a_i^j}{\sigma^j} - \frac{a_{i'}^j}{\sigma^j}|, |\frac{b_i^j}{\sigma^j} - \frac{b_{i'}^j}{\sigma^j}|) = \frac{1}{\sigma^j} d_H(x_i^j, x_{i'}^j) \quad [18]$$

Dans [CHA 02], la distance d entre deux hyper-rectangles est la somme sur toutes les variables des distances de Hausdorff entre les intervalles et on a donc :

$$d(z_i, z_{i'}) = \sum_{j=1}^{p} \frac{1}{\sigma^j} d_H(x_i^j, x_{i'}^j) \quad [19]$$

Dans [CHA 04], la distance d entre deux hyper-rectangles est le maximum sur toutes les variables des distances de Hausdorff entre les intervalles et on a donc :

$$d(z_i, z_{i'}) = \max_{j=1\ldots p} \frac{1}{\sigma^j} d_H(x_i^j, x_{i'}^j) \quad [20]$$

4. Bibliographie

[BOC 00] BOCK H.-H., DIDAY E., Eds., *Analysis of symbolic data. Exploratory methods for extracting statistical information from complex data*, Studies in classification, data analysis and knowledge organisation, Springer Verlag, Heidelberg, 2000.

[BOC 01] BOCK H.-H., Clustering algorithms and kohonen maps for symbolic data, *ICNCB Proceedings*, Osaka, 2001, p. 203–215.

[CHA 97] CHAVENT M., Analyse des données symboliques. Une méthode divisive de classification, PhD thesis, Université Paris-IX Dauphine, 1997.

[CHA 02] CHAVENT M., LECHEVALLIER Y., Dynamical clustering of interval data. Optimization of an adequacy criterion based on hausdorff distance, JAJUGA K., SOKOLOWSKI A., BOCK H.-H., Eds., *Classification, Clustering, and Data Analysis*, Berlin, 2002, Sringer Verlag, p. 53–60.

[CHA 04] CHAVENT M., An Hausdorff distance between hyper-rectangles for clustering interval data, *IFCS Proceedings*, Chicago, 2004.

[DEC 03] DE CARVALHO F. A. T., BRITO P., BOCK H.-H., Une méthode Type Nuées Dynamiques pour les données symboliques quantitatives, DODGE Y., MELFI G., Eds., *Méthodes et Perspectives en Classification*, Presses Académiques Neuchatel, 2003, p. 79-81.

[DES 04] DE SOUZA R. M. C. R., DE CARVALHO F. A. T., Clustering of interval data based on city-block distances, *Pattern Recognition Letters*, vol. 25, 2004, p. 353-365.

[DID 88] DIDAY E., The symbolic approach in clustering and related methods of data analysis : The basic choices, BOCK H.-H., Ed., *Classification and related methods of data anlysis*, Amsterdam, 1988, North Holland, p. 673–684.

Deux méthodes de classification de règles d'association en fouille de textes

Hacène Cherfi, Amedeo Napoli et Yannick Toussaint

LORIA (UMR 7503)
B.P. 239, F-54506 Vandœuvre-lès-Nancy cedex
{cherfi,napoli,yannick}@loria.fr

RÉSUMÉ. Un processus de fouille de données textuelles s'appuyant sur l'extraction de règles d'association engendre un très grand nombre de règles extraites. Il est alors nécessaire pour classifier les règles extraites de pouvoir disposer de critères fiables en rapport avec des connaissances du domaine. Dans cet article, nous présentons deux méthodes de classification : la première met en jeu des mesures statistiques tandis que la seconde, plus originale, repose sur un modèle de connaissances du domaine. Une discussion sur le bien-fondé de cette dernière approche termine cet article.

MOTS-CLÉS. Fouille de données, fouille de textes, extraction de règles d'association classification de règles, modèle de connaissances.

1 Introduction

Dans cet article, nous présentons le processus de *fouille de textes* (FDT) comme un processus d'extraction de connaissances à partir de données contrôlé et orienté d'une part par un *analyste*, expert du domaine des textes, et d'autre part par un modèle de connaissances du domaine des textes. L'objectif est d'enrichir progressivement les connaissances de l'analyste et celles du modèle, et réciproquement, de guider le processus d'extraction grâce au modèle de connaissances.

Dans ce qui suit, le processus de FDT cherche à extraire des règles d'association à partir de tableaux booléens $\texttt{Textes} \times \texttt{Termes}$, où `Textes` désigne un ensemble de textes et `Termes` un ensemble de termes-clés associés. Deux méthodes de FDT sont évoquées, qui reposent toutes deux sur l'extraction de règles d'association, par l'intermédiaire d'un algorithme d'extraction de motifs fréquents. La première méthode propose une classification des règles d'association extraites sur la base de mesures de qualité statistiques [CNT03], où chaque mesure met en valeur des éléments de nature différente, comme des informations rares, stables au bruit, des dépendances fonctionnelles, etc. Cette approche relève d'un processus purement statistique, ne tenant aucun compte des connaissances du domaine. La seconde méthode propose au contraire une classification qualitative des règles extraites, où une règle d'association peut prétendre être de bonne qualité si elle contient des éléments d'information potentiellement aptes à enrichir le modèle des connaissances du domaine [CJNT04]. Le modèle de

connaissances, noté $(\mathcal{K}, \sqsubseteq)$, se ramène à un ensemble *fini* de termes-clés muni d'un ordre partiel, qui doit servir à l'interprétation des résultats de la fouille ; réciproquement, les résultats de la fouille doivent contribuer à enrichir le modèle.

Dans cet article, nous présentons successivement les deux processus de fouille de textes et montrons en quoi ils diffèrent, le premier dépendant de critères numériques et le second dépendant de critères qualitatifs en rapport avec les connaissances du domaine des textes. Une brève discussion sur la problématique de la fouille de textes termine l'article.

2 Les règles d'association en FDT et les mesures statistiques

Soit $\mathcal{T} = \{\mathtt{t}_1, \mathtt{t}_2, \ldots, \mathtt{t}_m\}$ un ensemble de textes et $\mathcal{K} = \{\mathtt{k}_1, \mathtt{k}_2, \ldots, \mathtt{k}_n\}$ un ensemble de termes-clés associés à ces textes. Une règle d'association est une implication pondérée de la forme $\mathtt{A} \longrightarrow \mathtt{B}$ où $\mathtt{A} = \{\mathtt{t}_1, \mathtt{t}_2, \ldots, \mathtt{t}_p\}$ et $\mathtt{B} = \{\mathtt{t}_{p+1}, \mathtt{t}_{p+2}, \ldots, \mathtt{t}_q\}$. La règle $\mathtt{A} \longrightarrow \mathtt{B}$ s'interprète comme le fait que tous les textes contenant les termes-clés $\{\mathtt{t}_1, \mathtt{t}_2, \ldots, \mathtt{t}_p\}$ ont tendance à contenir aussi les termes-clés $\{\mathtt{t}_{p+1}, \mathtt{t}_{p+2}, \ldots, \mathtt{t}_q\}$, avec une certaine probabilité donnée par la *confiance* de la règle. Plusieurs algorithmes permettent de mettre en œuvre un tel processus d'extraction de règles à partir de motifs fréquents, par exemple *Close* et *Pascal* [BTP+02].

Le *support* et la *confiance* sont deux mesures associées aux règles d'association utilisées pour réduire le nombre de règles extraites. Le support d'une règle est donné par le nombre de textes contenant à la fois les termes-clés de $\mathtt{A}$ et de $\mathtt{B}$ — la réunion des termes-clés est notée $\mathtt{A} \sqcap \mathtt{B}$ — tandis que la confiance d'une règle est le rapport entre le nombre de textes contenant $\mathtt{A} \sqcap \mathtt{B}$ et le nombre de textes contenant $\mathtt{A}$ (ce qui reflète la probabilité conditionnelle $\mathtt{P}(\mathtt{B}/\mathtt{A})$). La confiance donne une mesure du pourcentage d'exemples et de contre-exemples de la règle. La présence d'un contre-exemple montre qu'il existe des textes possédant les termes de $\mathtt{A}$ mais pas nécessairement tous les termes de $\mathtt{B}$. Lorsque la confiance vaut 1, la règle est dite *exacte*, sinon elle est *approximative*. Deux valeurs de seuil sont définies, $\sigma_{\mathtt{s}}$ pour le support minimal et $\sigma_{\mathtt{c}}$ pour la confiance minimale.

Partant d'une règle comme $\mathtt{A} \longrightarrow \mathtt{B}$, si la règle est constituée de motifs $\mathtt{A}$ et $\mathtt{B}$ très fréquents, alors ces motifs sont partagés par presque tous les textes et les probabilités associées, $\mathtt{P}(\mathtt{A})$, $\mathtt{P}(\mathtt{B})$ et $\mathtt{P}(\mathtt{A} \sqcap \mathtt{B})$, sont fortes ou très fortes ; inversement, l'intérêt des connaissances impliquées par les motifs $\mathtt{A}$ et $\mathtt{B}$, du point de vue de la découverte de connaissances, est faible. Si la règle est constituée de motifs $\mathtt{A}$ et $\mathtt{B}$ rares, alors ces motifs sont partagés par un petit nombre de textes mais apparaissent conjointement : ils sont vraisemblablement liés dans le contexte du domaine des textes étudiés et peuvent présenter un intérêt du point de vue de la découverte de connaissances. Les mesures de support et de confiance ne permettent pas toujours de discerner à elles seules les règles porteuses de sens et à fournir une classification adéquate des règles d'association extraites. C'est pourquoi d'autres mesures sont également utilisées, comme l'intérêt, la conviction, la dépendance, la nouveauté et la satisfaction (pour des détails voir [CNT03, Che04]).

3 La vraisemblance d'une règle d'association

3.1 Le modèle de connaissances

Le modèle de connaissances noté $(\mathcal{K}, \sqsubseteq)$ est caractérisé par un ensemble de termes $\mathcal{K}$ muni d'une relation de *spécialisation* $\sqsubseteq$ (qui est un ordre partiel). Le principe de la classification par la

vraisemblance est le suivant : il faut rechercher pour les écarter toute règle d'association $\texttt{A} \longrightarrow \texttt{B}$ qui ne fait que traduire une relation $\texttt{A} \sqsubseteq \texttt{B}$. Par exemple, si $\texttt{pomme} \sqsubseteq \texttt{fruit}$, alors la règle d'association $\texttt{pomme} \longrightarrow \texttt{fruit}$ va être rejetée car elle est connue dans le modèle $(\mathcal{K}, \sqsubseteq)$. En revanche, la règle $\texttt{tarte-cerise} \longrightarrow \texttt{chocolat}$ exprime potentiellement une relation intéressante entre les termes `tarte-cerise` et `chocolat`, entre lesquels il n'existe *a priori* pas de relation de spécialisation $\sqsubseteq$.

Dans ce qui suit, nous nous restreignons aux règles d'association dites *simples*, où la prémisse et la conclusion sont réduites à un seul et unique terme-clé, comme $\texttt{a} \longrightarrow \texttt{b}$ par exemple. Une règle d'association $\texttt{a} \longrightarrow \texttt{b}$ est dite *triviale* si la relation $\texttt{a} \sqsubseteq \texttt{b}$ existe dans le modèle $(\mathcal{K}, \sqsubseteq)$.

La *vraisemblance* d'une règle $\texttt{a} \longrightarrow \texttt{b}$ par rapport à un modèle de connaissances $(\mathcal{K}, \sqsubseteq)$) comme la probabilité de trouver un chemin allant de `a` vers `b` dans $(\mathcal{K}, \sqsubseteq)$. Cette probabilité s'appuie sur le principe de la « propagation de l'activation » (*spreading activation theory* [CL75]) selon laquelle un marqueur d'information part d'un sommet (un concept) du modèle, par exemple $\texttt{k}_1$, et se propage à travers ce modèle avec une certaine force, à la recherche d'un autre élément, par exemple $\texttt{k}_2$. La force s'affaiblit de façon proportionnelle à la distance parcourue par le marqueur et au nombre de possibilités offertes à chaque branchement.

3.2 La définition de la vraisemblance d'une règle

Étant donné un modèle de connaissances du domaine $(\mathcal{K}, \sqsubseteq)$, nous définissons une table de probabilités de transitions qui va servir de base à la mesure de vraisemblance entre deux termes-clés. La probabilité de transition entre un terme-clé $\texttt{k}_\texttt{i}$ et un terme-clé $\texttt{k}_\texttt{j}$ est calculée sur la base du chemin de longueur minimale qui relie $\texttt{k}_\texttt{i}$ et $\texttt{k}_\texttt{j}$ dans le modèle. Il existe deux cas particuliers : (1) par convention, pour tout terme $\texttt{k}_\texttt{i}$, $\texttt{d}(\texttt{k}_\texttt{i}, \texttt{k}_\texttt{i}) = 1$; ceci pour prendre en compte la réflexivité de la relation de spécialisation et éviter des probabilités anormalement élevées en cas d'absence d'arc sortant ; (2) s'il n'existe pas de chemin entre un terme $\texttt{k}_\texttt{i}$ et un terme $\texttt{k}_\texttt{j}$, alors $\texttt{d}(\texttt{k}_\texttt{i}, \texttt{k}_\texttt{j}) = 2\texttt{N} + 1$ où $\texttt{N}$ est le cardinal de $\mathcal{K}$ ($\mathcal{K}$ est *fini*).

La probabilité de transition entre $\texttt{k}_\texttt{i}$ et $\texttt{k}_\texttt{j}$ définit la vraisemblance notée $\texttt{V}(\texttt{k}_\texttt{i}, \texttt{k}_\texttt{j})$ de la règle $\texttt{k}_\texttt{i} \longrightarrow \texttt{k}_\texttt{j}$ et repose sur le produit de deux facteurs : la distance de $\texttt{k}_\texttt{i}$ à $\texttt{k}_\texttt{j}$ et le *poids* de $\texttt{k}_\texttt{i}$ dans le modèle noté $\delta(\texttt{k}_\texttt{i})$. En outre, il faut encore tenir compte de deux éléments : (1) plus la distance entre deux termes-clés est grande, plus la valeur de la vraisemblance doit être faible, (2) le poids d'un élément $\texttt{k}_\texttt{i}$ dépend de l'ensemble des termes-clés du modèle, qu'ils soient atteignables ou non depuis $\texttt{k}_\texttt{i}$. Ainsi, la formule qui calcule la vraisemblance entre $\texttt{k}_\texttt{i}$ et $\texttt{k}_\texttt{j}$ est la suivante : $\texttt{V}(\texttt{k}_\texttt{i}, \texttt{k}_\texttt{j}) = [\texttt{d}(\texttt{k}_\texttt{i}, \texttt{k}_\texttt{j}) \times \delta(\texttt{k}_\texttt{i})]^{-1}$, où le poids de $\texttt{k}_\texttt{i}$ $\delta(\texttt{k}_\texttt{i}) = \sum_{\texttt{x} \in \mathcal{K}} 1/\texttt{d}(\texttt{k}_\texttt{i}, \texttt{x})$. Le poids $\delta(\texttt{k}_\texttt{i})$ d'un élément $\texttt{k}_\texttt{i}$ est dépendant du nombre d'arcs sortants associés à $\texttt{k}_\texttt{i}$ dans le modèle $\mathcal{K}$: plus ce nombre est élevé plus le poids est élevé. À l'inverse, lorsqu'il n'existe aucun arc sortant, l'élément lui-même devient prépondérant car $\texttt{d}(\texttt{k}_\texttt{i}, \texttt{k}_\texttt{i}) = 1$. Il faut encore remarquer que le poids $\delta(\texttt{k}_\texttt{i})$ est calculé une seule fois pour tout $\texttt{k}_\texttt{i}$ et que l'équation suivante est vérifiée : $\sum_{\texttt{x} \in \mathcal{K}} \texttt{V}(\texttt{k}_\texttt{i}, \texttt{x}) = 1$. Un exemple est proposé et détaillé dans [CNT05].

4 Discussion, conclusion et perspectives

Beaucoup de travaux en fouille de textes s'intéressent à la façon de gérer le très grand nombre de règles d'association extraites. La plupart de ces travaux abordent le problème du point de vue statistique, sans chercher à tenir compte de connaissances du domaine.

Dans cet article, nous proposons deux méthodes de classification de règles d'association pour la FDT, qui s'appuient pour l'une sur des mesures statistiques et pour l'autre sur une mesure de vraisemblance qui est fonction d'un modèle de connaissances du domaine. Le calcul et l'utilisation de la mesure de vraisemblance sont de nature différente de ce qui se pratique habituellement en FDT avec des mesures statistiques. La mesure de vraisemblance permet de classifier les règles d'association extraites et de focaliser l'attention de l'analyste sur des règles qui ne reflètent pas une relation de spécialisation dans le modèle. Ainsi, le comportement de la mesure de vraisemblance en terme d'apport de nouvelles connaissances pour enrichir une ontologie par exemple est cohérent avec ce qui peut être attendu d'un processus d'extraction de connaissances à partir de textes pour un analyste spécialiste du domaine des textes. Par ailleurs, l'approche présentée ici autorise un enrichissement mutuel entre modèle de connaissances et processus de FDT, ce qui montre qu'une telle approche peut être véritablement qualifiée d'approche pour l'extraction de connaissances à partir de textes guidée par les connaissances du domaine.

Le travail de recherche actuel peut être prolongé dans un certain nombre de directions : définition de la vraisemblance de règles complexes (avec des prémisses et des conclusions comptant plus d'un terme-clé), prise en compte dans le modèle de connaissances de relations de causalité, temporelles ou spatiales ... Il reste encore à approfondir les liens entre les aspects statistiques et les aspects qualitatifs, et à relier de façon plus significative les classifications issues des mesures statistiques et celles qui sont issues du modèle de connaissances : peu de travaux se sont jusqu'à présent intéressés cette tâche ; le domaine est donc ouvert et potentiellement fertile pour la fouille de textes.

Références

[BTP+02] Y. Bastide, R. Taouil, N. Pasquier, G. Stumme, and L. Lakhal. Pascal : un algorithme d'extraction des motifs fréquents. *Technique et science informatiques*, 21(1) :65–95, 2002.

[Che04] H. Cherfi. *Étude et réalisation d'un système d'extraction de connaissances à partir de textes*. Thèse d'université, Université Henri Poincaré (Nancy 1), 2004.

[CJNT04] H. Cherfi, D. Janetzko, A. Napoli, and Y. Toussaint. Sélection de règles d'association par un modèle de connaissances pour la fouille de textes. In M. Liquière and M. Sebban, editors, *Actes de la conférence sur l'apprentissage (CAp 2004), Montpellier*, pages 191–206. Presses Universitaires de Grenoble, 2004.

[CL75] A. Collins and E. Loftus. A spreading-activation of semantic processing. *Psychological Review*, 82(6) :407–428, 1975.

[CNT03] H. Cherfi, A. Napoli, and Y. Toussaint. Vers une méthodologie de fouille de textes s'appuyant sur l'extraction de motifs fréquents et de règles d'association. In R. Gilleron, editor, *Actes de la conférence sur l'apprentissage (CAp-03), Laval*, pages 61–76, 2003.

[CNT05] H. Cherfi, A. Napoli, and Y. Toussaint. Deux méthodes de classification de régles d'association pour la fouille de textes. Rapport de recherche, LORIA, Nancy, 2005.

Comparaison de textes sanskrits en vue d'une édition critique

Marc Csernel[1], **Patrice Bertrand**[2]

[1]*Inria Rocquencourt & Univ. Paris-Dauphine*
Domaine de Voluceau ; BP 105
78 153 Le Chesnay Cedex ; France
Marc.Csernel@inria.fr

[2]*GET – ENST Bretagne & Inria Rocquencourt*
UMR 2872 TAMCIC – Département LUSSI
CS 83818 ; 29 238 Brest Cedex 3 ; France
Patrice.Bertrand@enst-bretagne.fr

RÉSUMÉ. Malgré son titre un peu incongru, l'édition critique de textes sanskrits soulève des problèmes non triviaux d'analyse de données, liés à la structure même du sanskrit. Dans cet article, après avoir défini ce qu'est une édition critique, nous examinons tout d'abord les problèmes générés par la structure spécifique du sanskrit. Puis, en tenant compte de ces spécificités, nous décrivons comment mesurer la distance entre deux textes afin notamment de les classer et de déterminer la filiation d'un ensemble de manuscrits. De fait, les techniques que nous proposons sont très proches de celles employées lors des comparaisons de chaînes moléculaires.

MOTS-CLÉS : édition critique, sanskrit, distance intertextuelle.

1 Introduction

Une édition critique est un ouvrage dans lequel on fait apparaître toutes les variantes d'un texte littéraire (ou autre). L'élaboration d'une édition critique prend toute son importance et toute sa difficulté lorsqu'un texte est connu au travers d'un vaste ensemble de manuscrits dispersés à la fois dans le temps et dans l'espace. La retranscription des manuscrits est en effet affectée de modifications successives qui proviennent de façon volontaire ou non des scribes, et/ou des atteintes du temps rendant inutilisable tel ou tel morceau. En outre, un manuscrit donné peut éventuellement provenir de plusieurs sources : manuscrits recopiés par d'autres scribes, épigraphies, ...

Si l'édition critique prend en compte de nombreux manuscrits, alors elle peut revêtir pour le profane une allure rébarbative, puisqu'en poussant les choses à l'extrême, le lecteur se retrouve devant une page comportant quelques lignes faisant partie du texte de l'édition (appelé encore texte "maître"), et un grand nombre de lignes écrites sous la forme de notes de bas de pages, décrivant par le menu les variations du texte en fonction des différentes sources possibles.

Le choix du texte de l'édition relève de l'éditeur qui par comparaison des textes, peut aussi bien retenir un manuscrit particulier qu'un texte "moyen" qui résulte de son analyse. Or, la comparaison deux à deux des variantes d'un texte est un travail des plus fastidieux, et donc le choix du texte de l'édition est un authentique travail de bénédictin. C'est pourquoi depuis longtemps l'informatique s'est mise au service des "éditeurs critiques". Malheureusement l'écriture du sanskrit est dotée de caractéristiques qui rendent inutilisables les logiciels d'édition habituels. Par exemple et bien qu'il puisse être appliqué au sanskrit, Edmac [LAV 96] n'est qu'un ensemble de macros Tex qui facilitent la présentation de l'édition. Malgré ses succès concernant l'anglais ([OHA 93]), Collate [ROB 94,00] n'a pas réussi à produire des résultats satisfaisants en ce qui concerne le sanskrit. Le logiciel Anastasia [ANA 00] donne à l'université de Munster de bons résultats avec la graphie grecque en fournissant une édition critique électronique interactive de l'évangile selon St Jean ([JON 03]), mais ne peut servir à d'autres graphies.

Dans ce qui suit, après avoir exposé certaines spécificités graphiques du sanskrit, nous présentons les problèmes rencontrés pour créer un logiciel de génération "assistée" d'édition critique, en abordant plus particulièrement la comparaison des textes sanskrits. Puis, nous évoquons les « produits dérivés » de l'édition critique informatisée, en particulier la filiation des manuscrits à l'aide d'arbres phylogénétiques.

2 La graphie du sanskrit

La graphie du sanskrit est dotée de certaines particularités qui rendent spécifique la comparaison des textes. Tout d'abord, le sanskrit possède un alphabet de 46 lettres et se dactylographie suivant un système de translittération, le plus courant étant le *système Velthuis* avec lequel nous travaillons. La translittération consiste à écrire le sanskrit suivant l'alphabet latin, en faisant correspondre à chaque lettre sanskrit une séquence de lettres latines. Par conséquent, les comparaisons ne peuvent s'effectuer "lettre par lettre", celles-ci ne correspondant pas directement aux caractères latins, mais seulement "séquence par séquence", chaque séquence correspondant à la translittération latine d'un caractère sankrit (selon le codeVelthuis).
De plus, le sanskrit est une langue qui s'écrit souvent sans espace entre les mots. Ce n'est pas la seule : les épigraphies latines que nous pouvons encore admirer ne comportent pas d'espace, mais restent lisibles[†]. En ce qui concerne le sanskrit et les langues d'une graphie apparentée, cette absence peut rendre les textes ambigus (et c'est même un effet de style recherché). Une "désambiguïsation" automatique des textes dont la graphie dérive du sanskrit, a été étudiée ([DEL 03]). Comme nous le verrons au paragraphe suivant, l'absence de caractère séparateur entre les mots augmente considérablement la complexité algorithmique. Afin de diminuer cette complexité et de rendre possible l'identification des mots sur lesquels porte la différence entre les textes, nous allons utiliser un texte *lemmatisé*, c'est-à-dire une version du texte incluant des signes spécifiques afin d'indiquer les séparations entre les mots, les racines, les préfixes, les suffixes ... Cette version lemmatisée du texte de l'édition va servir de référence à la comparaison de tous les autres textes, et se nomme dans notre contexte le "padapatha".
Enfin, on pourrait croire que la lemmatisation permet de résoudre l'essentiel des problèmes liés à la comparaison des textes sanskrits, mais il n'en est rien. En effet, les mots sanskrits se comportent comme des protéines : ils se transforment lorsqu'ils s'accolent. C'est ce que l'on appelle des *Sandhi*. Un médiocre exemple de sandhi est fourni en français par la transformation du préfixe privatif **in** en **im** devant un **m**, un **b** ou un **p** (par exemple, ont dit **in**édit mais **im**muable)[‡]. Plus spécifiquement, une même séquence incluse à la fois dans le padapatha et dans un manuscrit, ne figure pas sous la même forme dans les deux textes en raison de la lemmatisation du padapatha et de l'existence de sandhi. Ainsi dans notre mauvais exemple français, il faudrait comparer "in+muable" avec "immuable". Disons enfin qu'il existe un dernier plaisir : les règles de copistes permettent d'utiliser certains caractères à la place d'autres sans que cela porte à conséquence …

3 Comparaison de textes sanskrits : élaboration d'une distance

En vue d'établir l'édition critique, nous devons comparer le texte lemmatisé "padapatha" avec les textes des manuscrits. La procédure de comparaison s'effectue séquentiellement selon l'ordre des mots du "padapatha". Les différences sont de trois types. Celles qui apparaissent entre deux mots, celles qui sont relatives aux assemblages de mots (tournures de phrases) et qui ressemblent à la différence existant entre "d'amour mourir me font" en lieu et place de "me font mourir d'amour", et enfin celles qui concernent les pans de phrase rajoutés ou oubliés par un scribe. L'identification de ces pans de phrases, consiste à déterminer les endroits où s'arrête les parties de texte oubliées ou rajoutées, ce qui entraîne une importante difficulté algorithmique.
Pour évaluer ces trois types de différences nous proposons d'utiliser l'algorithme **DIFF** [HUN 77], ou plus exactement certains de ses avatars. Rappelons que **DIFF**, grand classique des méthodes d'alignement de séquences, est basé sur la programmation dynamique et sur les distances d'édition. Ces distances sont des

[†] Les manuscrits écrits en minuscules carolingiennes, bien souvent ne comportent pas de blanc, mais deviennent en revanche rapidement incompréhensible passé les premiers mots.
[‡] La liste des sandhi est donnée par les grammaires sanskrites (voir par exemple [REN 96]).

fonctions simples des nombres et des coûts d'insertions, de suppressions et de substitutions nécessaires pour passer d'une chaîne à l'autre (cf. Crochemore [CRO 01]).
Pour comparer deux mots, nous calculons la longueur d'une plus longue séquence commune (LCS). Dans ce cas on utilise l'algorithme **DIFF** qui ignore l'opération de substitution en lui donnant un coût trop élevé. Nous proposons de calculer la distance entre deux mots *x* et *y* selon la formule suivante (voir aussi [CRO 01]) :

$$d(x,y) = \frac{|x| + |y| - 2 * LCS(x,y)}{|x| + |y|}.$$

Le tableau 1 illustre ce type de comparaison entre "**mour**ir" et " d'a**mour** ". Ce tableau se lit de la manière suivante: la première ligne contient les indices de colonne correspondant aux caractères du mot "d'amour". La première colonne comprend les même informations pour le mot " mourir ". La seconde ligne contient le premier des mots à comparer : M1="d'amour", et la seconde colonne contient le second mot à comparer : M2 = "mourir". Chaque case du tableau d'indice (**n,m**) contient le nombre de caractères commun entre les **n** premiers caractères de M1 et les **m** premiers caractères de M2. Le tableau est bâti suivant l'algorithme de la programmation dynamique. Le coin inférieur droit du tableau contient le nombre de caractères d'une plus longue séquence commune. Sur notre exemple, la longueur de cette séquence étant 4, la distance entre **mourir** et **d'amour** est (6+7-2*4) / (6+7) = 5/13 = 0.385.

	j	0	1	2	3	4	5	6
i		**D**	**‘**	**A**	**M**	**O**	**U**	**R**
0	**M**	0	0	0	1	1	1	1
1	**O**	0	0	0	1	2	2	2
2	**U**	0	0	0	1	2	3	3
3	**R**	0	0	0	1	2	3	4
4	**I**	0	0	0	1	2	3	4
5	**R**	0	0	0	1	2	3	4

Tableau 1

En ce qui concerne les différences entre tournures de phrase, nous avons adapté l'algorithme **DIFF** sans augmenter sa complexité algorithmique. Notre approche implique évidement d'avoir défini dans les deux textes comparés, une unité lexicale appelée MOT.
Nous proposons de reconnaître les phrases "équivalentes" avec un algorithme qui consiste d'abord à calculer les distances entre chaque paire de mots issus des deux phrases. Puis, l'algorithme calcule la distance entre les deux phrases comme étant le poids d'un chemin hamiltonien de poids minimum dans le graphe biparti complet qui est défini de la manière suivante :

- une partie de la bipartition représente une phrase, et chaque sommet un mot de cette phrase ;
- les arêtes, qui relient donc deux mots, sont valuées par la distance entre ces deux mots.

En transposant au français, il faut pouvoir retrouver "d'amour mourir me font" en lieu et place de "me font mourir d'amour", pour paraphraser le bourgeois gentilhomme [POQ 98]. Le tableau 2 donne les distances ainsi calculées entre les deux phrases. Nous pouvons constater que chaque ligne contient une distance nulle, et que par conséquent il existe un hamiltonien de poids nul : les deux phrases sont bien équivalentes.

	me	*font*	*mourir*	*d'amour*
d'amour	0.778	0.818	0.385	0.000
mourir	0.750	0.800	0.000	0.385
me	0.000	1.000	0.750	0.778
font	1.000	0.000	0.800	0.818

Tableau 2

Il en résulte que la définition de notre distance intertextuelle dépend de plusieurs facteurs : distances entre mots, distances entre phrases, dissimilarités entre paragraphes que nous n'avons fait qu'évoquer (pans de phrases oubliés...), ainsi que d'un aspect que nous n'avons pas encore mentionné, les dissimilarités générées par les méta-données accessibles (couleur de l'encre, style de graphie, ...).

4 Conclusion

La masse considérable d'informations traitées par les (récents) logiciels d'édition critique (cf. [OHA 93], [ROB 94,00], [LAV 96] et [ANA 00]) conduit à diverses questions d'analyse de données exploratoire qui sont souvent spécifiques de l'œuvre étudiée et du langage utilisé. Il est plus facile de répondre à ces questions à l'aide d'un logiciel offrant des possibilités de visualisation ([MON 02]). Dans ce texte, nous avons introduit des distances et des dissimilarités entre textes, qui chacune mesure l'écart entre deux textes selon un critère donné qui est souvent spécifique. Nous envisageons de construire des hypothèses de filiation des manuscrits, à l'aide notamment de représentations arborées ([BUN 71], [BAR 91]) basées sur une dissimilarité "synthétique" obtenue comme moyenne pondérée des dissimilarités relatives à un critère particulier. Le choix de cette pondération peut être guidée par l'exigence de stabilité de la représentation arborée qui en résulte. Un exemple d'une telle représentation arborée sera exposé lors de notre présentation.

5 Bibliographie

[ANA 00] Scholarly Digital Editions Leicester (UK). http://server30087.uk2net.com/hengwrt.

[BAR 91] BARTHELEMY J.-P. & GUENOCHE A. (1991) *Trees and Proximity Representations*. John Wiley & Sons (première édition française : *Les arbres et les représentations des proximités*, Paris : Masson 1988).

[BUN 71] BUNEMAN P. (1971) *Filiation of Manuscript,* Mathematics in Archeological and Historical Sciences, Edinburgh University Press.

[CRO 01] CROCHEMORE M., HANCART C. & LECROCQ T. (2001) *Algorithmique du texte*. Vuibert, Paris.

[DEL 03] DEL VIGNA C. & BERMENT V. (2003) *Ambiguïtés irréductibles dans les monoïdes de mots*, Actes des 9èmes journées montoises d'informatique théorique, Montpellier, Sept 2002.

[HUN 77] HUNT J.W. & SZYMANSKI T.G. (1977) *A fast algorithm for computing longest common subsequence*, CACM 20:5 1977.

[JON 03] Westfälische Wilhelms-Universität MünsterSchlossplatz 2 48149 Münster, http://nestlealand.uni-muenster.de/AnaServer?NAtranscripts+0+start.anv.

[LAV 96] LAVAGNINO J. & WUJASTYK D. (1996) *Critical Edition Typesetting: The EDMAC format for plain TeX.* (San Francisco and Birmingham: TeX Users Group 1996). 108 pages, ill.

[MON 02] MONROY C., KOCHUMANN R., FURUTA R., URIBINA E., MELGOZA E. & GOENKA A. (2002) *Visualization of Variants in Textual Collations to Analyze the Evolution of Literary Works in the Cervantes Project*, Proceedings of the 6th European Conference, ECDL 2002. (Rome, Italy, Sept. 2002). M. Agosti and C. Thanos, eds. Berlin: Springer, 2002. 638-53.

[OHA 93] O'HARA R. J. & ROBINSON P.M.W. (1993) *Computer-Assisted Methods of Stemmatic Analysis*, Occasional Papers of the Canterbury Tales Project 1 (1993): 53-74. (Publication 5, Office for Humanities Communication, Oxford Univ.) Online: http://rjohara.net/cv/1993CTP.html.

[POQ 98] POQUELIN J.B. dit Molière *Le Bourgeois Gentilhomme Acte II, Sc 4.* Larousse **:** Petits classiques Larousse 10 juillet 1998.

[REN 96] RENOU L. (1996) *Grammaire sanskrite : phonétique, composition, dérivation, le nom, le verbe, la phrase*. Maisonneuve réimpression, Paris.

[ROB 94] ROBINSON P.M.W. (1994) *Collate: A Program for Interactive Collation of Large Textual Traditions*, in S. Hockey and N. Ide (eds.), Research in Humanities Computing 3, 32-45.

[ROB 00] *About the Project Edition of Collat.* http://www.cta.dmu.ac.uk/projects/collate/intro.html.

Distance des transferts entre partitions

Lucile Denoeud et Alain Guénoche

ENST, 46 rue Barrault 75014 Paris, denoeud@infres.enst.fr
IML, 163 Av. de Luminy, 13009 Marseille, guenoche@iml.univ-mrs.fr

RÉSUMÉ. *La* distance des transferts *entre deux partitions P et Q sur un même ensemble est définie comme le plus petits nombre de transferts d'un élément d'une classe dans une autre pour passer de P à Q. Nous détaillons ici l'algorithme de calcul des valeurs de cette distance.*

MOTS-CLÉS : *Partitions, transferts*

1. Introduction

En 1981, W. Day a étudié la complexité du calcul d'une dizaine de distances d'édition sur l'ensemble des partitions de X à n éléments, basées sur des opérations ensemblistes. La plus simple d'entre elles est définie comme le nombre minimum de transferts, d'un élément d'une classe dans une autre, éventuellement vide, pour passer de P à Q. En guise d'évaluation, il se contente de mentionner : it is a minimum cost flow metric since its computation is equivalent to the solution of a minimum cost flow problem on a suitable defined network. Et de préciser plus loin que ce calcul est en $O(n^3)$, sans plus indiquer comment le mettre en oeuvre. Pour en savoir plus, nous avons récemment étudié cette distance (Charon et al. 2005) que nous avons baptisée *distance des transferts*.

2. Distance des transferts

Soit P et Q deux partitions de X ayant respectivement p and q classes ; on admettra que $p \leq q$. On note

$$P = \{C_1, .., C_p\} \text{ et } Q = \{C'_1, .., C'_q\}.$$

Le nombre minimum de transferts pour passer de P à Q, noté $\theta(P, Q)$, est obtenu en construisant une bijection entre les classes de P et celles de Q. Pour ce faire, on commence par ajouter $q - p$ classes vides à P, si bien que P est considérée comme une partition à q classes.

Soit Υ l'application de $P \times Q \longrightarrow \mathbb{N}$ qui fait correspondre à une paire de classes le cardinal de leur intersection. Classiquement, on note $n_{i,j} = |C_i \cap C'_j|$, $n_i = |C_i|$ et $n'_j = |C'_j|$ désignent les cardinaux des classes et les sommes ligne et colonne du tableau de contingence. Soit Δ l'application qui à toute paire de classes (C_i, C'_j) fait correspondre le cardinal de la différence symétrique, noté $\delta_{i,j}$; on a $\delta(i, j) = n_i + n'_j - 2 \times n_{i,j}$. Soit Θ l'application de $(P, Q) \longrightarrow \mathbb{N}$ qui à un couple de classes (C_i, C'_j) fait correspondre $t(i, j) = |C_i - (C_i \cap C'_j)| = n_i - n_{i,j}$.

On considère donc le graphe biparti complet $K_{q,q}$ dont les sommets sont les classes de P et de Q, et les arêtes sont pondérées soit par Υ, Δ ou Θ.

Proposition 1 *Soit T une bijection entre les classes de P et de Q. Les assertions suivantes sont équivalentes :*
- *T minimise le nombre de transferts,*
- *T est un couplage de poids maximum w_1 dans $K_{q,q}$ pondéré par Υ ;*
- *T est un couplage de poids minimum w_2 dans $K_{q,q}$ pondéré par Δ ;*
- *T est un couplage de poids minimum w_3 dans $K_{q,q}$ pondéré par Θ.*

$$\theta(P,Q) = n - w_1 = \frac{w_2}{2} = w_3$$

Preuve : Les éléments situés dans l'intersection des classes sont ceux qu'il est inutile de déplacer ; il faut donc maximiser la somme des cardinaux des intersections des classes couplées. La relation entre les pondérations implique qu'un couplage de poids maximum selon Υ est nécessairement minimum selon Δ. Dans la différence symétrique entre deux classes, on fait intervenir pour chaque élément transféré deux opérations, le retrait et l'ajout dans une autre classe. Il faut donc diviser le poids du couplage par 2. Enfin, dans l'application Θ seuls les retraits sont comptabilisés.

3. Calcul de la distance entre deux partitions

Le calcul de $\theta(P,Q)$ est donc un problème d'affectation (des classes de P aux classes de Q) qui consiste à chercher un couplage parfait (à q arêtes) de poids minimum dans $K_{q,q}$ pondéré par Δ ou par Θ. La construction du graphe biparti est en $O(n^2)$. Le problème du couplage de poids minimum dans un graphe biparti complet est bien connu en Recherche Opérationnelle. Nous détaillons les trois étapes :

3.1. *Réduction*

Soit M la matrice de coûts d'affectation, indexée en ligne sur les classes de P et en colonne sur celles de Q. Le couplage d'une classe de P coûtera au moins la valeur minimum de sa ligne et pour une classe de Q le minimum de sa colonne. La première étape consiste à *réduire* la matrice M en soustrayant à chaque ligne puis à chaque colonne sa valeur minimum ; soit M' la matrice obtenue après réduction et w la somme des valeurs retranchées. On fait ainsi apparaître au moins un 0 par ligne et par colonne, et w est une borne inférieure du poids du couplage minimum.

3.2. *Couplage parfait*

Soit $G_{q,q}$ le graphe biparti non pondéré, dont les arêtes correspondent aux valeurs nulles de la matrice M'. Si l'on peut construire un couplage parfait dans $G_{q,q}$, alors il est nécessairement de coût minimum dans $K_{q,q}$. La question de savoir s'il existe un couplage de cardinal q dans $G_{q,q}$ peut être résolu par un algorithme de flot maximum (Ford & Fulkerson, 1962) appliqué au graphe $G_{q,q}$. C'est celle à laquelle Day faisait référence, et sa complexité est en $O(q^3)$. Mais cette procédure très générale peut ici être remplacée par une procédure plus simple, parce que nous sommes dans un graphe biparti.

Etant donné un graphe connexe $G = (X, E)$ et un couplage $K \subset E$, on appelle *chaîne alternée* une chaîne simple dont les arêtes sont alternativement dans K et dans $E \setminus K$. Cette chaîne est dite *augmentante* si ses deux extrémités sont des sommets non couplés. On trouve dans Berge [1957]et dans ses ouvrages ultérieurs :

Theorème 1 *Un couplage K est maximum dans un graphe G si et seulement si G ne contient pas de chaîne alternée* augmentante.

Il est clair que si l'on peut construire une chaîne alternée augmentante, en échangeant les arêtes du couplage avec celles qui sont hors couplage, on augmente la cardinalité de K d'une unité. On en déduit une procédure assez simple qui conduit à un couplage de cardinalité maximum. Elle partage les sommets de X en deux classes, les sommets pairs et les sommets impaires. Soit K un couplage quelconque qui laisse au moins deux sommets non couplés (sinon K est maximum) et soit x un sommet non couplé ; on construit un arbre alterné de racine x considéré comme pair.

- Pour tout sommet pair x, et pour tout sommet y adjacent à x non placé dans l'arbre
  - Ajouter (x, y) dans l'arbre ; y est impair ;
- Pour tout sommet impair y
  - Si y est non couplé, on a trouvé une chaîne augmentante
  - Sinon, soit z le sommet couplé à y ;
    - Ajouter (y, z) dans l'arbre ; z est pair ;

Si par construction de l'arbre alterné à partir d'un sommet non couplé, aucun sommet non couplé n'est atteint, le couplage K est maximum. Cette propriété n'est pas suffisante dans les graphes qui possèdent des cycles de longueur impaire, mais il n'y en a pas dans un graphe biparti.

Les graphes $G_{q,q}$ ne sont pas nécessairement connexes ; s'il y a des composantes connexes à nombre impair d'éléments, il n'y a évidemment pas de couplage parfait, mais il faut appliquer la construction d'un arbre alterné à toutes les composantes même celles qui ont un nombre pair d'éléments.

Nous noterons AA la procédure correspondante qui renvoie le nombre d'arêtes d'un couplage maximum. On dira qu'une ligne (resp. une colonne) est couplée si l'une de ses valeurs nulles dans M' est dans ce couplage.

3.3. *Méthode hongroise*

Si $AA(G_{q,q}) < q$, il faut introduire de nouvelles arêtes dans le graphe, celles qui sont nécessaires pour augmenter la cardinalité du couplage et qui sont de poids minimum. Ce problème est résolu par la *méthode hongroise* (Kuhn, 1955, 1956). La procédure marque les lignes et les colonnes de M' de la façon suivante :

- Tant que $AA(G_{q,q}) < q$
  1. Choisir un couplage de cardinalité maximum dans $G_{q,q}$
  2. Marquer toutes les lignes non couplées
  3. Tant qu'on peut marquer une nouvelle ligne :
  (a) Marquer les colonnes qui possèdent un 0 (non utilisé) dans une ligne marquée
  (b) Marquer les lignes couplées à une colonne marquée.
  4. Soit M" la sous-matrice de M' restreinte aux lignes marquées et aux colonnes non marquées et r sa valeur minimum.
  5. Soustraire r à chaque valeur de M" (ce qui fait apparaître de nouveau 0).
  6. Ajouter r à chaque case de M' à l'intersection d'une ligne non marquée et d'une colonne marquée.
  7. $w \leftarrow w + r$.
  8. Mettre à jour $G_{q,q}$ (les arêtes correspondent aux 0 de M').
- Fin tant que.

Nous ne rentrerons pas plus dans les détails et les justifications de cette procédure, que l'on trouvera par exemple dans Faure et al. (2000).

Exemple 1 *Soit* $P = (1, 2, 3|4, 5, 6|7, 8)$ *et* $Q = (1, 3, 5, 6|2, 7|4|8)$. *Les deux tables correspondant aux intersections et aux différences symétriques sont éditées ci-dessous. Deux couplages extrèmaux sont portés en gras. Tous deux donnent* $\theta(P, Q) = 4$.

Υ	*1,3,5,6*	*2,7*	*4*	*8*	Δ	*1,3,5,6*	*2,7*	*4*	*8*
1,2,3	***2***	*1*	*0*	*0*		*3*	***3***	*4*	*4*
4,5,6	*2*	*0*	***1***	*0*		***3***	*5*	*2*	*4*
7,8	*0*	***1***	*0*	*1*		*6*	*2*	*3*	***1***
$\emptyset$	*0*	*0*	*0*	***0***		*4*	*2*	***1***	*1*

Au couplage de poids maximum pour la table Υ *correspond la suite de transferts :* $(1, 2, 3|4, 5, 6|7, 8) \rightarrow (1, 3|4, 5, 6|2, 7, 8) \rightarrow (1, 3, 5|4, 6|2, 7, 8) \rightarrow (1, 3, 5, 6|4|2, 7, 8) \rightarrow (1, 3, 5, 6|4|2, 7|8)$.

Appliquons l'algorithme à la matrice M *de la table de la différence symétrique. L'étape de réduction donne* $w = 7$. *La matrice résultante* M' *n'admet pas de couplage parfait mais un couplage de cardinal 3 dont les arêtes figurent en gras dans la matrice. Les lignes et les colonnes marquées à l'issue de la procédure possèdent une étoile. La valeur minimum (*$r = 1$*) de la sous-matrice* M" *est soustraite. Il en résulte la table de droite qui admet, entre autres, le couplage parfait édité en gras.*

M'	*1,3,5,6*	*2,7*	*4*	*8*		*1,3,5,6*	*2,7*	*4*	*8*
1,2,3	***0***	*0*	*1*	*1*		*0*	***0***	*2*	*2*
4,5,6	*1*	*3*	***0***	*2*	$\star$	***0***	*2*	*0*	*2*
7,8	*5*	*1*	*2*	***0***	$\star$	*4*	*0*	*2*	***0***
$\emptyset$	*3*	*1*	*0*	*0*	$\star$	*2*	*2*	***0***	*0*
			$\star$	$\star$					

A ce couplage de poids $w_2 = 8$ *correspond une autre suite optimale :* $(1, 2, 3|4, 5, 6|7, 8) \rightarrow (1, 2, 3, 7|4, 5, 6|8) \rightarrow (2, 3, 7|1, 4, 5, 6|8) \rightarrow (2, 7|1, 3, 4, 5, 6|8) \rightarrow (2, 7|1, 3, 5, 6|8|4)$.

Il est clair que l'algorithme d'évaluation de la distance des transferts, dont nous n'avons décrit qu'une partie, est assez délicat à implémenter ; un programme en C peut être obtenu auprès des auteurs. Nous l'avons utilisé pour étudier la distribution des valeurs de distance à partir d'une partition fixée et pour comparer d'autres indices de distance entre partitions proches du point de vue des transferts (Denoeud et al. 2004).

Cette distance entre deux partitions quelconques est bornée par $n - 1$. Pour une partition donnée, la partition à distance maximum n'est pas nécessairement la partition à une seule classe ou la partition à n singletons. Dans Charon et al. (2005), nous donnons pour deux partitions à p et q classes, la plus grande valeur de distance possible. Ces bornes permettent de calculer un *indice de distance*, normé entre 0 et 1.

4. Références

C. Berge (1957) Two theorems in graph theory, *Proc. Nat. Acad. Sci.*, 43, 842-844.

I. Charon, L. Denoeud, A. Guénoche, O. Hudry (2005) Comparing partitions by element transferts, submitted.

L. Denoeud, H. Garreta and A. Guénoche (2004), Comparison of distance indices between partitions, Proceedings of Applied Stochastic Models in Data Analysis on CD ROM, Brest, May 2005.

W. Day (1981) The complexity of computing metric distances between partitions, *Mathematical Social Sciences*, 1, 269-287.

R. Faure, B. Lemaire, C. Picouleau, *Précis de Recherche Opérationnelle*, Dunod, 2000, 134-137.

L.R. Ford and D.R. Fulkerson (1962) *Flows in Networks*, Princeton University Press.

H.W. Kuhn (1955) The Hungarian method for the assignment problem, *Naval Res. Logist. Quart.*, 2, 83-97.

H.W. Kuhn (1956) Variants on the Hungarian method for the assignment problems, *Naval Res. Logist. Quart.*, 3, 253-258.

Remerciements

Ce travail a été réalisé avec l'aide de l'ACI IMP-Bio.

Reconnaissance d'objets tridimensionnels par leurs caractéristiques clés

Laurent Desmecht, Marcel Rémon

Département de Mathématique,
Université de Namur B-5000 Namur,
laurent.desmecht@fundp.ac.be, marcel.remon@fundp.ac.be

RÉSUMÉ. Les algorithmes de reconnaissance d'objets 3D procèdent habituellement en comparant directement l'objet à reconnaître à une base de données de vues 2D d'objets de référence. Ainsi, dans les méthodes classiques 2D, la structure et la nature de l'objet ne sont pas prises en compte.

Il est important de développer un modèle qui tienne réellement compte de cette spécificité, décrite ici sous la notion de caractéristique clé. Ainsi, les algorithmes pourront disposer d'informations pertinentes permettant d'effectuer plus efficacement la discrimination et la reconnaissance d'objets.

MOTS-CLÉS : Classification, Reconnaissance d'objets, Discrimination, Analyse d'images

1 Introduction

La reconnaissance d'objets tridimensionnels est un problème très difficile. En utilisant une base de données de vues 2D d'objets de références, de nombreux problèmes se posent. En particulier, lorsque le nombre d'objets augmente, la taille de la base de données augmente et il devient de plus en plus difficile de reconnaître correctement les objets puisque les informations ne permettent plus de séparer correctement les classes des objets.

L'idée préconisée dans cette présentation est que ces problèmes peuvent être évités en utilisant d'autres sources d'information que les vues 2D. Ces informations supplémentaires viennent de la structure de l'objet, ou de son contexte habituel.

2 Un nouveau modèle mathématique pour les objets tridimensionnels

2.1 Les caractéristiques clés

Les objets 3D complexes ont une certaine structure. Par exemple, une théière est formée d'une anse, d'un récipient, etc. L'idée est que, pendant la reconnaissance, la détection d'un objet passe par la détection de ses sous-objets appelés caractéristiques clés. L'ensemble des caractéristiques clés définissent l'objet.

En pratique, cette structure peut être facilement obtenue à partir d'une représentation en fils de fer; les objets sont alors considérés comme un ensemble de petites faces et les caractéristiques clés en sont des sous-ensembles. Cette approche a de nombreux avantages : la taille de la base de données peut être réduite de manière considérable puisque ce système permet de générer, à la demande, tous les vues possibles sous un grand nombre de formes (contours, région, squelette...).

2.2 Pondération des caractéristiques clés

De plus, comme les caractéristiques clés peuvent être connues a priori, il est possible d'y attacher manuellement des informations supplémentaires tel qu'un poids pour en indiquer l'importance relative dans le procédé de reconnaissance : par exemple, les roues d'une voiture jouent un rôle de premier plan dans la reconnaissance humaine de l'objet "voiture". Ici l'algorithme accorde un poids relatif supérieur aux caractéristiques clés "roue" et "vitre". La pondération des caractéristiques clés se fait sur base subjective (pondération a priori). L'algorithme est conçu pour mettre à jour et affiner les pondérations, par apprentissages successifs.

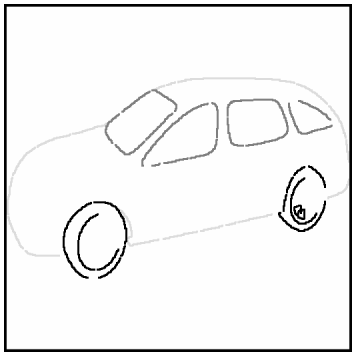


Figure 1: Exemple de choix de poids manuels, l'intensité de noir indique l'importance de la caractéristique clé.

L'idée générale est de proposer un cadre flexible dans lequel toutes les informations jugées pertinentes peuvent être encodées dans la base des objets de référence.

3 Reconnaissance d'un objet

La tâche principale d'un algorithme de reconnaissance de forme est d'identifier un objet se trouvant sur une image. En pratique, l'objet à reconnaître est comparé avec les objets de la base de données. L'objet qui est le plus proche de l'image de départ est considéré comme l'objet reconnu.

Une notion de mesure de dissimilarité est utilisée pour la comparaison des objets. L'idée est qu'un objet est identifié par ses caractéristiques clés. Ainsi, toutes les caractéristiques clés de l'objet de référence sont recherchées dans l'objet à analyser. Pour cela, la caractéristique clé de l'objet à reconnaître la plus proche est obtenue en minimisant une distance entre les caractéristiques clés. Finalement, la mesure de dissimilarité entre objets est obtenue par la moyenne pondérée des distances entre les caractéristiques clés. Cette pondération est donnée a priori dans la base de données.

$$\delta(\text{objet}, \text{modèle}) = \sum_{c \in \text{ modèle}} w_c \min_{c' \in \text{ objet}} d(c, c')$$

Où *c* est une caractéristique clé de l'objet modèle, *c'* est une caractéristique clé de l'objet à reconnaître et $\sum_{c \in \text{ modèle}} w_c = 1$ où w_c est le poids associé à *c*.

3.1 Distance sur les caractéristiques clés

Dans le cas où les caractéristiques clés sont modélisées comme un ensemble de contours, la possibilité de présence d'une caractéristique clé peut être estimée en comparant les contours de la caractéristique clé aux contours de l'objet à reconnaître.

Une notion de distance entre contours est nécessaire mais le choix de cette distance est délicat : la forme des objets varie. Ainsi deux contours provenant intuitivement du même objet peuvent avoir des apparences assez différentes.

Pour éviter ce problème, une distance non triviale est utilisée. Cette distance est une combinaison de trois autres distances comparant les contours selon différents critères.

$$d(C_1, C_2) = w_{\text{position}} d_{\text{position}}(C_1, C_2) + w_{\text{orientation}} d_{\text{orientation}}(C_1, C_2) + w_{\text{contexte}} d_{\text{contexte}}(C_1, C_2)$$

Où $w_{position}$, $w_{orientation}$ et $w_{contexte}$ sont les poids associés aux différentes distances (respectivement fixé à 0.4, 0.4 et 0.2 dans nos expérimentations).

La première distance $d_{position}(C_1, C_2)$ compare les contours en fonction de leurs positions. Cette distance est la distance euclidienne moyenne entre les deux contours :

$$d_{\text{position}}(C_1, C_2) = \int_0^{\ell} \sqrt{(f_x^1(t) - f_x^2(t))^2 + (f_y^1(t) - f_y^2(t))^2} dt$$

La position n'est pas suffisante : deux contours peuvent très bien avoir une distance proche alors qu'ils sont assez différents. C'est pourquoi une deuxième distance $d_{orientation}(C_1, C_2)$ est utilisée pour obtenir une certaine flexibilité. Les contours sont alors comparés en fonction de leurs allures: la distance est l'angle moyen entre les deux courbes.

$$d_{\text{orientation}}(C_1, C_2) = \int_0^{\ell} |\arctan(\frac{f_y'^1(t)}{f_x'^1(t)}) - \arctan(\frac{f_y'^2(t)}{f_x'^2(t)})| \, dt$$

Finalement, les deux premières distances comparent les contours au niveau local. Or, deux contours peuvent être très semblables et ne pas appartenir au même type d'objet. Une troisième distance est donc utilisée pour comparer les contours à un niveau plus global.

Cette distance compare les contours selon leurs positions par rapport au reste des contours des objets. Pour cela, la notion de "shape-contexts"[1] est utilisée: à chacun des contours est associé un histogramme log-polaire des autres contours de l'objet. Finalement, les histogrammes des contours sont comparés par un test Chi² à leurs équivalents issus de la base de données de référence.

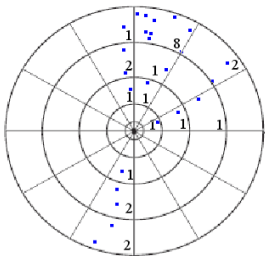


Figure 2 : Exemple de "shape context": chacun des points correspond à un contour

[1] La notion de "shape-contexts" est trop complexe pour être développée ici. Pour plus d'information voir [MAL 02]

4 Résultats

Nous avons construit une base de données de référence contenant huit objets de référence (une tasse, un rasoir, une voiture, un verre à vin, un révolver, un briquet, un ballon et une table). Sur chacun, nous avons surpondéré des caractéristiques clés et nous avons obtenu les résultats suivants à partir d'images réelles d'objets isolés, n'ayant pas servi à construire la base de référence.

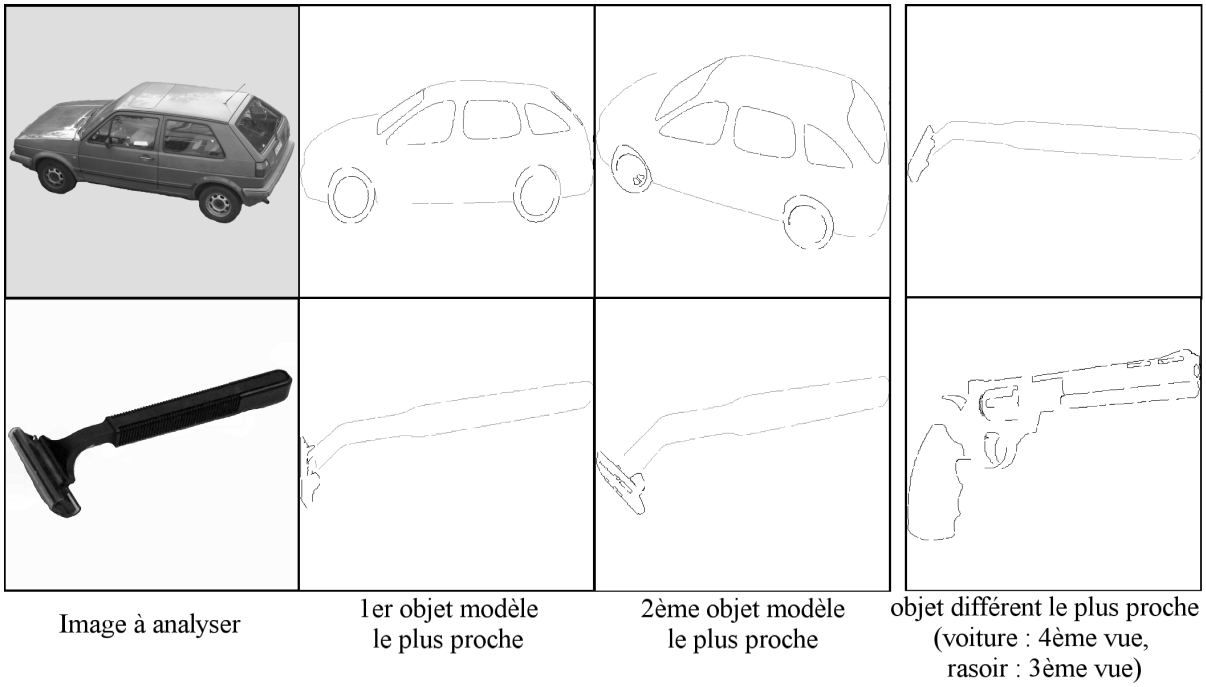


Figure 3 : Résultats

5 Conclusions et recherches futures

Les résultats obtenus sont encourageants et la flexibilité permise par ce modèle se révèle importante dans l'efficacité de l'algorithme.

Les recherches se poursuivent dans le cadre de la reconnaissance de scène plus complexe avec plusieurs objets. Le modèle est alors modifié pour y intégrer des propriétés de proximité d'objets ou de classes d'objets (ex: objets scolaires, tasses et couverts...).

6 Bibliographie

[CAN 86] CANNY J., "A computational Approach to Edge Detection*", IEEE Transactions on pattern analysis and machine intelligence*, 8(6), 1986.

[DES 04] DESMECHT L., "Reconnaissance d'objets 3D par leurs caractéristiques clés*", Mémoire à l'Université de Namur (FUNDP)*, 2004.

[MAL 02] MALIK J., BELONGIE S. and PUZICHA J., "A new descriptor for shape matching and object recognition*", IEEE Transactions on pattern analysis and machine intelligence*, 24(24), 2002.

Une nouvelle méthode pour l'estimation de nucléotides manquants en vue de l'inférence phylogénétique

Abdoulaye Baniré Diallo, Alpha Boubacar Diallo et Vladimir Makarenkov

Département d'informatique,
Université du Québec à Montréal,
Case postale 8888, succursale Centre-ville
Montréal (Québec) Canada, H3C 3P8

RÉSUMÉ. Cet article adresse le problème de la reconstruction d'arbres phylogénétiques à partir de séquences contenant des nucléotides manquants ou indéterminés. Nous proposons une méthode permettant d'estimer ces nucléotides manquants dans les séquences d'ADN ou d'ARN en se basant des modèles d'évolution de Jukes-Cantor [JUK 69] et de Kimura 2-paramètres [KIM 80]. Cette méthode permet d'améliorer la qualité de l'inférence phylogénétique comparativement aux méthodes "Ignore Missing Sites" (IMS) et "Proportional Distribution of Missing and Ambiguous Bases" (PDMAB) incluses dans le logiciel PAUP [SWO 01]. Par ailleurs, dans la partie application, nous montrons l'utilité de la nouvelle méthode en considérant un arbre phylogénétique pour un groupe de 10 eucaryotes.

MOTS-CLÉS : arbre phylogénétique, donnée manquante, modèle d'évolution

1 Introduction

La présence des nucléotides indéterminés dans les séquences d'ADN peut être due à la difficulté de séquencer certaines régions du génome d'un spécimen donné ou a une mauvaise conservation des spécimens. Ce problème peut représenter un grand obstacle pour la classification phylogénétique. Huelsenbeck [HUE 91] et Makarenkov et Lapointe [MAK 04] ont indiqué que les taxons qui contiennent de nombreux caractères inconnus diminuent considérablement la qualité de l'inférence phylogénétique. Dans cette étude nous proposons une nouvelle méthode, appelée *PEMV* (*Probabilistic estimation of missing values*), utilisant une approche probabiliste et visant à estimer les nucléotides manquants avant le calcul des distances d'évolution. Ici nous présentons cette méthode dans le cadre des modèles d'évolution de Jukes-Cantor [JUK 69] et de Kimura 2-paramètres [KIM 80], mais elle peut être généralisée à toute autre transformation séquences-distances. Dans la section suivante nous introduirons la nouvelle méthode et présenterons ses performances en inférant une phylogénie d'eucaryotes à partir de leur ADN mitochondriaux auxquels nous ajoutons des données manquantes. Nous comparons les performances de la nouvelle méthode à celles des méthodes *IMS* (« *Ignoring missing sites* ») et *PDMAB* (« *Proportional distribution of missing and ambiguous bases* »), qui sont incluses dans le populaire logiciel PAUP de Swofford [SWO 01]. La qualité d'inférence phylogénétique est estimée à l'aide de la distance topologique de Robinson et Foulds [ROB 80]. La méthode Neighbor Joining (NJ) de Saitou et Nei [SAI 87] a été utilisée dans les simulations pour reconstruire les arbres phylogénétiques.

2 La nouvelle méthode d'estimation des nucléotides manquants PEMV

La méthode *PEMV* est présentée ici dans le cadre des modèles d'évolution de Jukes-Cantor [JUK 69] et Kimura 2-paramètres [KIM 80]. Pour calculer les distances entre les paires de séquences, selon le modèle de Jukes-Cantor, la formule de correction suivante a été utilisée: $D = -3/4\ ln(1 - 4/3d)$, où d représente la distance observée. Selon le modèle de Kimura, la formule de correction est la suivante: $D = -1/2\ ln((1 - 2P - Q)\sqrt{(1-2Q)})$, où P représente le taux de transition et Q le taux de transversion entre les séquences.

Considérons un ensemble de séquences d'ADN (composées des nucléotides A, C, G et T). Admettons que le site k (i.e. colonne k) de la séquence i est inconnu (i.e. manquant). Pour calculer la distance entre la séquence i et toutes les autres séquences considérées, *PEMV* estime, à l'aide de l'équation 1, les probabilités P_{ik}(A), P_{ik}(C), P_{ik}(G) et P_{ik}(T) d'avoir respectivement le nucléotide A, C, G ou T au site k de la séquence i. La probabilité qu'une base manquante corresponde à un nucléotide spécifique dépend du nombre de séquences ayant le même nucléotide au site k, de même que de la distance entre i et toutes les autres séquences ayant des nucléotides connus au site k. On évalue en premier lieu le score de parité (i.e. la similarité) δ entre toutes les séquences (en ignorant les données manquantes). Ce score représente le rapport entre le nombre de paires de nucléotides distinctes et le nombre de sites comparables dans une paire de séquences. La probabilité P_{ik} se calcule comme suit :

$$P_{ik}(\mathrm{V}) = \frac{1}{N_k}\left(\sum_{\substack{j,\ \text{tels que} \\ C_{jk} = \mathrm{V}}} \delta_{ij} \ + \ \frac{1}{3} \sum_{\substack{j,\ \text{tels que} \\ C_{jk} \neq \mathrm{V}}} (1 - \delta_{ij}) \right) \tag{1}$$

où le caractère V remplace l'un des quatre nucléotides A, C, G ou T; N_k – est le nombre de valeurs existantes dans la colonne k; δ_{ij} – est le score de parité entre les séquences i et j; **C** – est la matrice des séquences des nucléotides,

Le théorème suivant caractérisant les probabilités P_{ik}(A), P_{ik}(C), P_{ik}(G) et P_{ik}(T) pour une séquence i et un site donné k peut être formulé (sa preuve n'est pas présentée ici):

Théorème. *Pour toute séquence i, tout site k de la matrice* **C***, tels que* c_{ik} *est un nucléotide manquant, nous avons* : $P_{ik}(A) + P_{ik}(C) + P_{ik}(G) + P_{ik}(T) = 1$.

Une fois les différentes probabilités P_{ik} trouvées, nous calculons la matrice de distances **D** entre toutes les séquences en appliquant l'équation 2. Dans le modèle de Jukes-Cantor, la distance *PEMV* non corrigée entre les séquences i et j se calcule comme suit :

$$d_{ij} = \frac{N_{ij}^{c} - N_{ij}^{m} \ + \ \sum_{k=1}^{N - N_{ij}^{c}} (1 - P_{ij}^{k})}{N}, \tag{2}$$

où d_{ij} – est la distance observée entre les séquences i et j ; N – est le nombre de sites considérés; N_{ij}^{m} - est le nombre de paires de nucléotides identiques dans i et j, N_{ij}^{c} - est le nombre de paires de nucléotides comparables (i.e. quand les deux nucléotides sont présents dans les sites correspondants) dans i et j ; P_{ij}^{k} - la probabilité, calculée en utilisant l'équation 1, d'avoir une paire de nucléotides identiques au site k dans i et j. La distance de Jukes-Cantor s'obtient à partir de d par l'application de la formule logarithmique.

Dans le cadre du modèle d'évolution de Kimura 2-paramètres nous calculons les taux de transition $P(i,j)$ et de transversion $Q(i,j)$ en utilisant l'équation 3 avant d'appliquer la formule logarithmique correspondante:

$$P(i,j) = \frac{P'(i,j) + \sum_{k=1}^{N-N_{ij}^{c}} P'(i,j,k)}{N}, \qquad Q(i,j) = \frac{Q'(i,j) + \sum_{k=1}^{N-N_{ij}^{c}} Q'(i,j,k)}{N}, \tag{3}$$

où $P'(i,j)$ – est le nombre de transitions entre les séquences i et j calculé en ignorant les sites incomplets; $P'(i,j,k)$ – est la probabilité de transition entre i et j au site k quand le nucléotide au site k est manquant soit dans i soit dans j; $Q'(i,j)$ – représente le nombre de transversions entre i et j calculé en ignorant les sites incomplets; $Q'(i,j,k)$ – est la probabilité d'obtenir une transversion entre i et j au site k lorsque ce nucléotide est manquant soit dans i soit dans j.

3 Application à des données réelles

La méthode introduite dans cet article a été utilisée pour reconstruire des arbres phylogénétiques inférés à partir des séquences de 10 espèces d'eucaryotes. Ces séquences de longueur 705 nucléotides chacune, ne contiennent pas de données manquantes et proviennent de l'ADN mitochondrial des espèces choisies. Elles sont disponibles sur la page web du Workshop on Molecular Biology à l'adresse URL suivante : <http://workshop.molecularevolution.org/resources/fileformats/fasta_dna_al.php>. La figure 1 présente deux arbres phylogénétiques possibles pour représenter l'évolution de ce groupe d'espèces et le pourcentage de bootstrap pour leurs branches internes. Le pourcentage de bootstrap permet d'estimer la robustesse de chacune des branches internes d'un arbre phylogénétique.

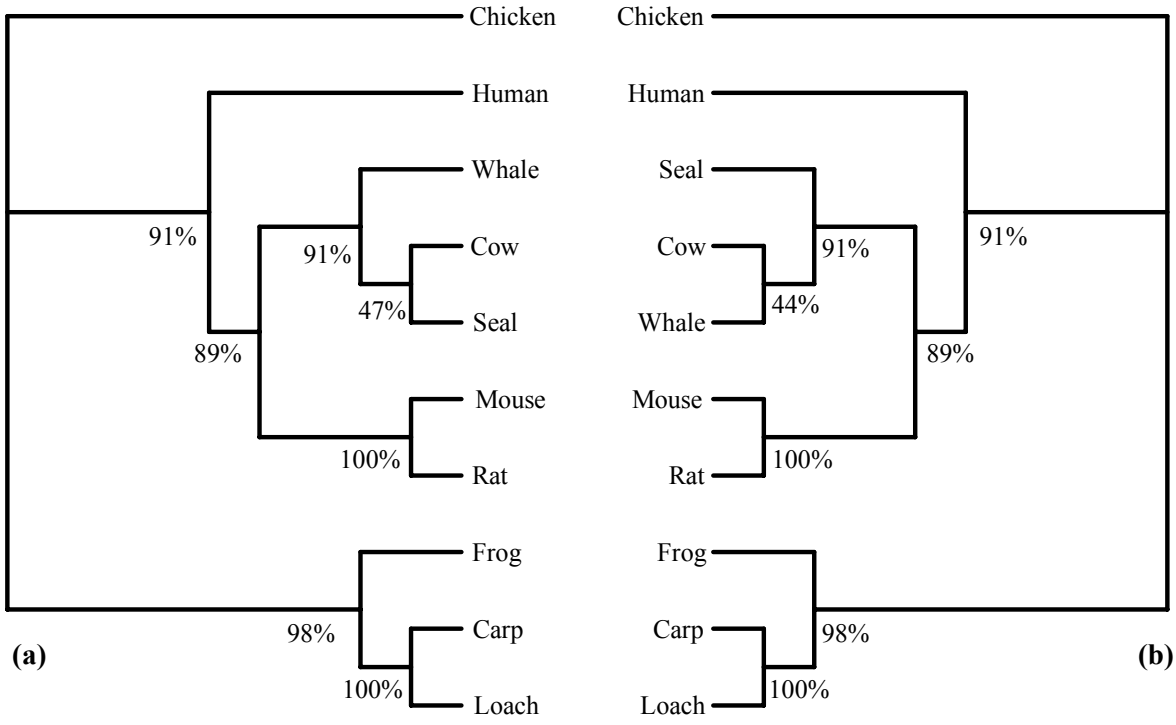


Figure 1. Deux arbres phylogénétiques inférés par la méthode NJ. Les séquences utilisées pour inférer ces arbres ne contiennent pas de données manquantes. Le modèle d'évolution de Kimura 2-paramètres a été utilisé pour calculer la matrice de distances entre les espèces.

Ces arbres sont légèrement différents au niveau du cluster regroupant les espèces Whale, Cow et Seal. Le premier arbre (figure 1a) soutient le groupe {Cow, Seal} avec le pourcentage de bootstrap de 47%, alors que le deuxième contient le cluster {Whale, Cow} dont le pourcentage de bootstrap est 44%.

Les deux topologies sont donc presque équiprobables. Donc, dans nos simulations nous les considérons comme les arbres initiaux corrects. Ces deux arbres ont été comparés aux phylogénies obtenues par les méthodes *IMS*, *PDMAB* et *PEMV* appliquées aux séquences d'ADN d'origine perturbées par l'ajout des données manquantes. Pour évaluer les performances des trois méthodes, de 0 à 80% de nucléotides ont été retirés des séquences initiales. Les nucléotides manquants ont été simulés par blocs pour mieux refléter la réalité génomique. Les séquences incomplètes obtenues, ont été soumises aux trois méthodes de calcul des matrices de distances d'évolution (*IMS*, *PDMAB* et *PEMV*) mentionnées ci-dessus. Pour chaque matrice de distance ainsi obtenue, nous avons reconstruit une phylogénie *T'* en utilisant la méthode NJ [SAI 87]. L'arbre phylogénique *T'* a été ensuite comparé à l'aide de la distance de Robinson et Foulds [ROB 81] aux deux arbres initiaux *T* présentés sur la figure 1 et la plus petite valeur (des deux valeurs possibles) de la distance de Robinson et Foulds a été retenue. Sur la figure 2, nous présentons les résultats moyens obtenus après 500 itérations (i.e. ensembles de séquences partielles différents). La figure 2a présente le pourcentage de la distance de Robinson et Foulds entre *T'* est le plus proche des deux arbres *T*. Deux arbres sont identiques si ce pourcentage est égal à 0 et leur différence s'accentue lorsqu'il augmente.

Les simulations fournissent des résultats similaires pour *IMS* et *PDMAB*. Pour tous les pourcentages de nucléotides manquants, *PEMV* assure une meilleure inférence phylogénétique comparativement aux méthodes de PAUP (figure 2a). Le nombre d'arbres identiques aux arbres initiaux était toujours plus élevé pour *PEMV* que pour *IMS* et *PDMAB* (figure 2b). Les performances de la nouvelle méthode sont les plus marquées pour 20, 30, 40 et 50% de valeurs manquantes.

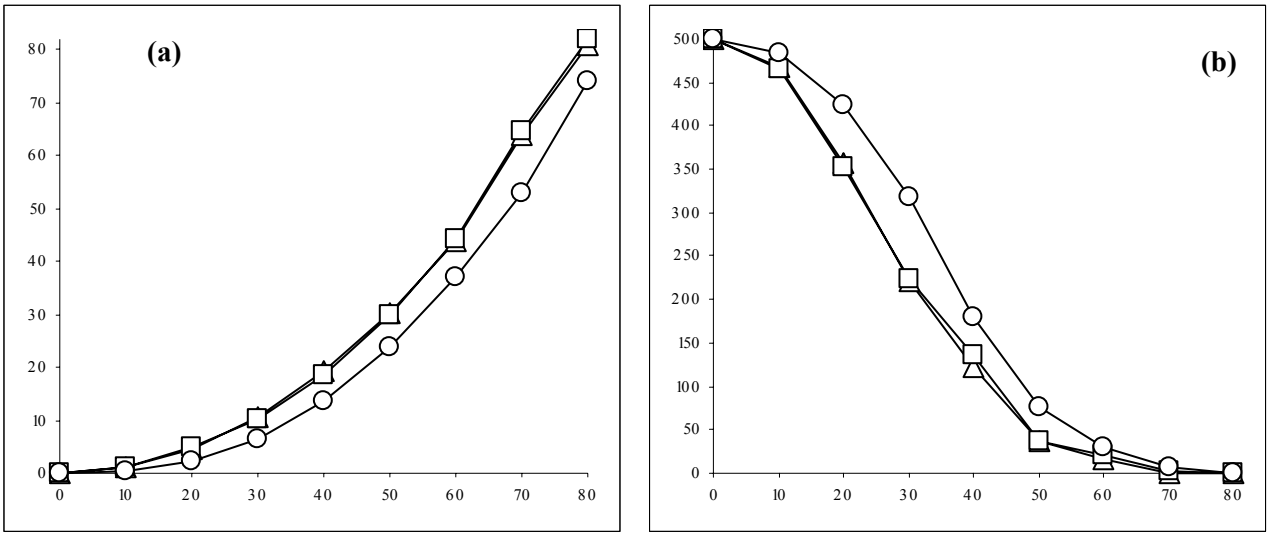


Figure 2. Les résultats moyens obtenus pour 500 itérations. Le pourcentage des données manquantes varie de 0 à 80% (axe des abscisses). Les courbes traduisent (a) les variations de la distance de Robinson et Foulds et (b) le nombre de fois que l'une des deux phylogénies initiales a été recouvrée. Les méthodes testées sont *IMS* (Δ), *PDMAB* (□) et *PEMV* (○).

4 Conclusion

Dans cet article nous avons décrit la méthode *PEMV* permettant d'estimer les probabilités des nucléotides manquants dans des séquences d'ADN et d'ARN en vue d'inférence d'un arbre phylogénétique. Dans nos simulations, le nombre de topologies identiques aux topologies initiales retrouvées par *PEMV* était supérieur à ceux des méthodes connues *PDMAB* et *IMS* de plus de 15, 30 et 25% pour, respectivement, 20, 30 et 40% de nucléotides manquants (figure 2b). Dans ces situations, l'élimination de sites manquants, comme le préconise la méthode *IMS*, ou leur estimation par la méthode *PDMAB* supprime des spécificités importantes des données traitées. La méthode *PEMV* a été incluse dans le logiciel T-Rex [MAK 01] disponible à l'URL suivant : <http://www.info.uqam.ca/~makarenv/trex.html>. Cette méthode a été

présentée ici dans le cadre des modèles de Jukes-Cantor [JUK 69] et de Kimura 2-parameter [KIM 80]. Il serait intéressant de continuer le développement de cette approche en la généralisant pour d'autres modèles d'évolution connus tels que : Tajima – Nei [TAJ 84], LogDet [STE 94] et d'autres. Il serait également nécessaire de comparer les performances de la stratégie *PEMV* à celles de maximum de vraisemblance et de maximum de parcimonie.

5 Bibliographie

[HUE 91] HUELSENBECK, J. P., "When are fossils better than existent taxa in phylogenetic analysis?" *Sys. Zool*., vol. 40, 1991, p. 458-469.

[JUK 69] JUKES, T. H., CANTOR, C.. "Mammalian Protein Metabolism", *chapter Evolution of protein molecules*, 1969, p. 21-132. Academic Press, New York.

[KIM 80] KIMURA, M., "A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences", *Jour. of Mol. Evol.,* vol. 16, 1980, p. 111-120.

[MAK 01] MAKARENKOV, V., "T-REX: reconstructing and visualizing phylogenetic trees and reticulation networks", *Bioinformatics*, vol. 17, 2001, p. 664-668.

[MAK 04] MAKARENKOV, V., LAPOINTE, F-J., "A weighted least-squares approach for inferring phylogenies from incomplete distance matrices", *Bioinformatics*, vol. 20, 2004, p. 2113-2121.

[ROB 81] ROBINSON D. R., FOULDS L. R. "Comparison of phylogenetic trees", *Mathematical Biosciences,* vol. 53, 1981, p. 131-147.

[SAI 87] SAITOU, N., M. NEI. "The neighbor-joining method: A new method for reconstructing phylogenetic trees", *Mol. Biol. Evol.* , vol. 4, 1987, p. 406-425.

[STE 94] STEEL, M. A. "Recovering a tree from the leaf colorations it generates under a Markov model", *Applied Math Letters*, vol. 72, 1994, p.19-24.

[SWO 01] SWOFFORD, D. L. "*PAUP**. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4.", *Sinauer Associates*, 2001, Sunderland, Massachusetts.

[TAJ 84] TAJIMA F, NEI M. "Estimation of evolutionary distance between nucleotide sequences." *Mol. Biol. Evol*., vol. 3, 1984, p. 269-285.

Caractérisation des ensembles critiques d'une famille de Moore finie

Jean Diatta

IREMIA, Université de la Réunion
15 avenue René Cassin - BP 7151
97715 Saint-Denis messag cedex 9, France
Jean.Diatta@univ-reunion.fr

RÉSUMÉ. *Nous donnons une condition nécessaire et suffisante pour qu'une partie d'un ensemble soit un quasi-fermé d'une famille de Moore finie sur cet ensemble. Nous situons cette condition par rapport à une autre, que l'on peut trouver dans un papier récent de Caspard et Monjardet. De plus, nous caractérisons les ensembles critiques d'une famille de Moore finie. Par ailleurs, nous montrons que l'identification des ensembles critiques des familles de Moore faiblement hiérarchiques sur un ensemble fini et contenant tous les singletons de cet ensemble, obtenue par Domenach et Leclerc, est une conséquence de notre résultat de caractérisation. Nous généralisons aussi cette identification aux familles de Moore k-faiblement hiérarchiques.*

MOTS-CLÉS : *Famille de Moore, Ensemble critique, Ensemble quasi-fermé, Hiérarchie faible, Opérateur de fermeture*

1. Introduction

La notion de famille de Moore est présente dans plusieurs domaines dont ceux mentionnés dans un papier récent de Caspard et Monjardet [CAS 03], et parmi lesquels on peut citer la classification, les bases de données relationnelles et la fouille de données. Cette omniprésence se traduit dans une certaine mesure par une grande variété de formalisations de cette notion selon les domaines dans lesquels elle est considérée : opérateurs de fermeture, treillis, systèmes complets d'implications, etc.. L'importance des applications de cette notion a conduit à de nombreux résultats intéressants dont, par exemple, l'existence d'une base canonique, prouvée initialement en termes de systèmes d'implications [GUI 86]. Cette base canonique est caractérisée par des quasi-fermés particuliers appelés ensembles critiques.

Dans cette communication, nous donnons une condition nécessaire et suffisante pour qu'une partie d'un ensemble soit un quasi-fermé d'une famille de Moore finie sur cet ensemble. Nous situons cette condition par rapport à une autre, que l'on peut trouver dans [CAS 03]. De plus, nous caractérisons les ensembles critiques d'une famille de Moore finie. Par ailleurs, nous montrons que l'identification des ensembles critiques des familles de Moore faiblement hiérarchiques sur un ensemble fini et contenant tous les singletons de cet ensemble, obtenue par Domenach et Leclerc [DOM 04], est une conséquence de notre résultat de caractérisation. Nous généralisons aussi cette identification aux familles de Moore k-faiblement hiérarchiques.

2. Familles de Moore et opérateurs de fermeture

Soit E un ensemble. Une *famille de Moore* sur E est une partie $\mathcal{F}$ de l'ensemble $\mathcal{P}(E)$ des parties de E, vérifiant :

(M1) $E \in \mathcal{F}$;

(M2) $\mathcal{F}' \subseteq \mathcal{F} \Longrightarrow \cap \mathcal{F}' \in \mathcal{F}$.

Si $\mathcal{F}$ est finie, donc si E est fini, alors la condition (M2) peut être remplacée par : $X, Y \in \mathcal{F} \Longrightarrow X \cap Y \in \mathcal{F}$. Par ailleurs, étant donnée un famille de Moore $\mathcal{F}$ sur E, l'application $\phi_{\mathcal{F}}$ définie sur $\mathcal{P}(E)$ par $\phi_{\mathcal{F}}(X) = \cap\{Y \in \mathcal{F} : X \subseteq Y\}$ est un *opérateur de fermeture* sur $\mathcal{P}(E)$, c'est-à-dire qu'elle vérifie les propriétés suivantes :

(C1) $X \subseteq \phi_{\mathcal{F}}(X)$;

(C2) $X \subseteq Y \Longrightarrow \phi_{\mathcal{F}}(X) \subseteq \phi_{\mathcal{F}}(Y)$;

(C3) $\phi_{\mathcal{F}}(\phi_{\mathcal{F}}(X)) = \phi_{\mathcal{F}}(X)$.

Pour un opérateur de fermeture ϕ sur $\mathcal{P}(E)$ et pour $X \subseteq E$, $\phi(X)$ est appelé la *fermeture* de X (selon ϕ).

Réciproquement, étant donné un opérateur de fermeture ϕ sur E, la collection $\mathcal{F}_{\phi}$ de sous-ensembles de E, définie par $\mathcal{F}_{\phi} = \{X \subseteq E : \phi(X) = X\}$ est une famille de Moore sur E. Il est en effet bien connu que les notions de famille de Moore et d'opérateur de fermeture sont deux notions équivalentes. D'autres formalisations de ce même et unique concept, ainsi que des liens entre elles, peuvent être trouvées dans [CAS 03]. Les éléments d'une famille de Moore ou, de manière équivalente, les points fixes d'un opérateur de fermeture, sont appelés des fermés.

3. Ensembles critiques

Dans ce qui suit, les ensembles finis seront parfois notés comme des mots : par exemple ab désignera la paire $\{a, b\}$; le contexte permettra de distinguer les élements de l'ensemble des singletons. On dira qu'un ensemble X intersecte proprement un ensemble Y si $X \cap Y \neq \varnothing$, $X \setminus Y \neq \varnothing$ et $Y \setminus X \neq \varnothing$.

Soit E un ensemble quelconque et $\mathcal{F}$ une famille de Moore finie sur E. Une partie Q de E est appelée un ensemble *quasi-fermé* de $\mathcal{F}$ si $Q \notin \mathcal{F}$ et $\mathcal{F} \cup \{Q\}$ est une famille de Moore sur E. Lorsque $\mathcal{F}$ est implicite dans le contexte, on dira simplement que Q est (un ensemble) quasi-fermé au lieu de Q est (un ensemble) quasi-fermé de $\mathcal{F}$. Etant donné $F \in \mathcal{F}$, Q est un ensemble F-quasi-fermé de $\mathcal{F}$ si Q est quasi-fermé et $\phi_{\mathcal{F}}(Q) = F$.

EXEMPLE 1 *Soit $E = \{a, b, c, d, e\}$ et $\mathcal{F} = \{\varnothing, a, b, d, de, bcd, abcde\}$. Alors ab est quasi-fermé. Par contre, ae n'est pas quasi-fermé car $de \cap ae = e \notin \mathcal{F} \cup ae$.*

Une partie K de E est un ensemble *critique* de $\mathcal{F}$ si K est $\phi_{\mathcal{F}}(K)$-quasi-fermé minimal [DAY 92]. Des ensembles critiques triviaux sont spécifiés par la proposition immédiate suivante.

PROPOSITION 1 *Tout élément minimal de $\mathcal{P}(E) \setminus \mathcal{F}$ est critique.*

EXEMPLE 2 *Considérons la famille de Moore de l'exemple 1. Alors c est critique. Par contre bc est quasi-fermé, mais n'est pas critique car $\phi_{\mathcal{F}}(c) = \phi_{\mathcal{F}}(bc) = bcd$.*

REMARQUE 1 *Si $Q \subset E$ est un ensemble quasi-fermé de $\mathcal{F}$ ayant k éléments, alors aucun sur-ensemble de Q ayant $(k+1)$ éléments n'est critique. En particulier, si $\varnothing$ n'est pas fermé, alors aucun singleton de E n'est critique.*

Le résultat ci-dessous identifie les ensembles quasi-fermés d'une famille de Moore finie avec les sous-ensembles non fermés qui n'intersectent proprement la fermeture d'aucune de leurs parties propres.

THÉORÈME 1 *Une partie Q de E est quasi-fermée si et seulement si $Q \notin \mathcal{F}$ et pour tout $X \subset Q$, ou bien $\phi_{\mathcal{F}}(X) \subset Q$ ou bien $Q \subset \phi_{\mathcal{F}}(X)$.*

Considérons la caractérisation suivante, dont on peut trouver une preuve dans [CAS 03] : une partie Q de E est quasi-fermée si et seulement si $Q \notin \mathcal{F}$ et pour tout $X \subset Q$, $\phi_{\mathcal{F}}(X) \subset \phi_{\mathcal{F}}(Q)$ implique $\phi_{\mathcal{F}}(X) \subset Q$; en d'autres termes, Q est quasi-fermée si et seulement si $Q \notin \mathcal{F}$ et pour tout $X \subset Q$, ou bien $\phi_{\mathcal{F}}(X) = \phi_{\mathcal{F}}(Q)$ ou bien $\phi_{\mathcal{F}}(X) \subset Q$. L'équivalence entre cette caractérisation et celle que nous proposons vient du fait que si

$Q \subset \phi_{\mathcal{F}}(X)$, alors $\phi_{\mathcal{F}}(Q) \subseteq \phi_{\mathcal{F}}(X)$, de telle sorte que $\phi_{\mathcal{F}}(X) = \phi_{\mathcal{F}}(Q)$ puisque $X \subset Q$. L'avantage de la formulation proposée dans cette communication est essentiellement algorithmique car elle montre que pour $X \subset Q$ avec $\phi_{\mathcal{F}}(X) \not\subseteq Q$, il suffit de vérifier si $Q \subset \phi_{\mathcal{F}}(X)$ (on n'a pas besoin de savoir si $\phi_{\mathcal{F}}(X) = \phi_{\mathcal{F}}(Q)$).

Une conséquence immédiate du Théorème 1 est la condition suffisante, ci-dessous, pour qu'une partie d'un ensemble fini soit critique par rapport à une famille de Moore sur cet ensemble.

COROLLAIRE 1 *Si E est fini, alors toute partie $K \in \mathcal{P}(E) \setminus \mathcal{F}$ de E telle que pour tout $X \subset K$ avec $|X| = |K| - 1$ (où $|Y|$ est le nombre d'éléments de Y), $\phi_{\mathcal{F}}(X) \subset K$, est critique.*

La combinaison du résultat du Théorème 1 et du fait que les ensembles critiques sont, par définition, des quasi-fermés minimaux parmi ceux de même fermeture, conduit à la caractérisation suivante des ensembles critiques d'une famille de Moore finie.

THÉORÈME 2 *Une partie K de E est critique si et seulement si $K \notin \mathcal{F}$ et pour tout $X \subset K$, ou bien $\phi_{\mathcal{F}}(X) \subset K$ ou bien $K \subset \phi_{\mathcal{F}}(X)$ et il existe $Y \subset X$ tel que $\phi_{\mathcal{F}}(Y)$ intersecte proprement X.*

Mentionnons aussi la caractérisation récursive des ensembles critiques (voir [CAS 99]), souvent utilisée pour la construction algorithmique de ceux-ci : une partie d'un ensemble fini E est un ensemble critique d'une famille de Moore sur E si et seulement si elle contient la fermeture de chacun de ses sous-ensembles propres critiques.

4. Familles de Moore (k-faiblement) hiérarchiques

Dans tout ce qui suit, E est un ensemble fini. Une famille de Moore sur E sera dite *hiérarchique* si deux quelconques X, Y de ses éléments sont toujours soit disjoints, soit emboîtés, c'est-à-dire, $X \cap Y \in \{\varnothing, X, Y\}$.

EXEMPLE 3 Soit $E = \{a, b, c, d, e\}$. Alors $\mathcal{F} = \{\varnothing, a, b, cd, e, ae, bcd, abcde\}$ est une famille de Moore hiérarchique sur E.

On notera qu'une famille de Moore hiérarchique non réduite à une chaîne (c'est-à-dire dont les éléments ne sont pas deux-à-deux comparables au sens de l'inclusion) contient nécessairement l'ensemble vide comme l'un de ses éléments. Les familles de Moore hiérarchiques correspondent en fait aux bien connues hiérarchies, modifiées en ne rejetant pas l'ensemble vide. Dans le même esprit, d'autres structures de classification comme, par exemple, les hiérarchies faibles [BAN 89], donnent aussi lieu à des familles de Moore particulières.

Une famille de Moore sur E sera dite *faiblement hiérarchique* si l'intersection de trois quelconques X, Y, Z de ses éléments est toujours égale à l'intersection de deux parmi ces trois éléments, c'est-à-dire, $X \cap Y \cap Z \in \{X \cap Y, X \cap Z, Y \cap Z\}$.

EXEMPLE 4 Soit $E = \{a, b, c, d, e\}$. Alors $\mathcal{F} = \{a, ab, ac, ad, ae, abcde\}$ est une famille de Moore faiblement hiérarchique non hiérarchique. Elle n'est pas hiérarchique car $ab \cap ae \notin \{\varnothing, ab, ae\}$.

Il est clair que toute famille de Moore hiérarchique est aussi faiblement hiérarchique, mais contrairement aux familles de Moore hiérarchiques, une famille de Moore faiblement hiérarchique non réduite à une chaîne ne contient pas nécessairement l'ensemble vide (voir Exemple 4). Par ailleurs, les familles de Moore faiblement hiérarchiques se généralisent naturellement en les familles de Moore k-faiblement hiérarchiques, en considérant l'intersection de $(k+1)$ au lieu de l'intersection de trois ([BAN 94], [DIA 97], [Bar 01], [BER 03]).

Une famille de Moore sur E sera dite *k-faiblement hiérarchique* si l'intersection de $k+1$ quelconques $X_1, \ldots, X_k, X_{k+1}$ de ses éléments est toujours égale à l'intersection de k parmi ces $k+1$, i.e., s'il existe $i \in \{1, \ldots, k+1\}$ tel que $\bigcap_{j \neq i} X_j \subseteq X_i$. Les familles de Moore faiblement hiérarchiques correspondent au cas où $k = 2$.

EXEMPLE 5 *Soit* $E = \{a, b, c, d\}$. *Alors* $\mathcal{F} = \{\varnothing, a, b, c, d, ab, ac, ad, bc, bd, cd, abc, acd, bcd, abcd\}$ *est une famille de Moore* 3*-faiblement hiérarchique non faiblement hiérarchique. Elle n'est pas faiblement hiérarchique car, par exemple,* $ab \cap ac \cap bc \notin \{ab \cap bc, ab \cap ac, ac \cap bc\}$.

Les familles de Moore k-faiblement hiérarchiques ont la propriété intéressante suivante, à savoir qu'elles sont engendrées par des ensembles constitués de parties ayant au plus k éléments ([BAN 89] (pour $k = 2$), [BAN 94], [DIA 97]) :

PROPOSITION 2 *Une collection* $\mathcal{F}$ *de parties de* E *est une famille de Moore* k*-faiblement hiérarchique si et seulement si pour tout* $F \in \mathcal{F}$ *il existe* $X \subseteq F$ *tel que* $|X| \leq k$ *et* $\phi_{\mathcal{F}}(F) = \phi_{\mathcal{F}}(X)$.

Nous tirons de ce résultat la condition suivante, nécessaire pour qu'une partie ayant plus de k éléments soit critique dans une famille de Moore k-faiblement hiérarchique.

PROPOSITION 3 *Soit* $\mathcal{F}$ *une famille de Moore* k*-faiblement hiérarchique sur* E *et* $K \subset E$ *tel que* $|K| > k$. *Si* K *est critique, alors il existe* $X \subset K$, $|X| \leq k$, *tel que* X *n'est ni fermé ni quasi-fermé.*

De la Proposition 3 découle le résultat suivant obtenu par Domenach et Leclerc [DOM 04] dans le cas particulier des familles de Moore faiblement hiérarchiques ($k = 2$) :

PROPOSITION 4 *Soit* $\mathcal{F}$ *une famille de Moore* k*-faiblement hiérarchique sur* E, *contenant toutes les parties de* E *d'au plus* $k - 1$ *éléments. Alors* $X \subset E$ *est critique si et seulement si* $X \notin \mathcal{F}$ *et* $|X| = k$.

5. Bibliographie

[BAN 89] BANDELT H.-J., DRESS A. W. M., Weak hierarchies associated with similarity measures : an additive clustering technique, *Bull. Math. Biology*, vol. 51, 1989, p. 113–166.

[BAN 94] BANDELT H.-J., DRESS A. W. M., An order theoretic framework for overlapping clustering, *Discrete Mathematics*, vol. 136, 1994, p. 21–37.

[Bar 01] BARTHÉLEMY J.-P., BRUCKER F., NP-hard Approximation Problems in Overlapping Clustering, *Journal of Classification*, vol. 18, 2001, p. 159–183.

[BER 03] BERTRAND P., JANOWITZ M. F., The k-Weak Hierarchical Representations : an extension of the Indexed Closed Weak Hierarchies, *Discrete Applied Mathematics*, vol. 127, 2003, p. 199–220.

[CAS 99] CASPARD N., A characterization theorem for the canonical basis of a closure operator, *Order*, vol. 16, 1999, p. 227–230.

[CAS 03] CASPARD N., MONJARDET B., The lattices of closure systems, closure operators, and implicational systems on a finite set : a survey, *Discrete Applied Mathematics*, vol. 127, 2003, p. 241–269.

[DAY 92] DAY A., The lattice theory of functional dependencies and normal decompositions, *Internat. J. Algebra Comput.*, vol. 2, 1092, p. 409—431.

[DIA 97] DIATTA J., Dissimilarités multivoies et généralisations d'hypergraphes sans triangles, *Math. Inf. Sci. hum.*, vol. 138, 1997, p. 57–73.

[DOM 04] DOMENACH F., LECLERC B., Closure systems, implicational systems, overhanging relations and the case of hierarchical classification, *Mathematical Social Sciences*, vol. 47, 2004, p. 349–366.

[GUI 86] GUIGUES J. L., DUQUENNE V., Famille non redondante d'implications informatives résultant d'un tableau de données binaires, *Mathématiques et Sciences humaines*, vol. 95, 1986, p. 5–18.

Classification et détection d'habitats benthiques à l'aide de signatures sonores

Sébastien Durand, Pierre Legendre

Département de sciences biologiques,
Université de Montréal,
Case postale 6128, succursale Centre-ville,
Montréal (Québec) Canada, H3C 3J7

RÉSUMÉ. Étudier la nature des grands fonds marins sur de vastes étendues a toujours été une tâche fastidieuse même en utilisant un sous-marin muni de lampes puissantes et de caméras. L'acquisition de données visuelles détaillées est limitée à de petites superficies à cause du manque de visibilité. Grâce à la télédétection sonore, l'étude spatiotemporelle des milieux benthiques profonds est dorénavant à la portée des écologistes. Nous avons élaboré une méthode d'analyse permettant de classifier les ondes sonores et de relier les groupes ainsi formés à des types d'habitats. Nous décrirons en détail certains aspects de la méthode : d'abord comment identifier les variables les plus représentatives des signaux sonores, puis déterminer l'influence de différentes variables du substrat marin sur les signaux sonores enregistrés.

MOTS-CLÉS : Analyse discriminante canonique, écho sonore, habitat, radiale Juan de Fuca, sonar, source hydrothermale, submersible téléguidé, télédétection.

1 Introduction

Depuis longtemps, l'homme tente de comprendre la nature des ondes sonores. Comme le font plusieurs autres espèces animales, nous utilisons le son afin de communiquer entre nous. Chez certains mammifères, le son sert aussi à des fins de navigation et même à l'identification d'objets éloignés. Le fond des abysses océaniques, composé de roc et de sédiments, baigne dans un milieu qui absorbe fortement la lumière. On y trouve des communautés animales et bactériennes qui vivent dans l'obscurité. L'aptitude des chauves-souris et des dauphins à détecter des objets à distance grâce à des échos sonores nous montre que les diverses textures et densités propres aux aires échantillonnées affectent d'une façon significative la nature des échos sonores qui peuvent être enregistrés par nos appareils. Par l'analyse de variables décrivant la forme et l'intensité de ces signaux, nous chercherons à identifier à distance les habitats d'un champ hydrothermal et, éventuellement, à cartographier les habitats du milieu benthique sur de grandes surfaces.

Il n'existe pas de méthode standard pour l'extraction, la transformation et le traitement des variables tirées de signaux sonores. Cela n'est pas surprenant. D'une part, les méthodes d'interprétation d'ondes sonores ne sont encore qu'en phase exploratoire ou sont gardées secrètes par des compagnies privées. D'autre part, de nombreux types d'habitats marins n'ont pas encore été étudiés et de nombreuses fréquences et tailles d'empreintes sonores n'ont pas encore été utilisées. Il serait donc difficile d'établir des standards pour une méthodologie aussi jeune. Le présent article constitue une introduction à l'approche analytique de ce nouveau champ d'études. Il décrit la méthode d'analyse des échos sonores que nous avons développée et utilisée dans nos travaux.

Au cours de nos travaux récents [DUR soumis] basés sur ceux de Clarke et Hamilton [CLA 99], nous avons [1] testé diverses méthodes de transformation des données ; [2] démontré visuellement l'importance de normaliser les données sonores en fonction de leur altitude d'acquisition (distance au fond marin); [3] démontré que la variation de la taille des empreintes sonores influe sur les capacités discriminantes des ondes sonores ; et [4] créé un jeu de variables permettant la discrimination efficace des signatures sonores acquises dans les habitats dominants qui se retrouvent à l'intérieur de notre site d'étude.

2 Méthode

2.1 Acquisition des données

Les données analysées dans cet article ont été récoltées par le ROPOS, un sous-marin canadien télécommandé depuis un navire de la garde côtière canadienne, le John P. Tully. Au cours de la mission *High Rise* en mai 2001, nous avons étudié le champ de sources hydrothermales Clambed situé par 2200 m de fond sur la radiale Juan de Fuca, à quelque 200 miles marins à l'ouest de l'île de Vancouver dans le Pacifique. Les caméras du ROPOS ont filmé les habitats de ce champ hydrothermal en même temps que nous récoltions des signaux sonores à l'aide d'un sonar Imagenex 881B. Le plan d'échantillonnage comportait 18 transects verticaux (1 à 10 m d'altitude) visant les cinq communautés principales ainsi que 12 transects horizontaux (1 à 5 m d'altitude) couvrant les principaux gradients naturels de ces milieux profonds. La fréquence d'échantillonnage moyenne était de trois signaux sonores par seconde.

2.2 Traitement des données et extraction des variables

Puisque les ondes sonores sont affectées par des phénomènes aléatoires comme le bruit ambiant, l'instabilité des capteurs du ROPOS et du sonar, ainsi que par la variabilité naturelle des signaux, le calcul de la moyenne de plusieurs échos successifs augmente la stabilité du signal sonore [HAM 99] et permet aussi l'obtention de données représentant un intervalle temporel choisit par le chercheur. Après avoir calculé la moyenne des signaux, on extrait une première variable, l'altitude, car c'est à partir de cette variable que nous effectuerons tous les filtrages initiaux, corrections et normalisation pour la profondeur. Puisqu'il n'existe aucune méthode standard d'extraction de la variable altitude, nous avons créé notre propre algorithme. Ainsi, à l'intérieur d'un écho sonore, nous avons utilisé comme marqueur temporel servant à l'estimation de l'altitude d'acquisition le point correspondant à 80 % de la valeur d'intensité maximum de l'onde sonore temporairement lissée.

Afin de permettre la comparaison des informations extraites de différents signaux sonores, une normalisation en fonction de l'altitude est nécessaire. En fait, lorsqu'une onde sonore se déplace dans l'eau, son intensité diminue, car non seulement l'onde s'étire de façon sphérique, mais l'eau servant de support absorbe aussi une partie de l'énergie de l'onde. Dans la plupart des cas, le taux d'absorption exacte au site d'échantillonnage n'est pas connu. Il est donc nécessaire d'en ajuster la valeur. Dans ce dessein, nous avons dessiné tous les signaux sonores côte à côte et attribué des niveaux de gris aux valeurs d'intensité qui les composent, puis nous avons comparé visuellement les signaux sonores normalisés acquis à des altitudes différentes. Nous avons modifié les valeurs de taux d'absorption jusqu'à ce que les niveaux de gris attribués aux signaux obtenus à différentes altitudes soient uniformes. Nous avons alors considéré que nous avions ajusté les taux d'absorption sonore.

Puisque nous extrayons des variables du premier et du deuxième écho enregistrés (première et deuxième réflexion sur le substrat), tous les échantillons sonores n'incluant pas ce dernier segment ont été éliminés d'entrée de jeu. Puisqu'aucun standard industriel n'indique quelles variables permettent une description optimale d'un signal sonore, nous avons dans un premier temps basé notre choix sur des variables facilement interprétables et, dans un deuxième temps, tenté de conserver les variables les plus différentes les unes des autres. Deux jeux de variables ont été utilisés et testés. Le premier décrit principalement la forme du premier et du deuxième écho par l'entremise de variables extraites (VE) telles que les coefficients d'aplatissement et d'asymétrie, la distance entre les centroïdes de diverses sections de l'écho,

l'aire sous la courbe, etc. Le deuxième jeu de données (variables d'intensité, VI) utilise la série temporelle d'intensités centrées réduites du premier écho créée lors de la numérisation du signal sonore.

Pour chacune des 28 variables extraites (VE), nous avons sélectionné, parmi 4 transformations possibles, celle qui produisait les distributions les moins asymétriques. La transformation retenue différait d'une variable à l'autre. Pour l'ensemble des 92 variables d'intensité (VI), une seule transformation a été retenue parmi 11 transformations testées [DUR soumis] pour transformer les variables k_j en k_j' :

$$p_j = k_j^{0.25} / max(k^{0.25})$$

$$k_j' = \text{arcsinus}(p_j^{0.5})$$

($k_j^{0.25}$) représente la racine quatrième des valeurs d'une variable d'intensité (k_j) décrivant un écho, suivie d'une transformation en proportion (p_j) par rapport à l'intensité maximale du signal $max(k^{0.25})$, puis de la transformation arcsinus($p_j^{0.5}$) [SOK 95]. Puisque nous avons divisé l'intensité des signaux sonores par leur valeur maximale, cette transformation ne conserve que la forme des échos comme source de variation.

3 Analyses et résultats

3.1 Transects verticaux

Afin d'évaluer le pouvoir de discrimination des variables formant les deux jeux de données (VI et VE), nous avons utilisé les enregistrements vidéo pour attribuer à chacun des signaux sonores un nom d'habitat parmi les cinq habitats dominants observés au site hydrothermal Clambed. Nous avons supprimé toutes les variables sonores ayant une variance intragroupe nulle ou une trop faible variance.

L'analyse des jeux de variables VI, VE, ainsi que l'union de ces deux groupes, fut réalisée de la façon suivante. 70 % des échantillons sonores (5352 signaux), tirés au hasard dans chaque habitat, furent utilisés pour construire un modèle prédictif à l'aide d'analyses discriminantes linéaires. Le pouvoir discriminant de chaque modèle fut évalué en termes de pourcentage de classification correcte des données restantes, soit 30 % ou 2293 signaux. En comparant les pourcentages de classification correcte obtenus pour les trois jeux de données, nous avons noté que l'utilisation conjointe des deux jeux de données dans la même analyse (VI et VE) pouvait améliorer les prédictions de 10 %. Puis, utilisant seulement ce jeu de variables, les signaux sonores furent séparés en fonction de leur altitude d'acquisition afin d'évaluer les effets qu'ont les variations de la taille de l'empreinte sonore sur la capacité de discrimination. À l'aide des 5706, 1499 et 441 échantillons retrouvés respectivement entre 1 et 4, 4 et 7, et 7 et 10 mètres. Les pourcentages de classification correcte obtenus étaient faibles pour les basses altitudes (70.4 et 68.4 %) et plus élevés pour les signaux obtenus au-delà de 7 mètres (84 %, Fig. 1). Cette grande différence supporte l'hypothèse que les variations de la taille de l'empreinte peuvent grandement influencer les résultats d'un survol sonore.

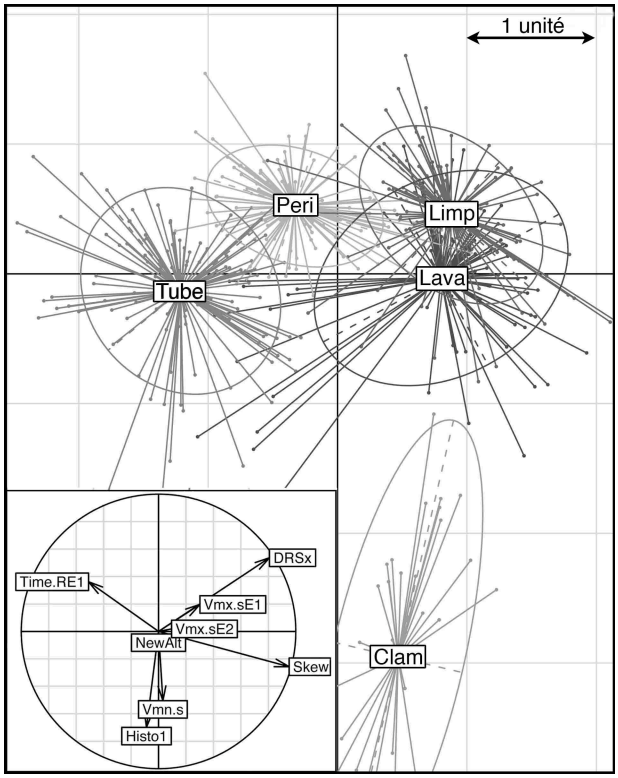


Figure 1. 30 % des échos échantillonnés à une altitude de 7 à 10 mètres projetés sur les deux premiers axes d'un modèle discriminant prédictif. Les variables VE dont les contributions aux axes (corrélations) sont illustrées en médaillon induisent la séparation des 5 habitats. L'espace discriminant est divisé en carrés de taille unité.

Afin de faciliter l'interprétation des résultats, après sélection ascendante pas à pas parmi les 120 variables dont nous disposions au départ, nous avons retenu 16 variables contribuant significativement à la discrimination. Malgré cette stricte sélection, les nouveaux pourcentages de classification correcte obtenus n'affichaient en moyenne qu'une baisse de 3 %. L'analyse plus approfondie des variables sélectionnées nous a permis non seulement d'identifier les caractéristiques propres à chaque signature sonore, mais aussi d'interpréter les signatures sonores en termes de texture et de densité.

3.2 Transect horizontaux

Après avoir démontré les capacités discriminantes des ondes sonores, il nous restait à montrer la capacité réelle des données sonores pour l'identification correcte des différents types d'habitats lors de survols horizontaux du champ de sources hydrothermales. Puisque nous avions le projet d'identifier visuellement les signatures sonores enregistrées en vue de l'analyse canonique de redondance (ACR), les transects furent survolés à une altitude variant d'un à cinq mètres ; une altitude plus élevée n'aurait pas permis d'identifier les habitats avec suffisamment de précision sur les enregistrements vidéo. À cette distance, le cône sonore sur le fond marin formait une empreinte variant de trois à quinze centimètres de diamètre. Au cours des 12 transects parcourant le site d'étude, nous avons enregistré en simultanéité 13676 signaux sonores et 90 minutes de survol vidéo. Dans un premier temps, nous avons calculé la moyenne des échos regroupés par intervalles d'une seconde et avons réalisé une description visuelle de la nature de l'aire de l'empreinte sonore, seconde par seconde. Une fois les données transformées et normalisées pour la profondeur et les variables sonores extraites, nous avons tenté de découvrir à quelles caractéristiques visuelles les différentes variables sonores semblaient être les plus sensibles. Par sélection ascendante pas à pas, un sous-groupe de variables visuelles fut extrait, puis utilisé dans une analyse de redondance qui a permis d'identifier à quels éléments du paysage sous-marin correspondaient les différentes variables sonores. Parce que les données sonores contiennent beaucoup de bruit, le coefficient de redondance bimultivariable de l'ACR est faible (R^2 = 10.4 %). Il reste très hautement significatif (P = 0.001 après 1000 permutations). Les relations entre les variables visuelles et sonores illustrées à la Figure 2 sont en accord avec les relations habitats-variables sonores trouvées lors de l'analyse des transects d'altitude [DUR soumis].

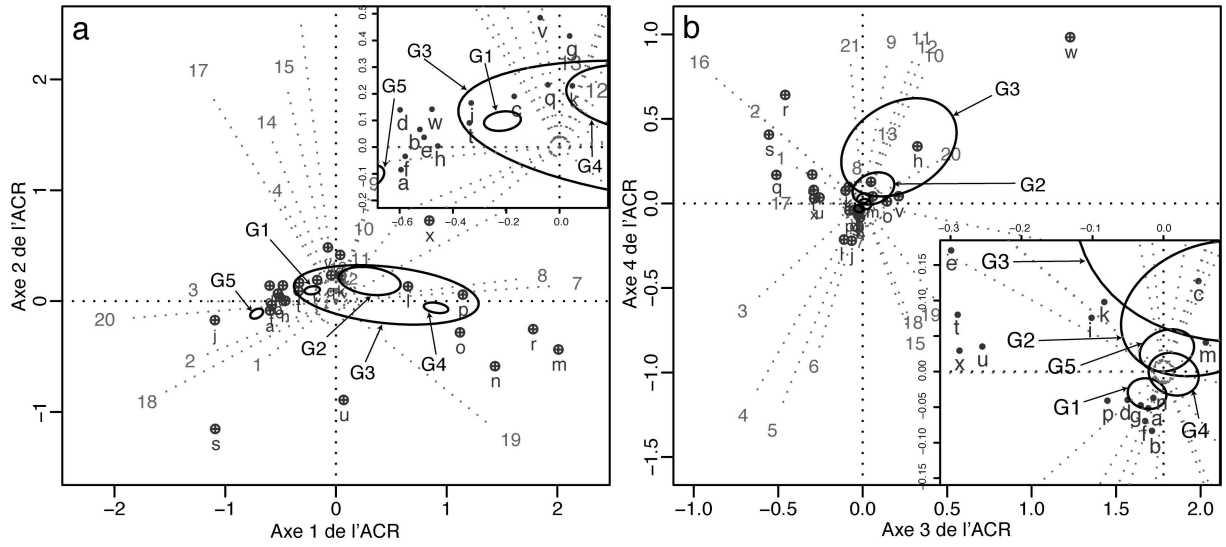


Figure 2. Projection des 21 variables sonores (variables VI : nos 1-13 ; variables VE : nos 14-21) et des 24 variables visuelles binaires (a à x) dans l'espace (a) des axes canoniques 1-2 et (b) des axes 3-4. Les cinq groupes (G1-G5) résultant de la partition par la méthode des *K* centroïdes sont représentés dans les graphiques par des ellipses de recouvrement à 95 %. À l'intérieur de chaque graphique, la zone centrale surchargée a été agrandie.

Nous avons ensuite partitionné les échos en 5 groupes de signaux sonores par la méthode des *K* centroïdes (*K-means*) [MAC 67] et retenu la solution présentant la plus faible variation intragroupe après 1000 démarrages aléatoires. Chaque groupe est représenté dans la Figure 2 par une ellipse de recouvrement à 95 %. Nous pouvons déterminer le type d'habitat représenté par chaque groupe (G1 à G5) en associant la position des ellipses aux variables visuelles (a à x) dans la figure. Cette étape était nécessaire pour nous permettre de cartographier les habitats détectés. Nos données de navigation contiennent cependant beaucoup d'erreur. Nous cherchons en ce moment à corriger ces données afin de pouvoir représenter sur une carte les différents groupes sonores que nous avons obtenus de la partition.

4 Discussion et conclusion

En nous basant sur les résultats obtenus au cours de cette étude exploratoire, nous croyons que l'utilisation du sonar à des fins de cartographie et de télédétection représente une avancée majeure pour l'étude des habitats benthiques. À ce jour, le plus grave problème relié à cette technologie est que les ondes enregistrées dans un même habitat sont très variables. De nombreuses combinaisons de variables à extraire des ondes sonores sont possibles; elles n'ont pas encore été toutes créées ou testées. Afin de standardiser la méthode, nous devons chercher à identifier les combinaisons décrivant le plus adéquatement les divers échos sonores en vue de la cartographie des fonds marins. Parce que les basses fréquences pénètrent davantage le substrat, l'utilisation de plusieurs fréquences pourrait faire varier les signatures sonores provenant d'un même habitat et enrichir nos jeux de variables. Malgré l'augmentation du temps de calcul qui en résultera, l'utilisation conjointe de différentes fréquences sonores devrait permettre l'identification plus efficace des habitats benthiques, comme c'est le cas pour les différentes longueurs d'onde utilisées en télédétection terrestre.

Une des grandes sources de variance entre les échos est la variation de la taille de l'empreinte sonore sur le fond marin. Ces changements de taille de l'unité d'échantillonnage peuvent réduire grandement la capacité discriminante d'un jeu de données. Il est impératif que tout survol sonore respecte cette source de variation et tente de la minimiser. Ce ne sera qu'une fois cette méthode de télédétection bien établie et standardisée, que nous pourrons construire une banque de signatures sonores pouvant servir à l'identification des types d'habitats. Cela minimiserait les coûts associés à la validation visuelle des données sonores enregistrées au cours d'une mission.

5 Bibliographie

[CLA 99] CLARKE P.A., HAMILTON L.J., *The ABCS Program for the analysis of echo sounder returns for acoustic bottom classification*, Report DSTO-GD-0215, Defence Science & Technology Organisation, Commonwealth of Australia, 1999.

[DUR soumis] DURAND S., LEGENDRE P., JUNIPER S.K., "Sonar backscatter differentiation of dominant macrohabitat types in a hydrothermal vent field", *Ecological Applications*, soumis.

[HAM 99] HAMILTON L.J., MULHEARN P.J., POECKERT R., "Comparison of RoxAnn and QTC-View acoustic bottom classification system performance for the Cairns area, Great Barrier Reef, Australia", *Continental Shelf Research* vol. 19, 1999, p. 1577-1597.

[MAC 67] MacQUEEN J., "Some methods for classification and analysis of multivariate observations", *in : Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Le Cam L.M., Neyman J., editors, University of California Press, Berkeley, vol. 1, p. 281-297, 1967.

[SOK 95] SOKAL R.R., ROHLF F.J., *Biometry – The principles and practice of statistics in biological research, Third Edition*, W.H. Freeman, New York, 1995.

Prise en compte de la durée de séjour dans la classification de données biographiques.

Estacio-Moreno Alexander[1,2], Artières Thierry[1], Gallinari Patrick[1]

[1]*Laboratoire d'informatique de Paris 6,*
Université Paris 6,
8 rue du Capitaine Scott
75015 Paris

[2]*UR 079 - 013,*
Institut de Recherche pour le Développement,
32 ave. Henri Varagnat
93143 Bondy CEDEX

RÉSUMÉ.

La mobilité, dans ses différentes dimensions (résidentielle, professionnelle, etc.), caractérise et différencie les individus et les groupes sociaux et devient un élément central pour l'analyse et la compréhension des dynamiques et des recompositions urbaines. Cependant, l'analyse des données biographiques qui décrivent les différentes formes de mobilité pose encore d'importants problèmes méthodologiques.

Nous présentons ici une méthode pour faire de la classification de données biographiques, utilisant un mélange de densités. Nous proposons d'utiliser des modèles semi-markoviens pour prendre en compte la durée de séjour dans les états. Le cadre est assez général pour qu'il soit appliqué à n'importe quel type de données séquentielles, où la durée de séjour dans les états est importante. Nous détaillons enfin l'application de cette méthode à l'étude de la mobilité résidentielle à partir des données d'une enquête rétrospective.

MOTS-CLÉS : Données biographiques, mobilité résidentielle, mélange de densités, modèles semi-markoviens.

1 Introduction

Les sources statistiques principales pour l'étude fine des phénomènes qui reposent sur les comportements démographiques, économiques et sociaux individuels et collectifs et de leur dynamique à différentes échelles spatiales et temporelles (mobilités) sont les enquêtes biographiques rétrospectives. Dans ces enquêtes sont recueillies, sur un échantillon d'individus, **des trajectoires** définies par les changements d'état des variables résidentielles, professionnelles, d'événements familiaux, etc.

Par l'analyse de données sur ces mobilités (données biographiques) on cherche à décrire et relier entre elles les différentes formes de mobilité pour comprendre leurs interactions et leur impact sur la réalité sociale. Ces dernières années ont vu des avancées significatives dans l'analyse des données biographiques. En statistique on peut utiliser une approche modélisatrice, avec des modèles log-linéaires, des modèles logit et probit, des modèles de survie de Cox ... ([COX 84, COU89]). Si l'on s'intéresse, par exemple, aux trajectoires résidentielles d'une certaine population, cette approche permet de répondre à la question : quelles sont les déterminants de l'ascension socio-résidentielle ? Egalement, on peut utiliser l'analyse typologique, qui est basée sur des méthodes désormais classiques en analyse des données : ACP, AFC, et classifications automatiques (nuées dynamiques etc., cf. [LEB 02], l'AHQ [DEV 80]). Cette approche quant à elle permettrait, dans l'exemple antérieur, de répondre à la question : Existe-il une (des) structure(s) dominante(s) dans les parcours résidentiels de la population ?

Nous nous plaçons dans le second type d'approche. Nous abordons la classification de trajectoires comme étant un problème d'estimation de densité de probabilité, et nous proposons d'utiliser un mélange de densités. En utilisant ensuite des données issues d'une enquête effectuée à Cali (Colombie) en 1998, nous montrons comment cette méthode est applicable à des données biographiques. Nous nous intéressons, tout particulièrement, à l'analyse de l'effet de l'introduction de la durée de séjour dans les états, lorsque les densités composantes du mélange sont des modèles de Markov.

2 Mélange de densités semimarkoviennes

Dans l'estimation de densités on essaie de modéliser une densité de probabilité $p(x)$ à partir des données observées $X=(x_1, x_2, \ldots, x_N)$, que l'on suppose issues de cette densité. Un mélange de densités est une distribution de la forme :

$$p(x)=\sum_{k=1}^{K} p(x/k)P(k), \quad \text{avec} \quad \sum_{k=1}^{K} P(k)=1 \quad \text{et} \quad 1 \le k \le K \qquad (1)$$

où K est le nombre de composantes du mélange, les $P(k)$ sont les paramètres du mélange (la probabilité a priori pour que la donnée x ait été générée par la composante k du mélange), et les $p(x/k)$ sont les densités composantes. Dans notre cas ces densités sont définies sur des séquences.

Soit $x_i=\{e_{i,1}, e_{i,2}, \ldots e_{i,T_i}\}$ la trajectoire de l'individu i ; où les e sont des états discrets de l'espace d'états E ($1 \le e \le m$), $e_{i,t}$, est l'état à l'instant t de la trajectoire de l'individu i et T_i, est la longueur de la trajectoire de l'individu *i.* Pour effectuer la classification par mélange de densités on peut réécrire (1) ainsi :

$$p(x_i/\Theta)=\sum_{k=1}^{K} p(x_i/\Theta_k)P(k) \qquad (2)$$

où K est le nombre de classes et Θ représente les paramètres du modèle $\{P(1), \ldots, P(K); \Theta_1, \ldots, \Theta_K\}$. Voir [EST 04] pour le cadre permettant d'effectuer la classification par mélange de densités, où l'on apprend les paramètres par l'algorithme EM [DEM 77], lorsque les densités sont des modèles de Markov.

Dans les données biographiques la durée passée dans les états est très importante. Cependant, les modèles de Markov classiques ne permettent pas de bien modéliser la durée passée dans un état donné. Dans un modèle de Markov la densité de durée (la probabilité de rester une durée d dans l'état e), notée $p(d/e)$, suit une distribution exponentielle qui ne dépend que de a_{ee} (la probabilité de boucler dans l'état e) :

$$p(d/e)=(a_{ee})^{d-1}(1-a_{ee}) \qquad (3)$$

Pour que le modèle rende compte de certains traits signifiants des trajectoires, par exemple pour les trajectoires résidentielles : les durées de séjour dans certains espaces géographiques (une région, une ville, …), il est préférable d'expliciter d'une façon analytique la densité de durée $p(d/e)$ dans le modèle. [FER 80] a spécifié pour chaque état du modèle, une densité de durée non paramétrique. Nous allons spécifier une densité de durée paramétrique dans un modèle de Markov. Pour faire intervenir explicitement les durées associées aux états on réécrit la trajectoire de l'individu i, x_i, ainsi :

$$x_i=\{(e_{i,1}, d_{i,1}), (e_{i,2}, d_{i,2}), \ldots (e_{i,NE_i}, d_{i,NE_i})\}, \qquad (4)$$

où : $e_{i,j}$ est le $j^{ème}$ état de la trajectoire et $e_{i,j} \neq e_{i,j-1}$, $d_{i,j}$ est la durée passée par l'individu i dans le $j^{ème}$ état de sa trajectoire et NE_i est le nombre d'états de la trajectoire. Donc, la vraisemblance d'une trajectoire conditionnée par son appartenance à une classe particulier Θ_k est donnée par :

$$p(x_i/\Theta_k)=\pi_k(e_{i,1})p_k(d_{i,1}/e_{i,1})\prod_{j=2}^{NE_i} a_k(e_{i,j}/e_{i,j-1})p_k(d_{i,j}/e_{i,j}) \qquad (5)$$

où $\pi_k(e_{i,1})$ est le vecteur de probabilité d'état initial et $a_k(e_t/e_{t-1})$ la matrice m x m de probabilités de transition.

Si la durée dans chaque état suit une loi Normale : $N(\mu, \sigma^2, d)=\frac{1}{(2\pi\sigma^2)^{1/2}}\exp\left\{-\frac{(d-\mu)^2}{2\sigma^2}\right\}$, alors,

$$p(d/e)=\frac{N\left(\mu(e),\sigma^2(e),d\right)}{\sum_{d'=d\min}^{d\max} N\left(\mu(e),\sigma^2(e),d'\right)} \text{ (Loi normale discrétisée avec } d\min \le d \le d\max \text{)} \quad (6)$$

Pour effectuer la classification, on apprend les paramètres du mélange de densités par l'algorithme EM :
Etape *E*: on calcule les probabilités a posteriori $p(i \in k / x_i, \Theta)$

$$p(i \in k / x_i, \Theta)=\frac{p(x_i / \Theta_k)P(k)}{\sum_{u=1}^{K} p(x_i / \Theta_u)P(u)} \quad (7)$$

Etape *M*: on actualise les paramètres courants Θ, en pondérant chaque individu par $p(i \in k / x_i, \Theta)$

$$P(k)^{Nouveau}=\frac{1}{N}\sum_{i=1}^{N} p(i \in k / x_i, \Theta) \qquad \pi_k^{Nouveau}(s_p)=\frac{\sum_{i=1}^{N} p(i \in k / x_i, \Theta)\delta(s_p, e_{i,1})}{\sum_{i=1}^{N} p(i \in k / x_i, \Theta)}$$

$$a_k^{Nouveau}(s_q / s_p)=\frac{\sum_{i=1}^{N} p(i \in k / x_i, \Theta) r_i^{s_p \to s_q}}{\sum_{i=1}^{N} p(i \in k / x_i, \Theta) r_i^{s_p \to}} \qquad \text{où} \quad p \neq q, \quad (8)$$

$$\mu_k^{Nouveau}(s)=\frac{\sum_{i=1}^{N} p(i \in k / x_i, \Theta)\sum_{j=1}^{NS_i^s} d_{i,j}^s}{\sum_{i=1}^{N} p(i \in k / x_i, \Theta) NS_i^s} \quad , \quad \sigma_k^{2\,Nouveau}(s)=\frac{\sum_{i=1}^{N} p(i \in k / x_i, \Theta)\left[\sum_{j=1}^{NS_i^s}\left[\left(d_{i,j}^s\right)-\left(\mu_k(s)\right)\right]^2\right]}{\sum_{i=1}^{N} p(i \in k / x_i, \Theta) NS_i^s}$$

$r_i^{s_p \to s_q}$ étant le compte des transitions depuis l'état s_p à l'état s_q dans la trajectoire de l'individu *i*, $r_i^{s_p \to}$ le compte des transitions depuis l'état s_p à n'importe quel état dans la trajectoire de l'individu *i*, $d_{i,j}^s$ la durée du $j^{ème}$ séjour de l'individu *i* dans l'état *s* et NS_i^s est le nombre de séjours de l'individu *i* dans l'état *s*. L'apprentissage du nombre de classes est un problème ouvert. Il existe cependant des méthodes qui essaient d'en donner des réponses : pénalisation de la vraisemblance (AIC, BIC), coude de la vraisemblance, etc.

3 Expériences et résultats

La méthode a été appliquée à 1749 trajectoires socio-résidentielles géographiques (changement de résidence à l'intérieur de Cali), ayant 5 changements en moyenne. Quatre modèles différents ont été utilisés comme densités du mélange : un modèle de Markov et trois modèles semi-markoviens avec des lois de durée Normale (N), Poisson (P) et Log-Normale (LN). Les critères BIC et AIC ont été testés pour déterminer le nombre de classes mais ils se sont montrés inappropriées pour ce type de densités composantes. Ce nombre a donc été déterminé par la méthode du coude de la vraisemblance. Nous avons établit une mesure d'Homogénéité Intra-classe, notée HI, permettant d'évaluer la cohésion des individus dans les classes. Egalement, nous mesurons l'instant de sortie de la censure (noté ISC) d'au moins le 50 % des individus de chaque classe. Cette mesure par classe permet d'analyser l'ensemble des classes selon la longueur des trajectoires. Le tableau 1 montre le résultat pour les mesures HI et ISC. La dernière ligne du tableau est le HI globale et l'écart type de l'ISC.

Globalement, une meilleure cohésion des individus aux classes est obtenue lorsqu'on introduit explicitement la durée de séjour. La définition d'une loi pour les durées de séjour des individus dans les états, fait qu'ils sont mieux attachés aux classes dont ils font partie. De plus, une meilleure différentiation des classes est observée à partir de la longueur moyenne des trajectoires : les ISC sont proches dans la

typologie sans durée explicite (M a un écart type très petit), et plus distants dans celles avec durée explicite (écart types sont plus élevés). La taille moyenne des classes est de 168 individus. L'interprétation des classes est facilitée avec l'introduction d'une loi de durée. Pour les 10 classes de la classification avec une loi de durée Normal (N) l'interprétation montre des parcours socio-résidentiels bien différentiées à Cali.

Mesure	HI				ISC			
Classe	M	N	P	LN	M	N	P	LN
1	0,86	0,95	0,98	0,96	36	18	46	36
2	1,00	0,95	1,00	0,91	47	47	34	47
3	0,75	0,99	0,94	0,94	40	35	45	49
4	0,85	1,00	1,00	0,98	39	51	10	25
5	0,65	0,99	1,00	0,91	36	21	31	28
6	0,82	0,97	0,99	0,94	45	45	19	42
7	0,98	1,00	0,94	0,92	36	47	36	51
8	0,71	1,00	0,99	1,00	49	55	53	51
9	0,95	0,99	0,96	1,00	43	32	31	55
10	0,84	0,99	1,00	0,93	37	41	20	40
HIG-ET	0,78	0,98	0,98	0,95	4,89	12,46	13,43	10,16

Tableau 1. HI - ISC

L'homogénéité Intra-Classe, qui reflète la facilité d'interprétation des classes, est calculée à partir des probabilités a posteriori, ainsi :

$$HI_k = \frac{\sum_{i \in k} \delta_i}{n_k}, \text{ où :}$$

$$\begin{cases} \delta_i = 0 & si \quad p(i \in k / x_i, \Theta) < 0{,}5 \\ \delta_i = 1 & si \quad p(i \in k / x_i, \Theta) \geq 0{,}5 \end{cases}$$

4 Conclusions

Nous avons abordé le problème de la prise en compte de la durée de séjour dans la classification de données biographiques. Nous avons présenté une méthode pour modéliser explicitement cette durée de séjour. Les densités composantes deviennent des modèles semi-markoviens. Nous avons montré, par un exemple, comment on peut apprendre les paramètres du mélange de densités semi-markoviennes avec l'algorithme EM. Nous avons réussi à relever l'importance des durées de séjour, ce qui est traduit par des classes plus stables (des individus mieux attachés aux classes) et mieux séparées selon la longueur de trajectoires. Il est apparu que les modèles semi-markoviens permettent d'obtenir des meilleurs résultats que les modèles markoviens (sans durée explicite). Cette méthode est applicable à n'importe quelle type de données séquentielles où la durée de séjour dans les états est importante.

5 Bibliographie

[COU 89] COURGEAU D. et LELIÈVRE E., (1989), *Analyse démographique des biographies*, INED, Paris, 268 p.

[COX 84] COX D. R. and OAKES D., (1984), *Analysis of survival data*, Chapman y Hall, Londres, 201 p.

[DEM 77] DEMPSTER A. P., LAIRD N. M., and RUBIN D. B. (1977). *Maximum likelihood from incomplete data via the EM algorithm*. Journal of the Royal Statistical Society, Series B, n° 34, pp.1-38.

[DEV 80] DEVILLE J.C., SAPORTA G., (1980), *Analyse harmonique qualitative*, in Data Analysis and Informatics, E. DIDAY et al. éditeurs, North Holland Publishing Compagny, pp. 375-389.

[EST 04] ESTACIO MORENO A., BARBARY O., GALLINARI P., PIRON M., (2004), *Classification de données biographiques : application à des trajectoires migratoires vers Cali (Colombie)*. In Revue de Statistique Appliquée, vol. LII (4). pp. 33-54.

[FER 80] FERGUSON J.D., (1980) *Variable duration models for speech*. In J.D. Ferguson, editor, Proc. Symposium on the Application of Hidden Markov Models to Text and Speech, Princeton, NJ. pp 143-179.

[LEB 02] LEBART L., MORINEAU A., PIRON M., (2002), *Statistique exploratoire multidimensionnelle*. Paris, 2002, éd. Dunod, 437 p

Extension de CART dans le cas bivarié : partition optimale du plan.

B. Fichet, J. Gaudart

LIF, Equipe Biomathématiques,
Universités Aix-Marseille, Faculté de Médecine,
27 Bd Jean Moulin 13385 Marseille cedex 5, France.

RÉSUMÉ. La méthode CART donne (de façon récursive) une partition de l'espace défini par les variables explicatives, par des hyperplans orthogonaux aux axes de cet espace. L'extension à la recherche d'une partition par un hyperplan oblique a été montrée NP-difficile, et la plupart des algorithmes proposés sont stochastiques et ne conduisent qu'à des solutions approximatives. Nous proposons ici un algorithme efficace pour obtenir une partition oblique optimale dans le cas particulier du plan.

MOTS-CLÉS : Classification et arbres de régressions, arbres de décision oblique, partition oblique optimale du plan, variables géographiques.

1 Introduction

Depuis les premiers textes de Breiman dans les années quatre-vingts, une abondante littérature a rendu compte des extensions de la méthode CART initiale (Classification And Regression Tree) [BRE 93]. Classiquement, l'algorithme CART recherche parmi les variables explicatives (numériques dans le cas qui nous intéresse) une variable et une bi-partition de celle-ci (en deux parties connexes) qui maximise la variance inter-groupes de la variable numérique à expliquer. Appliquée récursivement, cette procédure conduit à un arbre hiérarchique binaire appelé arbre de décision. La partition de l'espace défini par les variables explicatives est ainsi faite par des hyperplans orthogonaux aux axes de cet espace. L'extension à la recherche d'une partition par un hyperplan oblique (conduisant à un "arbre de décision oblique") a été décrite comme NP-difficile dans $\mathbb{R}^d$ par Heath *et al.* [HEA 93]. Dans la littérature les algorithmes proposés font appel à des procédures stochastiques (voir par exemple [MUR 94]) ou heuristiques [BRE 93].
Nous nous sommes plus particulièrement intéressé à la partition oblique du plan. Malgré l'abondante littérature, nous n'avons pas trouvé d'algorithme performant donnant une solution optimale dans $\mathbb{R}^2$.
Le présent travail a été motivé par la recherche de zones à risque de paludisme dans un village africain. La variable à expliquer Z est le pourcentage d'enfants infectés par groupe de maisons (concession), et les variables explicatives sont les coordonnées GPS de chaque concession (X =longitude, Y =latitude). D'autres méthodes, plus classiquement utilisées pour ce type d'application, sont largement décrites ailleurs [WAL 04].
Confronté ainsi à un système de coordonnées géographiques prises comme variables, rien a priori ne privilégie des partitions faites orthogonalement à ces variables (longitude et latitude). La recherche de partitions obliques découle de ces considérations. Nous proposons ici un algorithme efficace, simple et de complexité minimale permettant une partition optimale dans le cas particulier du plan.

2 Classification et arbre de régression

Soient dans le plan, muni d'un système $\{x, y\}$, choisi orthonormé sans perte de généralité, et d'une origine O fixés a priori, n points M_i de cordonnées (x_i, y_i). A tout point M_i est associé une variable aléatoire numérique Z_i, dite à expliquer, donnant l'observation z_i.
Dans l'esprit de CART nous cherchons une partition du plan selon une droite $\mathcal{D}$ oblique, telle que la partition induite sur les z_i maximise la variance intergroupe.

Pour une direction de $\mathcal{D}$ fixée, et donc une direction perpendiculaire $\boldsymbol{u}$ faisant un angle $\theta \in [0, \pi[$ avec l'axe $\boldsymbol{x}$, il convient de :

- projeter les points $\boldsymbol{M_i}$ sur la direction $(\boldsymbol{O}, \boldsymbol{u})$,
- sur les points ainsi projetés, calculer les variances intergroupes pour chaque partition en deux classes connexes dans cette direction. On détectera ainsi la partition qui maximise la variance intergroupe dans cette direction. Seul l'ordre induit par les points projetés importe ici, comme dans la méthode CART usuelle.

Pour une solution globale, il convient de balayer toutes les directions obliques d'angle $\theta \in [0, \pi[$. On peut, de façon heuristique, discrétiser l'ensemble $[0, \pi[$ en un nombre fini d'angles. Mais cette heuristique ne donne pas une solution optimale.

Il est évident que deux points $\boldsymbol{M_i}$ et $\boldsymbol{M_j}$ auront des coordonnées confondues en projection sur la direction $(\boldsymbol{O}, \boldsymbol{u})$ si et seulement si $\boldsymbol{M_i M_j}$ est orthogonale à $(\boldsymbol{O}, \boldsymbol{u})$ (figure 1a). Il existe donc un nombre fini de directions, dites critiques, définis par des angles θ_{ij}.

En posant $\varphi_{ij} = \boldsymbol{Arctg(a_{ij})}$, avec $\boldsymbol{a}_{ij} = \dfrac{\boldsymbol{y_j - y_i}}{\boldsymbol{x_j - x_i}}$ et $\varphi_{ij} \in \left[-\dfrac{\pi}{2} ; \dfrac{\pi}{2} \right[$, on a $\theta_{ij} = \dfrac{\pi}{2} + \varphi_{ij}$.

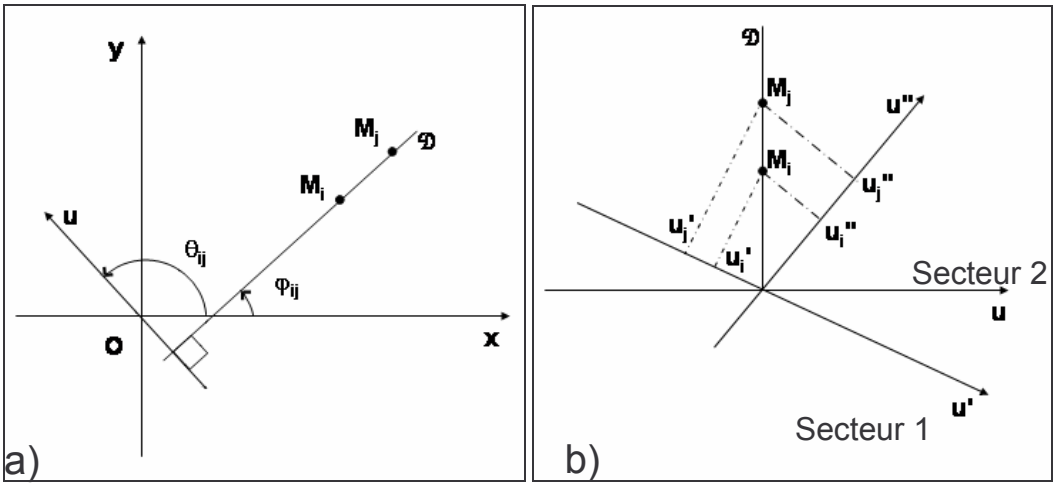


Figure 1: a) construction de l'angle critique θ_{ij} ; b) modification de l'ordre des coordonnées projetées sur les directions u' et u", lors du passage entre 2 secteurs. On pose u : direction d'angle critique θ_{ij} ;

u' et u": directions d'angles intermédiaires, respectivement du secteur 1 et du secteur 2.

u_i' u_j' ui" et u_j" sont les coordonnées des points M_i et M_j projetés sur les directions u' et u".

Ces angles critiques θ_{ij} définissent autant de secteurs angulaires. A l'intérieur de chaque secteur, l'ordre des coordonnées en projection sur la direction $(\boldsymbol{O}, \boldsymbol{u})$ ne dépend pas de cette direction. En effet, la différence de coordonnées $(\boldsymbol{u_i - u_j})$ en projection de $\boldsymbol{M_i}$ et $\boldsymbol{M_j}$ sur l'axe $(\boldsymbol{O}, \boldsymbol{u})$ vérifie :

$$(\boldsymbol{u_i - u_j})\boldsymbol{cos}(\varphi_{ij}) = (\boldsymbol{x_i - x_j})\boldsymbol{sin}(\theta - \theta_{ij}) \qquad [1]$$

avec $(\boldsymbol{u_i - u_j}) = (\boldsymbol{y_i - y_j})\boldsymbol{sin}(\theta)$ pour $\boldsymbol{x_i} = \boldsymbol{x_j}$ ce qui est équivalent à $\varphi_{ij} = -\dfrac{\pi}{2}$.

Ainsi la différence $(u_i - u_j)$ dépend continûment de l'angle θ . Elle ne peut pas changer de signe à l'intérieur d'un secteur angulaire puisqu'elle ne s'annule que pour $\theta = \theta_{ij}$. Il en ressort que les partitions (et donc les variances inter-groupes) ne sont pas modifiées à l'intérieur de chaque secteur. En outre le passage d'un secteur à un autre, via l'angle critique θ_{ij} (figure 1b), induit le même ordre à l'exception de la permutation de deux éléments (u_i, u_j) (voire la permutation de plusieurs couples (u_i, u_j) et (u_k, u_l) si $M_k M_l$ est parallèle à $M_i M_j$, ou encore la permutation complète d'un intervalle de plusieurs éléments $(u_i, u_j, u_k \ldots)$ si $M_i, M_j, M_k \ldots$ sont alignés). C'est en effet une conséquence immédiate de [1].

En ordonnant les angles critiques θ_{ij} distincts, $0 \leq \theta^{(1)} < \theta^{(2)} < \ldots < \theta^{(N)} < \pi$ avec $N \leq n \times (n-1)/2$, on pourrait choisir un angle dans chaque secteur angulaire.

Ces angles définissent autant de variables numériques ramenant ainsi l'algorithme à la procédure CART usuelle. Mais le nombre n de données est souvent grand : dans l'application ayant motivé ce travail, $n = 164$ concessions ont été étudiées. Le nombre d'angles distincts est donc grand (dans l'application numérique on trouve $N = 13270$ angles distincts), et cette procédure serait trop coûteuse en temps et en espace.

Un algorithme plus efficace et moins coûteux consiste à balayer les secteurs angulaires en tenant compte à chaque pas des calculs antérieurs.

En effet, le calcul de la variance intergroupe entre deux secteurs consécutifs ne s'effectue que sur une seule partition pour le couple permuté (voire quelques partitions pour un groupe de couples permutés). On hérite ainsi pour le calcul de la variance intergroupe des partitions précédentes, sauf celle correspondant à la nouvelle permutation, pour peu que l'on ait gardé les valeurs des variances intergroupes calculées pour le secteur angulaire précédant.

On peut voir que l'algorithme (présenté ici pour $\theta^{(1)} \neq 0$) est en $O(n^2 \log n)$ en temps et $O(n)$ en espace pour une partition:

- Ordonner les x_i.
- Calculer et ordonner les θ_{ij}, via les a_{ij}.
- Calculer la somme totale des z_i.
- Pour chaque bi-partition sur l'axe x :
  - Calculer la somme des z_i des deux groupes, afin de calculer les variances intergroupes, en utilisant à chaque nouveau calcul le résultat précédant. On a ainsi les résultats pour le premier secteur.
  - Stocker l'ordre des x_i et les sommes des z_i par groupe.
  - Initialiser un optimum et une partition optimale.
- Pour les secteurs angulaires suivants :
  - Calculer seulement la variance intergroupe correspondant à la partition induite par le (les) couple(s) permuté(s).
  - Stocker le nouvel ordre calculé et les nouvelles sommes des z_i par groupe.
  - Réinitialiser, si nécessaire, l'optimum et la partition optimale.
- La procédure se termine lorsque tous les secteurs angulaires ont été balayés.

Si l'on souhaite obtenir la variance totale il faut calculer la somme totale des z_i^2, et les sommes des z_i^2 par groupe pour avoir les variances locales de chaque partition.
L'algorithme continue dans chaque partition optimale ainsi constituée jusqu'à l'obtention de l'arbre complet de régression, comme dans la procédure CART usuelle.

3 Simulation

Nous avons utilisé la méthode CART usuelle sur 20 points simulés (X et Y suivant des lois uniformes, Z suivant un mélange de lois de Poisson, reproduisant le cadre de l'étude épidémiologique initiale), ainsi que l'extension proposée implémentée sous Matlab 7.0.1.
La méthode CART usuelle (avec comme règles d'arrêt : effectif minimum du nœud parent = 5; effectif minimum du nœud enfant = 2) permet la partition du plan en 5 zones différentes (figure 2a), avec un pourcentage de variance expliquée par la partition finale de l'arbre de 84,7%. L'algorithme proposé nous permet d'obtenir des partitions obliques optimales du plan en 5 zones différentes (figure 2b), avec un pourcentage de variance expliquée amélioré (88,1%).

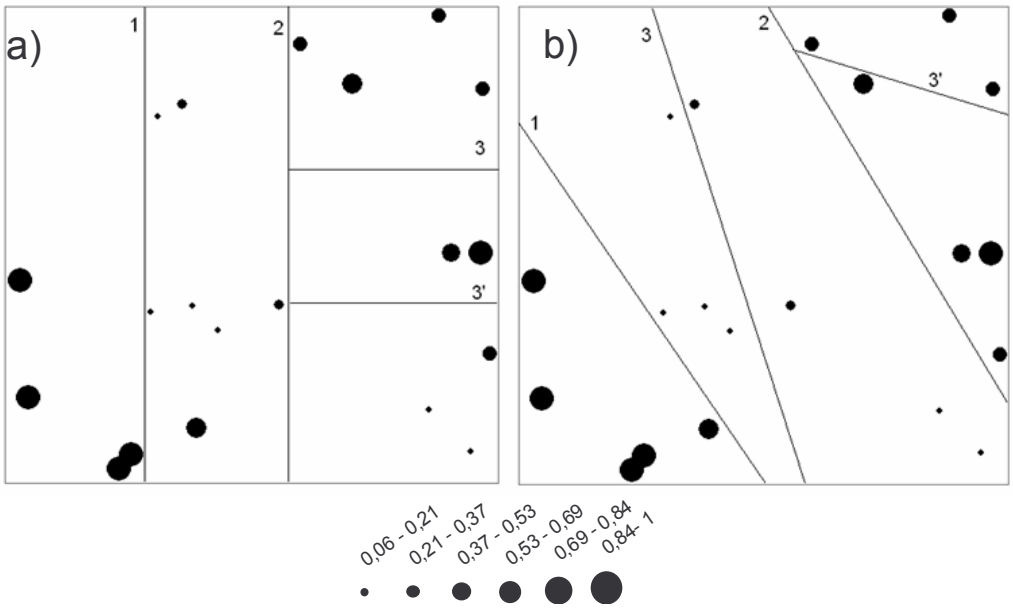


Figure 2: Partitions obtenues à l'aide de CART (a) et de l'arbre de décision oblique (b) :

(n =20; effectif minimum du nœud parent =5; effectif minimum du nœud enfant =2).

L'échelle de taille représente les valeurs simulées de la variable à expliquer discrétisées en 6 classes égales.

Les numéros représentent les nœuds de l'arbre de régression.

4 Bibliographie

[BRE 93] BREIMAN L., FRIEDMAN J., OLSHEN R., STONE C., *Classification and regression trees*, Chapman & Hall, 1993.

[HEA 93] HEATH D., KASIF M., SALZBERG S., "Induction of oblique decision trees", *Proceedings of the 13th International Joint Conference on Artificial Intelligence*, Ed. Bajcsy R., Chambery, France, 1993, p. 1002-1007.

[MUR 94] MURTHY S., KASIF M., SALZBERG S., "A system for induction of oblique decision trees", *Journal of Artificial Intelligence Research*, Vol. 2, 1994, p. 1-32.

[WAL 04] WALLER L., GOTWAY C., *Applied spatial statistics for public health data*, Wiley, 2004.

Une classification des graphes sans $K_{3,3}$ plongeables sur le plan projectif ou sur le tore

Andrei Gagarin, Gilbert Labelle et Pierre Leroux

LaCIM
Université du Québec à Montréal
Case postale 8888, succursale Centre-Ville
Montréal (Québec) Canada, H3C 3P8
gagarin@lacim.uqam.ca, labelle.gilbert@uqam.ca, leroux.pierre@uqam.ca

RÉSUMÉ. *Nous présentons des théorèmes de structure classifiant les graphes 2-connexes sans subdivision de $K_{3,3}$ qui peuvent être plongés sur le plan projectif ou sur le tore. Ces résultats sont formulés en termes d'une opération naturelle de substitution de réseaux bipolaires planaires dans les arêtes de certains graphes appelés noyaux projectifs-planaires ou toroïdaux. On en déduit un algorithme linéaire de reconnaissance des graphes toroïdaux sans $K_{3,3}$ ainsi que des formules de dénombrement dans le cas étiqueté.*

MOTS-CLÉS : *graphe toroïdal, graphe projectif-planaire, substitution dans les arêtes, dénombrement des graphes, séries génératrices.*

1. Introduction : l'opération $\mathcal{G} \uparrow \mathcal{N}$

Un *réseau bipolaire* est un graphe connexe N contenant deux sommets distingués a et b tel que le graphe $N \cup ab$ soit 2-connexe. Les sommets a et b sont appelés les *pôles* du réseau. Les sommets qui ne sont pas des pôles sont dits *internes*. Un réseau N est dit *fortement planaire* si le graphe $N \cup ab$ est planaire. Notons $\mathcal{N}_P$ la classe des réseaux bipolaires fortement planaires.

Le résultat de la *substitution* d'un réseau bipolaire N dans une arête e d'un graphe G est l'ensemble d'un ou deux graphes obtenus en remplaçant l'arête e par le réseau N, identifiant les pôles de N aux extrémités de e. On suppose que l'ensemble sous-jacent de sommets internes de N est disjoint des sommets de G. Étant donné un graphe G_0 ayant k arêtes, $E = \{e_1, e_2, \ldots, e_k\}$, et une liste $(N_1, N_2, \ldots, N_k)$ de réseaux bipolaires disjoints, on définit la *composition* $G_0 \uparrow (N_1, N_2, \ldots, N_k)$ comme l'ensemble des graphes qui peuvent être obtenus en substituant, pour $j = 1, 2, \ldots, k$, le réseau N_j dans l'arête e_j de G_0. Le graphe G_0 est appelé le *noyau*, et les N_i's sont appelés les *composantes* des graphes obtenus. Étant donné une classe de graphes $\mathcal{G}$ et une classe de réseaux $\mathcal{N}$, la classe de graphes obtenus comme des compositions $G_0 \uparrow (N_1, N_2, \ldots, N_k)$, où $G_0 \in \mathcal{G}$ et $N_i \in \mathcal{N}$, $i = 1, 2, \ldots, k$, est notée $\mathcal{G} \uparrow \mathcal{N}$. On dit que la composition $\mathcal{G} \uparrow \mathcal{N}$ est *canonique* si pour chaque graphe $G \in \mathcal{G} \uparrow \mathcal{N}$ il existe un noyau unique $G_0 \in \mathcal{G}$ et des composantes uniques (à l'orientation près) $N_1, N_2, \ldots, N_k \in \mathcal{N}$ tels que $G \in G_0 \uparrow (N_1, N_2, \ldots, N_k)$.

2. Théorèmes de structure

Le théorème de Kuratowski dit qu'un graphe G est non-planaire si et seulement s'il contient une subdivision de K_5 ou de $K_{3,3}$. Nous nous intéressons ici à la classification des graphes 2-connexes non-

planaires qui ne contiennent pas de subdivision de $K_{3,3}$ et qui sont plongeables sur le plan projectif (dits *projectifs-planaires*) ou sur le tore (dits *toroïdaux*). Nos résultats sont formulés en termes de compositions de la forme $\mathcal{G} \uparrow \mathcal{N}_P$, où $\mathcal{G}$ est une classe de graphes spéciaux appelés *noyaux projectifs-planaires* ou *toroïdaux*. Voici un premier exemple, pour les graphes projectifs-planaires, basé sur les résultats structurels et algorithmiques de [GAG 02].

Théorème 1 ([GAG 04a]) *Un graphe 2-connexe non-planaire G sans subdivision de $K_{3,3}$ est projectif-planaire si et seulement si $G \in K_5 \uparrow \mathcal{N}_P$. De plus, la composition $K_5 \uparrow \mathcal{N}_P$ est canonique.*

Ainsi le seul noyau pour les graphes projectifs-planaires sans $K_{3,3}$ est le graphe K_5 lui-même. Passons aux noyaux pour les graphes toroïdaux sans $K_{3,3}$.

Les graphes M est M^* sont définis par les figures 1(i) et (ii) respectivement. Le réseau bipolaire noté $K_5\backslash e$ est construit à partir de K_5 en sélectionnant deux sommets a et b (les pôles) et en enlevant l'arête $e = ab$ (voir la figure 2(i)). Une *couronne* de réseaux $K_5\backslash e$, H, est obtenue en substituant des réseaux $K_5\backslash e$ dans des arêtes d'un cycle C_j de longueur $j \geq 3$, de sorte que les arêtes restantes de C_j ne soient pas adjacentes (voir un exemple avec $j = 5$ à la figure 2(ii)). Par définition, l'ensemble des *noyaux toroïdaux* consiste en les graphes K_5, M, M^*, et toutes les couronnes de réseaux $K_5\backslash e$. Notons $\mathcal{T}_C$ la classe des noyaux toroïdaux. Dans [GAG 04b] nous raffinons des résultats de [GAG 02] pour obtenir le

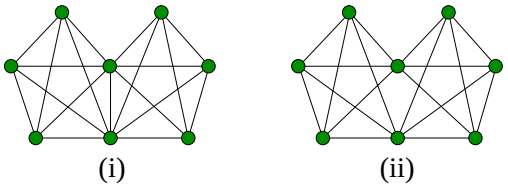


FIG. 1. (i) Graphe M, (ii) Graphe M^*.

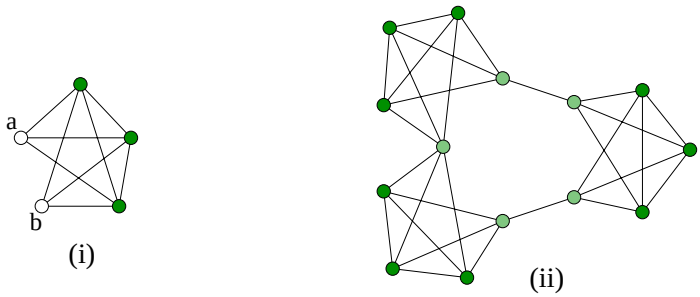


FIG. 2. (i) Réseau $K_5\backslash e$, (ii) Une couronne obtenue à partir de C_5.

théorème suivant.

Théorème 2 ([GAG 04b]) *Un graphe 2-connexe non-planaire G sans subdivision de $K_{3,3}$ est toroïdal si et seulement si $G \in \mathcal{T}_C \uparrow \mathcal{N}_P$. En plus, la composition $\mathcal{T} = \mathcal{T}_C \uparrow \mathcal{N}_P$ est canonique.*

Les théorèmes 1 et 2 impliquent qu'un graphe projectif-planaire sans subdivision de $K_{3,3}$ est aussi toroïdal. Cependant, un graphe projectif-planaire contenant une subdivision de $K_{3,3}$ peut être non-toroïdal. Les caractérisations des théorèmes 1 et 2 peuvent être utilisées pour reconnaître les graphes projectifs-planaires ou toroïdaux sans subdivision de $K_{3,3}$ en temps linéaire. Un tel algorithme peut être déduit des algorithmes correspondant de [GAG 02].

3. Formules de dénombrement

Une *espèce* de graphes est une classe de graphes étiquetés qui est fermée par rapport aux réétiquetages de leurs sommets. Pour une espèce $\mathcal{G}$ de graphes, nous introduisons sa fonction génératrice (exponentielle) $\mathcal{G}(x, y)$ comme

$$\mathcal{G}(x,y) = \sum_{n\geq 0} g_n(y)\frac{x^n}{n!}, \text{ où } g_n(y) = \sum_{m\geq 0} g_{n,m}y^m, \tag{1}$$

et où $g_{n,m}$ est le nombre de graphes dans $\mathcal{G}$ sur un ensemble donné de n sommets et ayant m arêtes. La variable formelle y est un compteur d'arêtes. Voir [BER 98] pour plus de détails sur la théorie combinatoire des espèces de structures.

Seulement les sommets internes d'un réseau bipolaire sont pris en compte pour le dénombrement. Ainsi, la fonction génératrice d'une espèce $\mathcal{N}$ de réseaux bipolaires est définie par

$$\mathcal{N}(x,y) = \sum_{n\geq 0} \nu_n(y)\frac{x^n}{n!}, \text{ où } \nu_n(y) = \sum_{m\geq 0} \nu_{n,m}y^m, \tag{2}$$

et où $\nu_{n,m}$ est le nombre de réseaux dans $\mathcal{N}$ construits sur un ensemble de sommets internes de taille n et ayant m arêtes.

Une espèce $\mathcal{N}$ de réseaux bipolaires est appelée *symétrique* si pour chaque $N \in \mathcal{N}$, son *opposé* $\tau \cdot N$ qui s'obtient en interchangeant les pôles a et b est aussi dans $\mathcal{N}$.

Lemme 1 (T. Walsh [WAL 82]) *Soient $\mathcal{G}$ une espèce de graphes et $\mathcal{N}$ une espèce symétrique de réseaux bipolaires telles que la composition $\mathcal{G} \uparrow \mathcal{N}$ soit canonique. Alors on a l'égalité suivante pour les fonctions génératrices :*

$$(\mathcal{G} \uparrow \mathcal{N})(x,y) = \mathcal{G}(x, \mathcal{N}(x,y)). \tag{3}$$

Corollaire 1 *Sachant que $K_5(x,y) = x^5y^{10}/5!$, la fonction génératrice $\mathrm{PP}(x,y)$ des graphes projectifs-planaires non-planaires étiquetés 2-connexes sans subdivision de $K_{3,3}$ est donnée par*

$$\mathrm{PP}(x,y) = K_5(x, \mathcal{N}_P(x,y)) = \frac{x^5\mathcal{N}_P^{10}(x,y)}{5!}. \tag{4}$$

Corollaire 2 *La fonction génératrice $\mathcal{T}(x,y)$ des graphes toroïdaux non-planaires étiquetés 2-connexes sans subdivision de $K_{3,3}$ est donnée par*

$$\mathcal{T}(x,y) = (\mathcal{T}_C \uparrow \mathcal{N}_P)(x,y) = \mathcal{T}_C(x, \mathcal{N}_P(x,y)). \tag{5}$$

Soit $\mathcal{P}$ l'espèce de graphes planaires 2-connexes. Des méthodes pour calculer la fonction génératrice $\mathcal{P}(x,y)$ sont décrites dans [BEN 02] et [BOD 03]. La fonction génératrice $\mathcal{N}_P$ de la classe de réseaux bipolaires fortement planaires peut être obtenue comme suit (voir [GAG 04a] ou [WAL 82] pour plus de détails) :

$$\mathcal{N}_P(x,y) = (1+y)\frac{2}{x^2}\frac{\partial}{\partial y}\mathcal{P}(x,y) - 1. \tag{6}$$

Ensuite, il faut calculer la fonction génératrice $\mathcal{T}_C(x,y)$ des noyaux toroïdaux pour obtenir $\mathcal{T}(x,y)$. Nous avons $\mathcal{T}_C = K_5 + M + M^* + \mathrm{CC}$, où CC désigne la classe des couronnes de réseaux $K_5 \backslash e$. Remarquons que

$$M(x,y) = 280\frac{x^8y^{19}}{8!}, \quad M^*(x,y) = 280\frac{x^8y^{18}}{8!}. \tag{7}$$

L'espèce CC des couronnes de réseaux $K_5\backslash e$ peut être dénombrée comme suit. On remarque que, pour une couronne de réseaux $K_5\backslash e$, les arêtes "restantes" (dans lesquelles aucune substitution n'a eu lieu) du cycle C_j forment un couplage de ce cycle. Il s'ensuit que

$$\begin{aligned} \mathrm{CC}(x,y) &= \sum_{j\geq 3}\frac{(j-1)!}{2}\sum_{k=0}^{j/2} t_{j,k}\frac{y^k(x^3y^9/3!)^{j-k}x^j}{j!} \\ &= -\frac{1}{144}(12x^4y^9 + 12x^5y^{10} + x^8y^{18} + 72\ln(1 - \frac{x^4y^9}{6} - \frac{x^5y^{10}}{6})), \end{aligned} \tag{8}$$

où $t_{j,k}$ est le nombre de couplages de taille k dans un cycle de longueur j (voir [GAG 04b] pour plus de détails). En appliquant (5), on obtient la série $\mathcal{T}(x,y)$. La table 1 donne le nombre total τ_n de graphes dans $\mathcal{T}$, pour $n \leq 17$, obtenu en posant $y = 1$ dans $\mathcal{T}(x,y)$.

n	τ_n
5	1
6	120
7	10920
8	989520
9	99897840
10	11940037200
11	1737017325120
12	307410206405280
13	64915089945797520
14	15941442348672800960
15	4446392119411980978240
16	1382470831306742435905920
17	472436578501629382684767360

TAB. 1. Le nombre τ_n de graphes étiquetés toroïdaux non-planaires 2-connexes sans $K_{3,3}$, ayant $n \leq 17$ sommets.

4. Bibliographie

[BEN 02] E. A. Bender, Zh. Gao, and N. C. Wormald, "The number of labeled 2-connected planar graphs", Electron. J. Combin. **9** (2002), Research Paper 43, 13 pp. (electronic).

[BER 98] F. Bergeron, G. Labelle, and P. Leroux, *Combinatorial Species and Tree-like Structures*, Cambrige Univ. Press, 1998.

[BOD 03] M. Bodirsky, C. Gröpl, and M. Kang, "Generating labelled planar graphs uniformly at random", in *Proceedings of the* 30*-th International Colloquium on Automata, Languages and Programming*, LNCS 2719, Springer, 2003, 1095–1107.

[GAG 02] A. Gagarin and W. Kocay, "Embedding graphs containing K_5-subdivisions", *Ars Combin.* **64** (2002), 33–49.

[GAG 04a] A. Gagarin, G. Labelle, and P. Leroux, "Counting labelled projective-planar graphs without a $K_{3,3}$-subdivision", arXiv :math. CO/0406140, (2004), soumis.

[GAG 04b] A. Gagarin, G. Labelle, and P. Leroux, "Characterization and enumeration of toroidal $K_{3,3}$-subdivision-free graphs", arXiv :math. CO/0411356, (2004), soumis.

[WAL 82] T. R. S. Walsh, "Counting labelled three-connected and homeomorphically irreducible two-connected graphs", *J. Comb. Theory Ser. B* **32** (1982), 1–11.

Extraction de Règles en Incertain par la Méthode Statistique Implicative

Régis Gras *, Raphaël Couturier, Fabrice Guillet*, Filippo Spagnolo*****

** LINA– Ecole Polytechnique de l'Université de Nantes*
La Chantrerie BP 60601 44306 Nantes cedex
*** Institut Universitaire de Technologie de Belfort, BP 527, rue E. Gros, 90016 Belfort cedex :*
**** G.R.I.M. (Gruppo di Ricerca sull'Insegnamento delle Matematiche), Department of Mathematics, University of Palermo:*

RÉSUMÉ. En relation avec des approches classiques de l'incertain de Zadeh et autres auteurs, l'analyse statistique implicative (A.S.I.) peut apparaître innovante, particulièrement pour l'opérateur d'implication. L'article montre en effet que la notion de variables à valeurs intervalles et celle de variables-intervalles sont efficaces dans la détermination de la distribution des variables et dans la recherche de règles entre variables floues. De plus, elles apportent de riches informations sur la qualité de ces règles, tout en permettant d'étudier le rôle des variables supplémentaires dans l'existence de ces règles. Cette nouvelle perspective épistémologique de l'incertain ouvre d'intéressantes perspectives d'application.

MOTS-CLÉS . Règles, hiérarchies orientées, analyse statistique implicative, variable-intervalle

1 Introduction

Partant du cadre défini et formalisé par [ZAD 97], [LOF 01] et [DUB 87], ce texte vise à étudier les proximités formelle et sémantique des cadres de l'incertain et de l'analyse statistique implicative (A.S.I.) entre variables à valeurs intervalles et variables-intervalles [GRA 01b]. On ne rappellera pas les formalisations classiques des notions premières et de chaque opérateur de la logique floue. On s'intéressera plus particulièrement à l'opérateur « implication » à l'aide duquel on extrait des règles d'association.

Le texte présent s'inscrit dans le cadre des travaux initiés par Gras [GRA 79] sur une méthode d'analyse de données, l'analyse statistique implicative (A.S.I.) qui vise à extraire et représenter des règles d'association à partir de bases de données. Nous considérons celles qui croisent des sujets (ou des objets) et des variables, présentant des modalités nettes ou floues. Une règle entre deux variables ou entre conjonctions de variables est établie sur la base de la rareté statistique du nombre de ses contre-exemples, dans l'hypothèse de l'indépendance a priori des variables en jeu [GRA 79], [LER 81]. La qualité de la règle sera évidemment d'autant plus grande que ce nombre de contre-exemples sera statistiquement petit sous cette hypothèse, eu égard aux occurrences des variables et des instances totales.

Dans le §2, nous présentons la problématique. Puis dans le §3, nous construisons, de façon peu classique, une distribution floue à partir de données objectives vs subjectives. Dans le § 4.1, nous abordons la recherche de règles d'association dans une situation « floue » en nous appuyant auparavant sur la notion de variables modales. Enfin, dans le § 4.2, nous revenons sur la construction des règles en ramenant les variables floues à des variables-intervalles.

2 Problématique

Bien que les applications de la logique floue soient nombreuses en intelligence artificielle (en matière de diagnostic médical ou de reconnaissance des formes, etc.), plusieurs questions restent souvent latentes : comment obtient-on des distributions des degrés d'appartenance dans le cas de variables numériques ? Sur quelles connaissances sont-elles établies ? Sont-elles données a priori et mises à l'épreuve du réel ou sont-elles construites ? Dans ce dernier cas, quel processus d'extraction de connaissances y conduirait, quel type de règle alors extrairait-on ? Quelle signification donnerait-on à la règle associant 2 sous-ensembles ou attributs flous ?.C'est une des problématiques du data-mining.

3 Deux méthodes de construction de distributions floues par extraction de connaissances

Notre objectif est de « *fuzzifier* » les données en quantifiant le degré d'appartenance d'un sujet d'une population à un intervalle numérique donné. La méthode de type « clustering » que nous proposons consiste tout d'abord, à partir du choix d'un indice de similarité [LER 81], d'extraire la proximité entre des variables nettes (par exemple des attributs) et des variables floues. Les données à traiter seront donc de deux ordres :

- d'une part, des variables **objectives, consensuelles,** à valeurs numériques réparties sur des intervalles auxquels on associe respectivement un **attribut net** ,
- et d'autre part, un **attribut flou,** attribué **subjectivement** à chaque sujet et les mêmes intervalles..

Pour cela, selon le procédé défini dans [GRAS 01], nous choisissons de transformer l'ensemble des valeurs observées sur les sujets en sous-intervalles disjoints de variance inter-classe maximale afin de pouvoir attribuer à chaque sous-intervalle un attribut net de même désignation que celle attribuée aux attributs flous. Enfin, pour chaque classe de similarité entre attribut net et attribut flou, nous déterminons le **degré d'appartenance** des sujets à une classe floue à partir de la mesure normalisée de **typicalité** associée à chaque individu. Cette notion de typicalité, définie dans [GRA 01a], rend compte d'un degré de responsabilité dans la proximité nette $\leftrightarrow$ floue d'attributs, soulignant l'accord entre net et flou. Ainsi, nous disposerons d'une mesure vérifiant les axiomes de Zadeh relatifs au concept de « possibilité ». Mais son avantage par rapport à la détermination subjective classique est qu'elle est établie à l'épreuve statistique de la réalité et qu'elle varie avec la dilatation de l'ensemble des sujets.

Dans une seconde approche, on dispose de la distribution des valeurs floues prises par chaque sujet sur un intervalle. On cherche à en déduire une distribution des degrés d'appartenance sur cet intervalle. L'objectif final est de définir une variable symbolique, qui soit l'histogramme d'un intervalle sur lequel on pourra déterminer des sous-intervalles optimaux selon le critère de la variance. Soit f_1, f_2, …,f_n les distributions respectives des n sujets (équidistribués par exemple) sur un intervalle A. La fonction $f=(f_1+f_2+ \ldots.+f_n)/n$ intègre en un histogramme sur A la distribution des fonctions d'appartenance. Il suffit ensuite de discrétiser A en une suite de points pondérés selon f ; enfin, d'appliquer sur A la méthode des nuées dynamiques ([DID 72]) pour obtenir une variable-intervalle **a** dont on pourra étudier les relations implicatives avec les autres variables du même type.

4 Règles d'association pour des variables numériques

Les distributions des variables floues sont supposées connues selon 2 variables observées sur les mêmes sujets : par exemple, taille et poids. On veut étudier alors, comme en A.S.I., les règles de déduction entre le prédicat taille et le prédicat poids, présentant des modalités, l'un **Taille** = {petit, moyen, grand}, l'autre **Poids** = {léger, moyen, lourd}. On dispose de données sous forme d'un tableau numérique des degrés d'appartenance aux modalités d'attributs flous, valeurs relatives à un échantillon de 20 sujets. Les 3 premiers sujets constituent *TAB.1*. L'un d'entre eux, i_1, n'est donc pas très grand et pas très lourd, l'autre i_2 assez grand et assez lourd, le dernier i_3 plutôt grand et plutôt lourd.

	taille			**poids**		
	petit T_1	*moyen* T_2	*grand* T_3	*léger* P_1	*moyen* P_2	*lourd* P_3
i_1	8/15	5/15	2/15	7/14	4/14	3/14
i_2	1/14	6/14	7/14	2/15	5/15	8/15
i_3	0	7/16	9/16	1/16	6/16	9/16

TAB. 1 – Valeurs prises par les modalités sur les 3 sujets

4.1 Un premier traitement de variables numériques

On propose ici un traitement implicatif, selon l'A.S.I., en considérant les 6 variables tailles-poids comme des variables numériques. On obtient le graphe implicatif en utilisant l'indice de [LAG 98], réactualisé par [REG 04]. à partir des 20 sujets. Ainsi, les implications T3 ⇒ P3 et P1 ⇒ T1 sont valides au seuil 0.90 et signifient que les propositions grand ⇒ lourd et léger ⇒ petit, règle qui est sémantiquement contraposée de la première, sont acceptables. Une autre implication à un seuil >0.6 apparaît : P2 ⇒ T1.

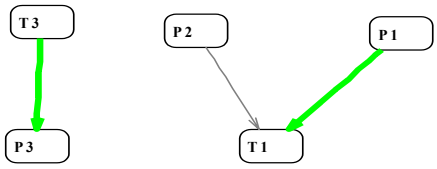


FIG. 1 – Graphe implicatif taille x poids

4.2 Second traitement par des variables à valeurs intervalles

Ce second traitement (cf. [GRA 01b]) va permettre de prendre en compte de façon plus fine les nuances des observations prises selon des sous-ensembles flous et de répartir leurs valeurs de façon optimale sur un intervalle numérique [0 ;1], selon une partition dont l'utilisateur définit le nombre de classes pour chacun de 20 sujets.

Nous disposons d'un nouveau tableau 2 donnant les distributions des 6 modalités des 2 attributs « taille » et « poids » relativement à chacun des individus et les valeurs binaires prises par 2 variables supplémentaire « Femme », « Homme ». En voici les 2 premières lignes

	Taille petite t	**Taille moyen. m**	**Taille grande T**	**Var supplé. Femme**	**Var. supplé. Homme**	**Poids léger L**	**Poids moy. o**	**Poids gran. P**
i1	0,7	0,4	0,3	1	0	0,8	0,3	0,1
i2	0,2	0,5	0,8	0	1	0,1	0,4	0,9

TAB. 2 – Distributions des attributs flous « taille » et « poids »

Par exemple., le sujet i_1 admet un degré d'appartenance 0.7 à la classe des sujets petits, 0.4 à celle de ceux de taille moyenne et 0.3 à la classe de ceux de grande taille. De plus, ce sujet est une femme et la distribution de ses degrés d'appartenance aux 3 classes de poids, sont respectivement 0.8, 0.3 et 0.1. Le traitement va emprunter cette fois la méthode des variables à valeurs intervalles. Comme dans le § 3, chaque modalité conduira à la construction de sous-intervalles optimaux (ici 3, soit **t1, t2** et **t3** pour **t**), c'est-à-dire la détermination de sous-intervalles optimisant, du moins localement sinon globalement, l'inertie inter-classe. Utilisant ensuite le logiciel de traitement CHIC de ce type de variable ([COU 01]), on établit les règles telles que : si un sujet relève de l'un des 3 intervalles $\mathbf{t_i}$ de la modalité t de l'attribut « taille » alors généralement il relève de l'intervalle $\mathbf{p_j}$ de la modalité **p** de l'attribut « poids ». Ainsi, si par exemple., il a tendance à être plutôt petit, alors il a généralement tendance à être plutôt léger. **Le graphe implicatif** au niveau 0.90 est également donné par CHIC :

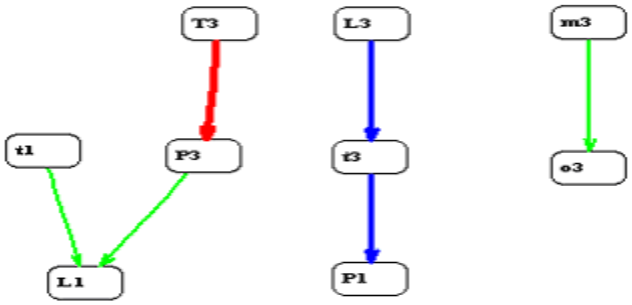


FIG. 2 – Graphe implicatif : taille x poids

On voit par exemple que :

- l'individu de grande taille (**T3**) admet généralement un poids important (**P3**) et donc n'est pas considéré comme léger (**L1**). Ce sont les hommes qui apportent, et de très loin (risque de se tromper = 0.07), la plus importante contribution ;

- l'individu de poids plutôt léger (**L3**) est généralement de petite taille (**t3**) ; dans ce cas, il n'est que très rarement considéré lourd (**P1**). Les femmes sont les plus contributives au chemin (risque = 0.25) ;

-les 2 variables **t1** et **L1**, liées par la règle **t1** $\Rightarrow$ **L1**, correspondent à des fréquences rares. Si donc, on rencontre un sujet petit alors il est généralement léger. Le sexe Homme contribue à cette règle.

5 Conclusion

A l'aide de l'A.S.I., nous avons cherché à objectiver la notion de degré d'appartenance Situant le modèle d'implication entre attributs par rapport à des modèles classiques, nous avons mis en évidence par un graphe, les relations implicatives entre des modalités de variables numériques. Nous avons, semble-t-il, amélioré la formalisation de la sémantique en faisant référence à des variables-intervalles. Les règles les plus consistantes ont pu être extraites selon leur qualité. Enfin, la relation entre des variables extrinsèques (supplémentaires) et ces règles a permis d'enrichir notre connaissance sur ces règles. Des applications à des situations réelles tenteront de valider cette nouvelle approche de l'incertain.

6 Bibliographie

[BER 04] BERNADET M., *Qualité des règles et des opérateurs en découverte de connaissances floues. Mesure de qualité pour la fouille de données*, Cepaduès, RNTI-E-1, p 169-192

[COU 01] COUTURIER R.,*Traitement de l'analyse statistique implicative dans CHIC* , Actes des Journées « Fouille des données par l'analyse statistique implicative », IUFM Caen, 2001, 33-50

[DID 72] DIDAY E., *Nouvelles méthodes et nouveaux concepts en classification automatique et reconnaissance des formes*, Thèse d'Etat, Université de Paris VI, 1972.

[DUB 87] DUBOIS D. et PRADE H., *Théorie des possibilités. Applications à la représentation des connaissances en informatique*, Masson

[GRA 79] GRAS R, *Contribution à l'étude expérimentale et à l'analyse de certaines acquisitions cognitives et de certains objectifs didactiques en mathématiques*, Thèse d'Etat, Rennes 1

[GRA 96] GRAS R., AG ALMOULOUD S., BAILLEUL M.,LARHER A.,POLO M., RATSIMBA-RAJOHN H. et TOTOHASINA A., *L'implication Statistique*, La Pensée Sauvage, Grenoble

[GRA 01a] GRAS R., KUNTZ P. et BRIAND H., *Les fondements de l'analyse statistique implicative et quelques prolongements pour la fouille de données*, Mathématiques et Sciences Humaines, n° 154-155, p 9-29, ISSN 0987 6936

[GRA 01b] GRAS R., DIDAY E., KUNTZ P. et COUTURIER R., *Variables sur intervalles et variables-intervalles en analyse implicative*, Actes du 8ème Congrès de la SFC de Pointe à Pitre, 17-21 décembre 2001, 166-173

[LAG 98] LAGRANGE J.B., *Analyse implicative d'un ensemble de variables numériques*, Revue de Statistique Appliquée, 1998, 71-93.

[LER 81] LERMAN I.C., *Classification et analyse ordinale des données*, Dunod, 1981.

[LOF 01] LOFTI A. et ZADEH L.A., *From computing with numbers to computing with words from manipulation of measurements to manipulation of perception, in Proceedings* "Human and machine perception". Ed. V. Cantoni, V. Di Gesù, A. Setti e D. Tegolo, Kluwer Academic, New York, 2001.

[REG 04] REGNIER J.C. et GRAS R : *Statistique de rangs et analyse statistique implicative*, Revue de Statistique Appliquée, à paraître en 2005

[SPA 04] SPAGNOLO F. et GRAS R., *A new approach in Zadeh's classification : fuzzy implication through statistic implication*, NAFIPS 2004, 23nd Conference of the North American Fuzzy Information Processing Society, june 27-30, Banff, AB Canada

[ZAD 97] ZADEH L.A., *Toward a Theory of Fuzzy Information Granulation and its Centrality in Human Reasoning and Fuzzy Logic*, Fuzzy Sets and Systems 90, pp. 111-127.

L'analyse d'un sous-ensemble de lignes ou colonnes en analyse des correspondances

Michael Greenacre[1] & **Rafael Pardo**[2]

[1]*Departament d'Economia i Empresa*
Universitat Pompeu Fabra
Ramon Trias Fargas, 25-27
08005 Barcelona (Catalunya), Espanya

[2]*Fundación BBVA*
Paseo de Recoletos, 10
28001 Madrid, España

RÉSUMÉ. Nous montrons comment l'analyse des correspondances s'applique à un sous-ensemble de lignes ou colonnes d'un tableau croisé ou – dans le cas de l'analyse des correspondances multiples – à un sous-ensemble de modalités. L'idée centrale est de considérer d'une part le sous-ensemble de la matrice des profils (non des données originales), et d'autre part de conserver les masses et la métrique du tableau originale (non du sous-ensemble). Ainsi il est possible de simplifier l'interprétation du tableau en le décomposant en sous-tableaux et en décomposant en même temps l'inertie totale en termes associés à ces sous-tableaux. Nous illustrons la méthodologie avec deux exemples d'analyse des correspondances simples et multiples.

MOTS-CLÉS : Analyse des correspondances, analyse des correspondances multiples, positionnement multidimensionnel pondéré, analyse en composantes principales généralisée, sous-ensemble.

1 Introduction

Dans la pratique de l'analyse des correspondances d'un tableau à I lignes et J colonnes, on analyse habituellement toutes les lignes et colonnes, en définissant parfois quelques lignes ou colonnes comme éléments supplémentaires, c'est à dire des points qui n'interviennent pas dans le calcul des axes mais qui sont projetés sur l'espace principal a posteriori pour faciliter l'interprétation du graphique. Ici nous voulons également ignorer certains éléments (lignes ou colonnes) mais proposons un traitement différent des éléments actifs.

L'approche classique serait de calculer les profils dans le sous-tableau de données, comme si le sous-tableau était le tableau original, avec ses propres masses et métrique. Nous proposons de calculer les profils ainsi que les masses des lignes et colonnes du tableau original et d'analyser le sous-tableau, en utilisant non pas les masses du sous-tableau pour pondérer les points et définir la métrique, mais les masses originales.L'analyse consiste alors à appliquer l'algorithme du *« weighted metric multidimensional scaling»* (que l'on peut traduire par : positionnement multidimensionnel pondéré, ou analyse en composantes principales généralisée) aux sous-profils avec leurs masses et métrique originales.

Cette idée simple a beaucoup d'avantages. Par exemple, si on décompose le tableau original en sous-tableaux, il est alors possible de décomposer l'inertie totale en termes associés chaque tableau. De plus, l'interprétation d'un grand tableau est simplifiée parce qu'il est possible d'interpréter un seul sous-tableau

à la fois, ce qui est particulièrement utile quand il existe des sous-ensembles naturels de lignes ou de colonnes.
La même idée peut être appliquée de manière presque identique à l'analyse des correspondances multiples, où les graphiques avec des dizaines de points sont fréquemment très difficiles à lire. Dans ce cas, il est possible de choisir toutes les colonnes que coïncident avec une modalité de réponse, par exemple 'très en accord' (« *strongly agree* ») et de représenter les individus par rapport à ces modalités seulement. De la même manière il est possible d'analyser les modalités de non-réponse (*non-response, « don't know »*) pour étudier la relation entre questions de ce phénomène.

2 Méthodologie

2.1 Analyse en composantes principales généralisée

La solution plus générale dont nous avons besoin est la suivante. Supposons que la matrice à représenter est $\mathbf{Y}$ ($n \times m$), centrée par rapport aux lignes ou colonnes ou les deux. Si les lignes sont des individus ou groupes d'individus et les colonnes sont des variables (dans notre cas, des modalités de réponses) le centrage se fera par rapport au centre de gravité (centroïde) des lignes en soustrayant les moyennes des colonnes. Soit $\mathbf{D}_r$ ($n \times n$) et $\mathbf{D}_w$ ($m \times m$) les matrices diagonales des masses-lignes et poids-colonnes, où les masses donnent une importance différente à chaque ligne et les poids-colonnes servent à normaliser les contributions des variables dans le calcul des distances entre les lignes. Nous pouvons supposer que la somme des masses des lignes est égale à 1. Nous cherchons le sous-espace de faible dimension qui s'ajuste le mieux aux lignes en minimisant la fonction suivante :

$$\mathrm{In}(\mathbf{Y} - \hat{\mathbf{Y}}) = \sum_{i=1}^{n} r_i \, (\mathbf{y}_i - \hat{\mathbf{y}}_i)^{\mathsf{T}} \mathbf{D}_w \, (\mathbf{y}_i - \hat{\mathbf{y}}_i) \qquad (1)$$

où $\hat{\mathbf{y}}_i$ est la i-ième ligne de $\hat{\mathbf{Y}}$, approximation dans le sous-espace qui est la plus proche à la i-ième ligne $\mathbf{y}_i$ de $\mathbf{Y}$. La fonction In(·) est l'*inertie* de la différence entre la matrice originale et son approximation dans l'espace réduit. L'*inertie totale*, qui est une mesure de la dispersion des points-lignes dans l'espace total de dimension m, est égal à In($\mathbf{Y}$).
Il est bien connu (voir, par exemple, [GRE 84], *Appendice*), que la solution peut être trouvée efficacement et sous une forme compacte en utilisant la décomposition matricielle qui s'appelle «décomposition généralisée en valeurs singulières» (GSVD) de la matrice $\mathbf{Y}$. Normalement comme seulement une SVD ordinaire (non généralisée, sans poids…) est disponible, par exemple dans le logiciel R ([VENSMI 03]), on trouve la solution en pré-transformant la matrice $\mathbf{Y}$, puis en calculant la SVD, et finalement en post-transformant les résultats pour trouver la solution généralisée. L'algorithme est le suivant (voir également [GRE 04]):

1. $\mathbf{S} = \mathbf{D}_r^{1/2} \mathbf{Y} \mathbf{D}_w^{1/2}$ (2)
2. $\mathbf{S} = \mathbf{U} \mathbf{D}_\alpha \mathbf{V}^{\mathsf{T}}$ (3)
3. Coordonnées principales des lignes: $\mathbf{F} = \mathbf{D}_r^{-1/2} \mathbf{U} \mathbf{D}_\alpha$ (4)
4. Axes principaux des lignes: $\mathbf{D}_w^{-1/2} \mathbf{V}$ (5)
5. Coordonnées standards des colonnes: $\mathbf{G} = \mathbf{D}_w^{1/2} \mathbf{V}$ (6)

2.2 Analyse des correspondances d'un sous-ensemble

L'analyse des correspondances (CA) est l'algorithme ci-dessus appliqué au tableau de fréquences $\mathbf{N}$. On divise $\mathbf{N}$ par son total n pour obtenir le tableau de correspondances $\mathbf{P} = (1/n)\,\mathbf{N}$. Soient $\mathbf{r}$ et $\mathbf{c}$ les totaux

marginaux de **P**, et $\mathbf{D}_r$ et $\mathbf{D}_c$ les matrices diagonales de **r** et **c**. On considère le tableau comme un ensemble de lignes (le raisonnement est identique si on le considère comme un ensemble de colonnes), et on calcule les profils-lignes en divisant **P** par ses totaux-lignes: $\mathbf{D}_r^{-1}\mathbf{P}$. CA est alors l'analyse des profils-lignes dans $\mathbf{D}_r^{-1}\mathbf{P}$, où les distances entre profils sont mesurées selon la métrique du chi-deux définie par $\mathbf{D}_c^{-1}$ et les profils sont pondérés par leurs masses dans $\mathbf{D}_r$. La moyenne pondérée des profils-lignes est égale au vecteur $\mathbf{c}^\mathsf{T}$ des fréquences marginales des colonnes, et la solution est donnée par (2)-(6) ci-dessus avec la matrice centrée **Y** égale à $\mathbf{D}_r^{-1}\mathbf{P}-\mathbf{1}\mathbf{c}^\mathsf{T}$, $\mathbf{D}_r$ égale à l'actuelle $\mathbf{D}_r$ et $\mathbf{D}_w$ égale à $\mathbf{D}_c^{-1}$.

On applique maintenant cette théorie à un sous-ensemble du tableau, en gardant la même pondération des lignes et colonnes, mais en traitant un sous-ensemble de colonnes de la matrice. On évite le re-calcul des profils-lignes. En effet soit **H** un sous-ensemble de colonnes de $\mathbf{D}_r^{-1}\mathbf{P}$ et soit **h** le sous-vecteur de **c** correspondant de poids-colonnes, comme précédemment, la moyenne pondérée des lignes sera égale à **h** : $\mathbf{H}^\mathsf{T}\mathbf{r} = \mathbf{h}$. L'analyse des correspondances du sous-tableau (que nous appelons s-CA) est une analyse en composantes principales généralisée de **H** avec des masses-lignes **r** dans $\mathbf{D}_r$ comme précédemment, et la métrique définie par $\mathbf{D}_h^{-1}$ sur $\mathbf{D}_h$ est la matrice diagonale de **h**. La solution de s-CA en découle en utilisant les formules (2)-(6) avec **Y** égale à $\mathbf{H} - \mathbf{1}\mathbf{h}^\mathsf{T} = (\mathbf{I} - \mathbf{1}\mathbf{r}^\mathsf{T})\mathbf{H}$, $\mathbf{D}_r$ égale à l'actuelle $\mathbf{D}_r$ et $\mathbf{D}_w$ égale à $\mathbf{D}_h^{-1}$. La matrice (2) qui se décompose maintenant est donc :

$$\mathbf{S} = \mathbf{D}_r^{1/2}(\mathbf{I} - \mathbf{1}\mathbf{r}^\mathsf{T})\mathbf{H}\mathbf{D}_h^{-1/2} \tag{7}$$

et les coordonnées principales et standards des lignes et colonnes sont:

$$\mathbf{F} = \mathbf{D}_r^{-1/2}\mathbf{U}\mathbf{D}_\alpha \qquad \mathbf{G} = \mathbf{D}_h^{-1/2}\mathbf{V} \tag{8}$$

Un autre résultat intéressant de cette approche est que l'inertie totale du tableau original se décompose en termes associés à chacun des sous-tableaux. Par exemple, si on définit une partition complète des colonnes de **Y** par des matrices $\mathbf{H}_1, \mathbf{H}_2, \ldots, \mathbf{H}_p$ l'inertie de **Y** est égale à la somme des inerties des sous-tableaux $\mathbf{H}_k$. [GREPAR 04], un document de travail avec plus de détails, peut être téléchargé à partir de la toile. Dans la conférence nous donnerons deux applications de cette méthodologie : un dans le cadre de l'analyse des correspondances simples de tableaux juxtaposés et un autre pour un tableau disjonctif en analyse des correspondances multiples (MCA). Ces exemples démontreront l'utilité de cette approche et la simplification qui en résulte dans l'interprétation des graphiques des grands tableaux.

3 Bibliographie

[GRE 84] GREENACRE M.J., *Theory and Applications of Correspondence Analysis*, Academic Press, 1984.

[GRE 04] GREENACRE M.J. Weighted metric multidimensional scaling. Conférence présentée au *German Classification Society Meeting*, Dortmund, mars 2004. Working Paper 777, Dept Economics and Business, Universitat Pompeu Fabra, Barcelona.

[GREPAR 04] GREENACRE M.J., PARDO, R. Subset correspondence Analysis: visualizing relationships among a selected set of response categories from a questionnaire survey. Working paper number 793, Dept of Economics and Business, Universitat Pompeu Fabra, Barcelona.

[VENSMI 03] VENABLES W.N., SMITH D.M. *An Introduction to* R. `www.r-project.org`.

Classification directe et croisée sur les données continues

F.-X. Jollois*, M. Nadif**

** CRIP5*
Université René Descartes Paris 5
45 rue des Saints Pères
75016 Paris, France

*** LITA*
Université de Metz
Île du Saulcy
75000 Metz, France

RÉSUMÉ. Alors que les méthodes usuelles de classification cherchent une partition soit sur l'ensemble des instances, soit sur l'ensemble des attributs, il existe des méthodes de classification croisée et directe qui permettant d'obtenir des blocs homogènes. Les méthodes de classification croisée atteignent cet objectif à partir d'une partition des instances et une partition des attribut, recherchées simultanément. Les méthodes de classification directe s'appliquent directement sur les données et permettent d'obtenir des blocs de données homogènes de toute taille, ainsi que des hiérarchies des classes en lignes et en colonnes. Combinant les avantages des deux méthodes, nous présentons ici une méthodologie permettant de travailler sur de grandes bases de données.

MOTS-CLÉS : Classification directe, Classification croisée..

1 Introduction

Lorsque le but d'une classification est d'obtenir une structure en blocs homogènes, nombre d'utilisateurs appliquent des algorithmes de classification simple sur l'ensemble des instances et sur l'ensemble des attributs séparément, les blocs résultent du croisement des deux partitions obtenues. Une telle démarche ne permet pas d'expliquer la relation spécifique pouvant exister entre un groupe d'instances et un groupe d'attributs. Ainsi, lorsque les données sont comparables, il est préférable d'appliquer des algorithmes de classification croisée tel que l'algorithme *Croeuc* [GOV 83] qui cherche simultanément une partition en lignes et une partition en colonnes. Les centres des blocs ainsi obtenus constituent une matrice de taille réduite offrant un résumé des données. Une autre façon de résumer l'information consiste à utiliser un algorithme de classification directe, comme *Two-way splitting* [HAR 75], qui cherche à obtenir des blocs de données homogènes et de toute taille.

Malgré sa rapidité et son efficacité à traiter des tables de grande taille, l'algorithme *Croeuc* présente un défaut majeur ; il nécessite la connaissance des nombres de classes en lignes et en colonnes. Par contre, l'algorithme *Two-way splitting* s'affranchit de cette hypothèse et offre en plus une hiérarchie sur les instances et les attributs. Mais sa complexité rend son utilisation impossible sur des tables de données de grande taille. Nous présentons donc ici une combinaison de ces deux algorithmes afin de pallier les inconvénients de chacun.

2 L'algorithme *Croeuc*

Dans la suite, la matrice des données est définie par $\mathbf{x} = \{ x_i^j ; i \in I \text{ et } j \in J\}$, où I est l'ensemble des n instances (lignes, objets, observations), et J est l'ensemble des d attributs (colonnes, attributs). On cherche à optimiser un critère $W(\mathbf{z}, \mathbf{w}, \mathbf{g})$, où $\mathbf{z} = (z_1, ..., z_s)$ est une partition de I en s classes, $\mathbf{w} = (w_1, ..., w_m)$ est une partition de J en m classes et $\mathbf{g} = (g_k^\ell)$ est une matrice s x m, qui peut être vue comme un résumé de la matrice de données $\mathbf{x}$. Une définition plus précise de ce résumé et du critère W dépendra de la nature des données. La recherche des partitions optimales $\mathbf{z}$ et $\mathbf{w}$ est effectuée en utilisant un algorithme itératif. Lorsque les données sont continues, en prenant la somme des distances euclidiennes au carré comme une mesure de la déviation entre la matrice $\mathbf{x}$, et la structure décrite par $\mathbf{z}$, $\mathbf{w}$ et $\mathbf{g}$, l'algorithme *Croeuc* consiste à trouver une paire de partition ($\mathbf{z}$, $\mathbf{w}$) et le paramètre $\mathbf{g}$ correspondant, tel que le critère suivant soit minimisé :

$$W(\mathbf{z}, \mathbf{w}, \mathbf{g}) = \sum_{k=1}^{s} \sum_{\ell=1}^{m} \sum_{i \in z_k} \sum_{j \in w_\ell} \left(x_i^j - g_k^\ell \right)^2 ,$$

où g_k^ℓ est le centre du bloc x_k^ℓ. Il est facile de voir que pour $\mathbf{z}$ et $\mathbf{w}$ fixés, la valeur optimale de g_k^ℓ est donnée par la moyenne de tous les x_i^j du bloc *(k,ℓ)*. Les différentes étapes de *Croeuc* sont :

1. Les paramètres initiaux ($\mathbf{z}^{(0)}$, $\mathbf{w}^{(0)}$, $\mathbf{g}^{(0)}$) peuvent être choisis au hasard.
2. Calculer ($\mathbf{z}^{(q+1)}$, $\mathbf{w}^{(q+1)}$, $\mathbf{g}^{(q+1)}$) à partir de ($\mathbf{z}^{(q)}$, $\mathbf{w}^{(q)}$, $\mathbf{g}^{(q)}$) :
   a. Calculer ($\mathbf{z}^{(q+1)}$, $\mathbf{w}^{(q)}$, $\mathbf{g}^{(q+0.5)}$) à partir de ($\mathbf{z}^{(q)}$, $\mathbf{w}^{(q)}$, $\mathbf{g}^{(q)}$).
   b. Calculer ($\mathbf{z}^{(q+1)}$, $\mathbf{w}^{(q+1)}$, $\mathbf{g}^{(q+1)}$) à partir de ($\mathbf{z}^{(q+1)}$, $\mathbf{w}^{(q)}$, $\mathbf{g}^{(q+0.5)}$).
3. Répéter l'étape 2 jusqu'à la convergence

Dans les étapes 2(a) et 2(b), pour trouver $\mathbf{z}^{(q+1)}$ et $\mathbf{w}^{(q+1)}$ optimaux, nous cherchons à minimiser alternativement le critère en ligne $W(\mathbf{z},\mathbf{g}/\mathbf{w}) = \sum_{k=1}^{s} \sum_{i \in z_k} \sum_{\ell=1}^{m} \#w_\ell \left(u_i^\ell - g_k^\ell \right)^2$, avec $u_i^\ell = \frac{\sum_{j \in w_\ell} x_i^j}{\#w_\ell}$, et le critère en colonne $W(\mathbf{w},\mathbf{g}/\mathbf{z}) = \sum_{\ell=1}^{m} \sum_{j \in w_\ell} \sum_{k=1}^{s} \#z_k \left(v_k^j - g_k^\ell \right)^2$ avec $v_k^j = \frac{\sum_{i \in z_k} x_i^j}{\#z_k}$. Ici, # représente la cardinalité. L'étape 2(a) est effectuée par l'application de l'algorithme *k*-means, utilisant la matrice n x m (u_i^ℓ). Alternativement, l'étape 2(b) est obtenue par l'application de l'algorithme *k*-means utilisant cette fois-ci la matrice s x d (v_k^j). Ainsi, à la convergence, nous obtenons des blocs homogènes en réorganisant les lignes et les colonnes selon les partitions $\mathbf{z}$ et $\mathbf{w}$. De cette manière, chaque bloc *(k,ℓ)*, défini par les éléments x_i^j pour $i \in z_k$ et $j \in w_\ell$, est caractérisé par g_k^ℓ .

L'intérêt de cet algorithme a été mis en évidence par comparaison avec *k*-means appliqué séparément sur les instances et les attributs [NAD 04]. Par sa simplicité et sa rapidité, cet algorithme peut s'appliquer sur des données comparables de grande taille. Malheureusement, il requiert la connaissance du nombre de classes en lignes et en colonnes.

3 L'algorithme *Two-way splitting*

Lorsque les données sont directement comparables d'un attribut à un autre, Hartigan [HAR 75] propose un algorithme divisif, *Two-way splitting*, qui choisit à chaque étape entre une division de l'ensemble des instances et une division de l'ensemble des attributs. Ce choix est basé sur la réduction au maximum de l'hétérogénéité du groupe d'instances ou de variables divisé. Afin de respecter les contraintes hiérarchiques imposées pour cet algorithme, les divisions effectuées à une étape ne sont jamais remises en cause aux étapes suivantes. Cet algorithme ne nécessite pas de savoir à l'avance le nombre de blocs que l'on veut obtenir. Il peut être décrit succinctement de la manière suivante :

1. Fixer un seuil minimum de variance T, et démarrer avec les instances dans une seule classe et les attributs dans une seule classe.
2. Calculer les variances moyennes de chaque classe en lignes et de chaque classe en colonnes. Les lignes et les colonnes ayant une variance inférieure à T ne sont pas prises en compte. Ainsi, une classe en lignes ne contenant que des instances avec une variance inférieure à T ne sera plus découpée. De même en colonnes.
3. Choisir la classe en lignes ou en colonnes qui a la plus grande variance.
4. Découper cette classe en deux en utilisant une variante de k-means, en ne retenant que les blocs où la variance est supérieure à T.
5. Recommencer à partir de l'étape 2, jusqu'à ce que toutes les variances de chaque bloc soient inférieures au seuil fixé par l'utilisateur.

Cet algorithme permet de mettre en évidence des structures plus fines que celles de *Croeuc*. Notons que nous disposons de plus d'une hiérarchie en lignes et une hiérarchie en colonnes.

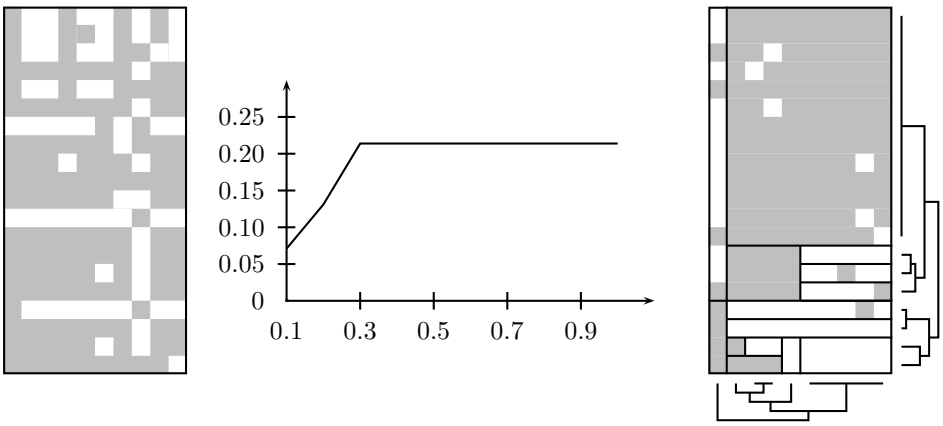


Figure 1 : Matrice initiale binaire (gris : 1, blanc : 0), variances moyennes pour chaque seuil et matrice réordonnée selon les résultats de *Two-way splitting* avec le seuil 0.2.

Pour illustrer cet algorithme, nous l'avons appliqué sur des données binaires de taille nxd = 20x10. Pour le choix du seui, nous avons testé les différentes valeurs de 0.1 à 1.0 avec un pas de 0.1. Les moyennes des variances de chaque bloc sont reportées, pour chaque seuil, dans la figure 1. A partir de 0.3, *Two-way splitting* ne sépare plus les données, et garde un seul bloc. C'est pourquoi nous avons choisi de retenir un seuil de 0.2. Nous présentons la matrice réordonnée selon les blocs et les hiérarchies en lignes et en colonnes fournies par *Two-way splitting*.

Malheureusement, cet algorithme nécessite le calcul à chaque étape d'un grand nombre de variances, ce qui rend son utilisation sur des données de grande taille inadaptée. De plus, le choix du seuil n'est pas automatique, et nécessite soit une connaissance préalable de celui-ci, soit un test de plusieurs valeurs candidates.

4 Combinaison des algorithmes et illustration

Comme l'algorithme *Two-way splitting* est bien adapté lorsque les données sont comparables, il est possible de le combiner avec un algorithme de type *Croeuc*. En effet, ce dernier propose une matrice d'information réduite, contenant les moyennes de chaque bloc. Nous obtenons donc un tableau sur lequel *Two-way splitting* peut parfaitement s'appliquer. De plus, grâce à cette combinaison, nous pouvons nous affranchir du problème du nombre de classes pour *Croeuc*, en choisissant s et m assez grands. Ensuite, nous pouvons appliquer *Two-way splitting* sur les centres obtenus et avoir ainsi une structure claire de la matrice d'information (et donc des données de base).

Pour illustrer notre démarche, nous l'appliquons sur des données simulées de taille *n*x*d* = 5000x500, suivant 4 classes en lignes et 3 classes en colonnes. Comme le critère optimisé par *Croeuc* est associé à un modèle de mélange Gaussien croisé [GOV 03] ; les données de chaque bloc *(k,ℓ)* suivent une loi normale de centre μ_k^ℓ (les centres sont présentés dans figure 2) et de variance σ^2 supposée égale dans tous les blocs. Les proportions sont choisies égales (en lignes et en colonnes). Pour résumer la matrice de données, nous avons choisi d'appliquer *Croeuc* en prenant *s* = 10 et *m* = 5. Il est bien sûr possible de choisir d'autres nombres de classes, en fonction de la granularité désirée.

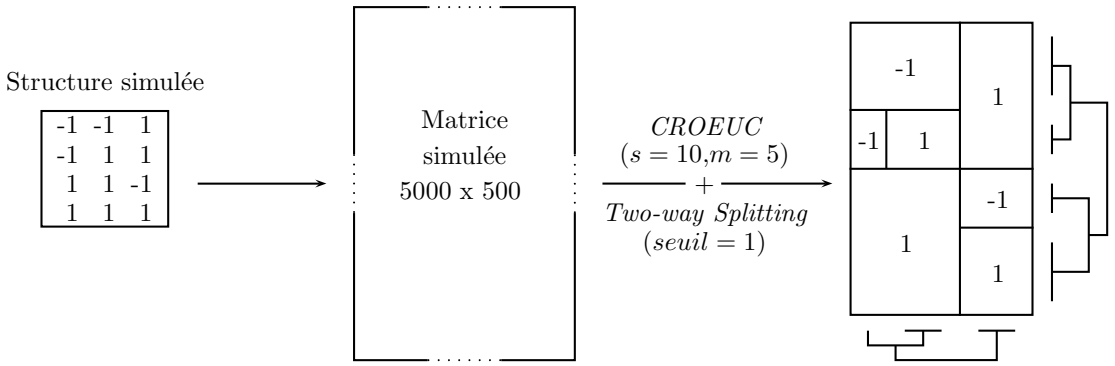


Figure 2 : Schéma de simulation (avec les moyennes des 4x3 gaussiennes, à gauche) et structure obtenue à l'aide de *Two-way splitting* appliqué sur les centres donnés par *Croeuc* (à droite).

Appliqué avec un seuil de 1, *Two-way splitting* fournit un découpage en bloc approprié, et identique à ceux fournis par les seuils inférieurs (0.1, 0.2, 0.3, …, 0.9). Nous avons donc décidé de garder ce seuil de 1. Dans la figure 2, on voit très bien que la structure proposée par *Two-way splitting* est la même que celle simulée.

5 Conclusion et Perspectives

Après avoir présenté un algorithme de classification croisée, *Croeuc*, nous avons présenté un algorithme de classification directe, *Two-way splitting*. Profitant des avantages des deux méthodes, nous avons proposé une méthodologie d'identification de blocs homogènes en appliquant *Two-way splitting* sur la matrice résumée obtenue à l'aide de *Croeuc*. Les premiers résultats obtenus sont très encourageants et nous persuadent de l'intérêt évident d'une telle démarche. Notons que celle-ci est applicable uniquement dans le cas où les données sont comparables telles que les données de type biopuces. Actuellement, nous testons cette méthode sur des données réelles.

6 Bibliographie

[GOV 83] GOVAERT G., Classification Croisé, Thèse d'Etat, Université de Paris 6, France, 1983.

[GOV 03] GOVAERT G., NADIF M., Clustering with Block Mixture Models, *Pattern Recognition*, 36, 463-473, 2003.

[HAR 75] HARTIGAN J., Direct Splitting, Chap. 14, 251-277. Dans *Clustering Algorithms*, John Wiley & Sons, New York, 1975.

[NAD04] NADIF M., JOLLOIS F.-X., GOVAERT G., Block clustering for large continuous data sets, *AISTA 2004*, Luxembourg, 15-18 novembre, 2004.

Quality control and data correction in high-throughput screening

Dmytro Kevorkov and Vladimir Makarenkov

Département d'informatique,
Université du Québec à Montréal,
Case postale 8888, succursale Centre-ville
Montréal (Québec) Canada, H3C 3P8
Courriels : kevorkov@lacim.uqam.ca, makarenkov.vladimir@uqam.ca

---

ABSTRACT

High-throughput screening (HTS) plays a central role in modern drug discovery, allowing for testing of more than 100,000 compounds per screen. Often compounds are screened only once and quality of measurements has a significant influence on the hit selection. We have developed an effective statistical procedure to control quality of HTS assays. It is intended to detect systematic errors in HTS measurements and make correction of experimental data. An experimental assay from the McMaster HTS laboratory was examined in this paper. We identified and studied systematic errors present in this dataset. The correction of the experimental data significantly improved the hit distribution for this assay.

KEYWORDS: High-throughput screening, quality control, systematic error, data correction.

---

1 Introduction

Quality control is an important part of the data analysis in the high-throughput screening (HTS). HTS is a modern and effective technology for drug discovery, allowing for rapid screening of large compound collections against a variety of putative drug targets. A typical HTS operation in the pharmaceutical industry allows for screening of more than 100,000 compounds per screen and generates approximately 50 million data points per year [HEU 02]. HTS operates with samples in microliter volumes that are arranged in two-dimensional plates. A typical HTS plates contain 96 (12×8) or 352 (22×16) samples. The hit selection procedure in primary HTS screens is mainly performed by automatic routines. Hits can be defined as positive signals that represent biologically or chemically active compounds. These compounds are potential targets for drag discovery. Quality of measurements is extremely important for a correct hit selection. Random and systematic errors can cause a misinterpretation of HTS signals. They can induce either underestimation (false negatives) or overestimation (false positives) of measured signals. Various methods dealing with quality control, data correction and hit selection are available in the scientific literature. These methods are discussed in details in the papers by Heuer et al. [HEU 02], Gunter et al. [GUN 03], Brideau et al. [BRI 03], Heyse [HEY 02], and Zhang et al. [ZHA 99, ZHA 00]. However, statistical methods that analyze and remove systematic errors in HTS assays are poorly developed

compared to those dealing with microarrays. There are various sources of systematic errors. Some of them are mentioned in the article of Heuer et al. [HEU 02]:

- Systematic errors caused by ageing, reagents evaporation or decay of cells can be recognized as smooth trends in the plate's means/medians;
- Errors in liquid handling and malfunction of pipettes can also generate localized deviations of expected data values;
- Variation in incubation time, time drift in measuring different wells or different plates, and reader effects may be recognized as smooth attenuations of measurement over an assay.

Heuer et al. [HEU 02] and Brideau et al. [BRI 03] demonstrated examples of systematic signal variations that are present in all plates of an assay. For instance, Brideau et al. [BRI 03] illustrates a systematic error caused by the positional effect of detector. Throughout the entire screening campaign involving more than 1000 plates, signal values in Row A were on average 14% lower than those in Row P (see Brideau et al. [BRI 03], Figure 1).

In this article, we present a new effective statistical procedure, allowing one to detect systematic errors in HTS assays and minimize their impact on the hit selection procedure.

2 Experimental procedure and results

2.1 Experimental data

We have selected for evaluation an HTS assay created and made available at the HTS Laboratory of the McMaster University. This assay was proposed as a benchmark for Data Mining and Docking Competition (http://hts.mcmaster.ca/HTSDataMiningCompetition.htm). It consists of a screen of compounds that inhibit the *Escherichia coli* dihydrofolate reductase. The dataset and the detailed description of the experimental procedure are available at: http://hts.mcmaster.ca/Competition_1.html. The assay consists of 1248 plates. Each plate contains measurements for 80 testing compounds arranged in 8 rows and 10 columns.

2.2 Data classification

The experimental data should be properly classified to carry out a correct statistical analysis. The screened samples can be divided into two groups. The first group contains inactive samples. The majority of samples in primary screens are inactive. Since they are inactive, they should have close average values (measured for a sufficient number of plates) and the values variability is caused mainly by random or systematic errors. The second group contains active samples and outliers. Their values differ substantially from the inactive ones. The values of active samples are caused by a biochemical reaction and can be classified as hits. The values of outliers are caused by random errors and should be disqualified from the analysis.

2.3 Statistical analysis and data correction

Ideally, inactive samples have similar values and generate a plane surface. In a real case, random errors produce residuals. For a large number of plates considered, the residuals should compensate each other during the computation of the mean values at each well. Systematic errors generate repeatable local artifacts and smooth global drifts at the assay surface.

To carry out the analysis and correction of experimental HTS data, the following steps have been carried out:

- Logarithmic transformation of raw HTS data;
- Plate normalization for all samples;
- Analysis of hit distribution for raw data;
- Hit and outlier elimination;
- Plate normalization of inactive samples (Normalization I);

- Well normalization of inactive samples (Normalization II);
- Data correction;
- Plate normalization for all samples;
- Analysis of hit distribution for corrected data;

Because the data under study have a Gaussian distribution, we first performed a logarithmic transformation of raw data, and then normalized values in plates using Normalization to Zero Mean and Unit Standard Deviation, which is as follows:

$$x'_i = \frac{x_i - \mu}{\sigma}, \tag{1}$$

where $x_i = \log(x_{raw})$ - input element, x_{raw} - raw element, x'_i - normalized output element, μ - mean value, σ - standard deviation. The output data conditions will be $\mu_{x'} = 0$ and $\sigma_{x'} = 1$. This pre-processing is necessary to compare and sum measurements in different plates; the detailed discussion can be found in the paper by Kevorkov and Makarenkov [KEV 04].

The presence of systematic error in an assay can be detected by analyzing its hit distribution surface. From a statistical point of view, hits should be distributed evenly over the hit distribution surface. We selected as hits experimental values that deviated from the plate means for more than 1σ; this is a common strategy for hit selection in HTS data analysis. The significant variation of the hit numbers shown in Figure 1a proves the presence of systematic errors in the experimental dataset.

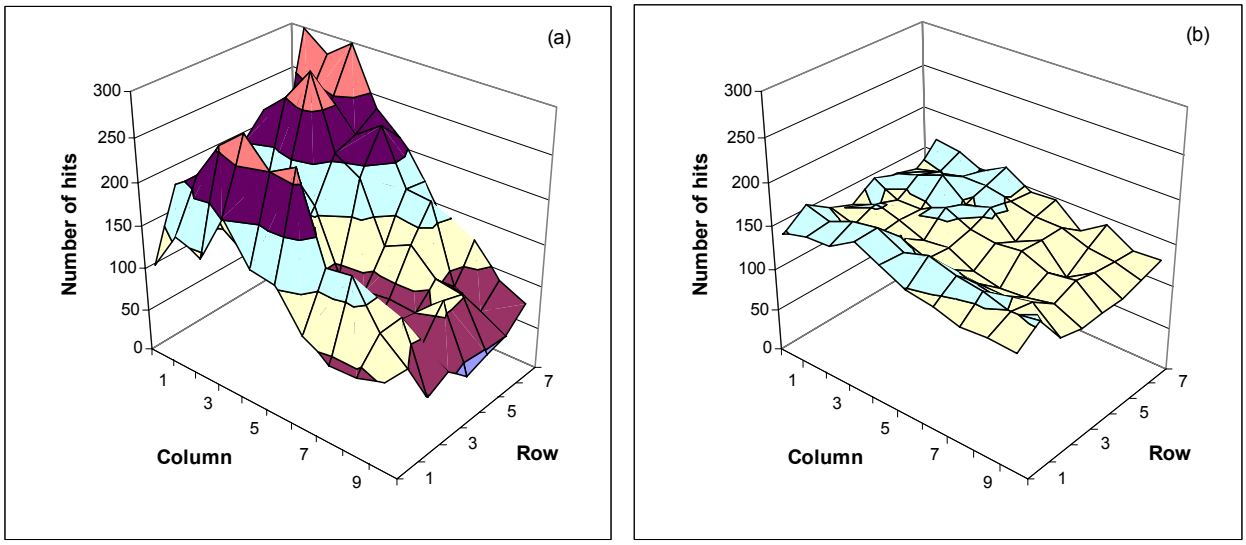


Figure 1. Hit distribution surfaces for the McMaster data (1248 plates): (a) raw data and (b) corrected data

In order to remove systematic error we analyzed inactive samples only. As we mentioned above they should have similar average values. As inactive were selected the values that deviated from the plate mean for less than 3σ; hit and outliers beyond the 3σ threshold were removed from the analysis. We normalized the inactive samples in plates using the Normalization to Zero Mean and Unit Standard Deviation (Normalization I). The normalized values of inactive samples should be close to zero.

Then, we analyzed the arrays of values for each specific well over all plates. An example of such an array is presented in Figure 2. It demonstrates fluctuation of inactive values after Normalization I for the well located in column 1 and row 8. The deviation of the mean value (equal to -0.37) from zero level can be interpreted as an impact of systematic error. The fluctuations of values around of the array mean can be interpreted as an influence of random errors. The values at the wells containing an important systematic error demonstrated substantial deviations from zero level. In order to remove the effect of systematic errors we normalized (using the Normalization to Zero Mean and Unit Standard Deviation) well values over all 1248 plates (Normalization II). Normalization II moved the mean well value to zero level and fitted the standard deviation to unity.

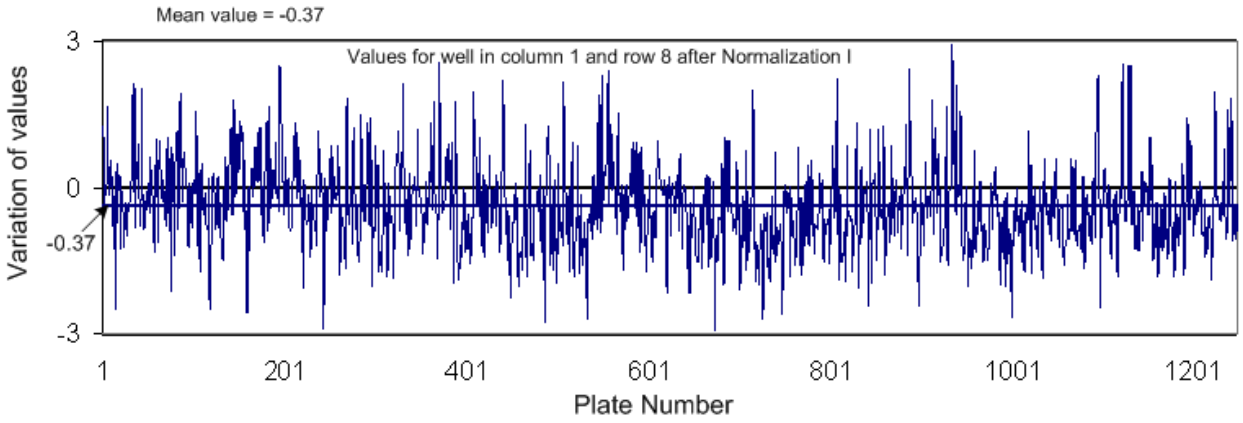


Figure 2. Variation of values in different plates for well in column 1 and row 8 (1248 plates).

The parameters used in Normalization I and II were employed to modify the complete dataset, i.e. both active and inactive samples were transformed by the same procedures and using the same parameters as the dataset of inactive samples. Then, we normalized the corrected values in plates and reexamined the hit distribution surface presented in Figure 1b. The hit distribution for the corrected data demonstrates significant improvements compared to the raw data. The minimum and maximum numbers of hits for the corrected dataset were 110 and 177, respectively, compared to 36 and 298 for the raw data.

The hit distributions by columns and rows are presented in Figure 3. Figure 3a (white columns) shows that values for the raw dataset systematically decrease from column 3 to column 10. The average number of hits per well in column 3, computed over 1248 plates, was 209. In contrast, the average number of hits per well in column 10 was only 65. Figure 3b demonstrates an important difference between the number of hits on the edges and in the middle of the plates for the raw data. The average number of hits per well in row 2 was 202, whereas in row 4 there were only 98 hits per well. Such a difference is unlikely due to random errors and, in our opinion, caused by systematic error of the measurements. The corrected dataset (black columns) provides a much better hit distribution: the average number of hits per well in column 3 was 154 and in column 10 was 123, whereas the average number of hits per well in row 2 was 163 and in row 4 was 130. Thus, the comparison of hit distributions for raw and corrected datasets demonstrated that the impact of systematic errors on raw data was significantly minimized after the correction proposed.

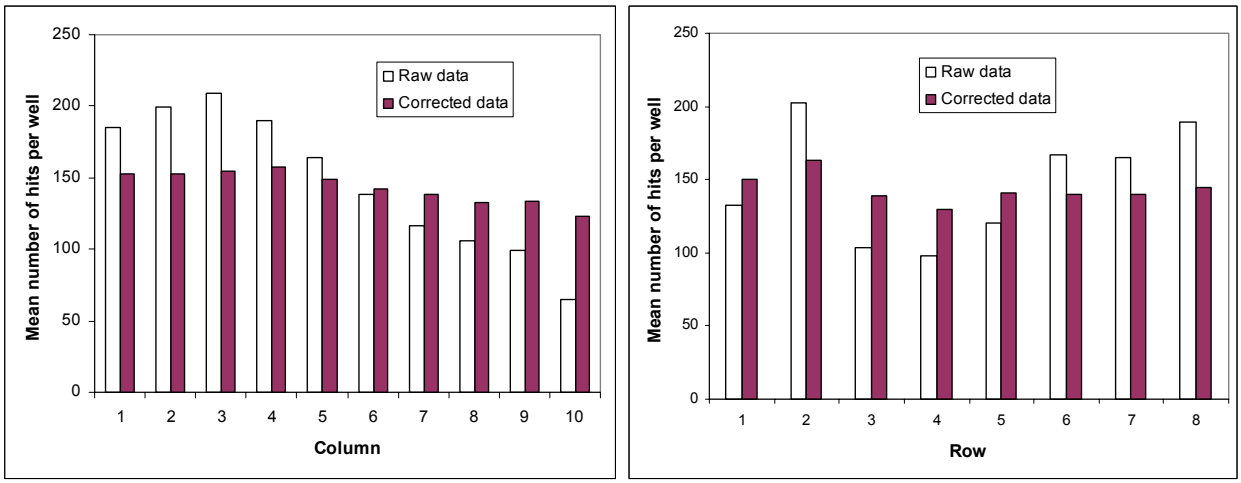


Figure 3. Hit distribution by columns and rows for the McMaster HTS assay (1248 plates).

3 Conclusion

In this paper we described a method for the quality control and data correction in high throughput screening. The described method enables one to analyze experimental HTS data, detect the presence of systematic error and correct trends and local fluctuations of the background surface. We examined experimental data from the McMaster HTS laboratory. The analysis of the hit distribution showed the presence of systematic errors in this dataset. Using the new method, we examined and corrected HTS assay containing 1248 plates. The analysis of the raw and corrected datasets demonstrated that the proposed procedure is able to improve the hit distribution. The impact of systematic error on measured values was significantly minimized after the correction. The software allowing researches to carry out the background evaluation analysis of HTS data is available upon request (distributed as a Windows console application and its C++ source code). A graphical version of this software will be freely available at our website (http://www.info.uqam.ca/~makarenv/hts.html) at the beginning of the year 2005.

4 Bibliography

[BRI 03] BRIDEAU C., GUNTER B., PIKOUNIS W., PAJNI N., LIAW A., "Improved statistical methods for hit selection in high-throughput screening", *J Biomol Screen*, vol. 8, 2003, p. 634-647.

[GUN 03] GUNTER B., BRIDEAU C., PIKOUNIS B., PAJNI N., LIAW A., "Statistical and graphical methods for quality control determination of high throughput screening data", *J Biomol Screen*, vol. 8, 2003, p. 624-633.

[HEU 02] HEUER C., HAENEL T., PRAUSE B., "A novel approach for quality control and correction of HTS data based on artificial intelligence", *The Pharmaceutical Discovery & Development Report 2003/03*, 2002, PharmaVentures Ltd., [Online] Retrieved from http://www.worldpharmaweb.com/pdd/new/overview5.pdf.

[HEY 02] HEYSE S., "Comprehensive analysis of high-throughput screening data", *Proc SPIE*, vol. 4626, 2002, p. 535-547.

[KEV 04] KEVORKOV D., MAKARENKOV V., "Statistical analysis of systematic errors in high-throughput screening", submitted to *J Biomol Screen*, 2004.

[ZHA 00] ZHANG J.H., CHUNG T.D.Y., OLDENBURG K.R., "Confirmation of Primary Active Substances from High Throughput Screening of Chemical and Biological Populations: A Statistical Approach and Practical Considerations", *J Comb Chem*, vol. 2, 2000, 258-265.

[ZHA 99] ZHANG J.H., CHUNG T.D.Y., OLDENBURG K.R., "A Simple Statistic Parameter for Use in Evaluation and Validation of High Throughput Screening Assays", *J Biomol Screen*, vol. 4, 1999, 67-73.

Enumération des graphes de k-arches étiquetés

Cédric Lamathe

LaCIM
Université du Québec à Montréal
Case postale 8888, succursale Centre-Ville
Montréal (Québec) Canada, H3C 3P8
lamathe@lacim.uqam.ca

RÉSUMÉ. *Dans cet article*[1]*, nous nous intéressons aux graphes de k-arches, une généralisation des arbres, contenant les k-arbres comme sous-classe. Nous montrons que le nombre de graphes de k-arches sur n sommets étiquetés, pour un entier fixé $k \geq 1$, est donné par $\binom{n}{k}^{n-k-1}$. A notre connaissance, c'est une suite d'entiers inédite à ce jour. Nous établissons ce résultat via une bijection entre les graphes de k-arches et certains mots dont les lettres sont des sous-ensembles de taille k de l'ensemble des sommets. Cette bijection généralise le code de Prüfer pour les arbres. Nous retrouvons également la formule de Cayley n^{n-2} qui énumère les arbres étiquetés à n sommets.*

MOTS-CLÉS : *Enumération, graphes de k-arches, bijection.*

1. Introduction

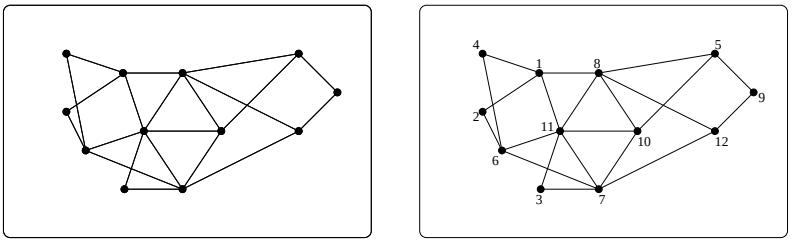
On définit récursivement la famille des *graphes de k-arches*, pour $k \geq 1$, comme la plus petite classe de graphes simples tels que :

1. un $(k-1)$-simplexe (*i.e.*, un graphe complet à k sommets) est un graphe de k-arches ;
2. si un graphe simple G possède un sommet v de degré k tel que le graphe $G - \{v\}$ obtenu à partir de G en enlevant v ainsi que les arêtes incidentes à v est un graphe de k-arches, alors G est un graphe de k-arches.

La figure 1 montre un graphe de 2-arches (on dit *graphe d'arche* dans ce cas) et un graphe de 2-arches étiqueté aux sommets, chacun d'eux étant construit sur 12 sommets. Lorsque $k = 1$, les graphes de 1-arches coïncident avec les arbres (de Cayley). De manière constructive, pour obtenir un graphe de k-arches à $n+1$ sommets à partir d'un possédant n sommets, il faut choisir k sommets et les joindre au nouveau sommet. Le terme *arche* évoque le fait d'attacher le nouveau sommet sur les k sommets sélectionnés.

Il n'existe que peu d'articles traitant des graphes de k-arches, à l'exception de [TOD 89], où ces graphes sont introduits, et [LEC 02]. Cependant, il existe une très vaste littérature sur une sous-classe des graphes de k-arches, celle des k-arbres, étudiée depuis la fin des années 60. La différence essentielle entre k-arbres et graphes de k-arches est que dans le cas des k-arbres, on impose que le nouveau sommet v soit attaché à k sommets mutuellement adjacents (*i.e.*, cela forme un graphe complet à $k+1$ sommets). On trouve l'énumération étiquetée des k arbres dans [BEI 69b, MOO 69] et dans [BEI 69a, PAL 69] pour le cas particulier $k = 2$. Harary et Palmer [HAR 73] traitent le cas de l'énumération non étiquetée

1. Cet article est la version conférence de l'article "The number of labelled k-arch graphs", J. Int. Seq., Vol **7**, article 04.3.1 (2004)



a) b)

FIG. 1. Graphe d'arches sur 12 sommets a) non étiqueté, b) étiqueté aux sommets.

des 2-arbres, de même que Fowler et al. [FOW 02] qui donne également des formules asymptotiques. Mentionnons également [LAB 04, BOU] pour l'énumération de variations des 2-arbres et [LAM 04], où plusieurs spécialisations des 2-arbres sont considérées. Finalement, Labelle et al. [LAB 03] propose une classification des 2-arbres (exter)planaires selon leurs symétries. En dehors des aspects énumératifs des 2-arbres, Leclerc et Makarenkov [LEC 98] donne une correspondance entre les fonctions d'arbres et les 2-arbres dans le contexte des métriques d'arbres et des dissimilarités d'arbres (voir [BAR 88, BAR 91]).

Dans [TOD 89, LEC 02], les graphes de k-arches sont définis comme des graphes maximaux kd-acycliques, où un graphe est dit kd-acyclique, s'il ne contient pas de kd-cycle (voir [LEC 02], définition 3.1). Todd montre que cette définition est équivalente à celle récursive donné plus haut. Dans [LEC 02], Leclerc utilise des graphes de k-arches valués aux arêtes pour coder des distances ou fonctions d'arbres (voir [BAR 88, BAR 91]).

Rappelons que le nombre de k-arbres sur n sommets étiquetés est donné par ([BEI 69b, MOO 69])

$$a_n^k = \binom{n}{k} (k(n-k)+1)^{n-k-2}. \qquad [1]$$

Lorsque $k = 1$, on retrouve la formule bien connu de Cayley $a_n = n^{n-2}$ comptant le nombre d'arbres (de Cayley) étiquetés.

Le but de ce travail est d'obtenir le nombre de graphes de k-arches étiquetés aux sommets. Nous établissons que ce nombre est

$$\binom{n}{k}^{n-k-1}.$$

Pour ce faire, nous proposons en section 2 une bijection entre les graphes de k-arches sur n sommets et des mots de longueur $n-k-1$ dont les lettres sont des sous-ensembles de k éléments de l'ensemble des étiquettes des sommets. Cette bijection généralise le code de Prüfer pour les arbres ([PRU 18]).

2. Le nombre de graphes de k-arches étiquetés

Dans cette section, nous proposons une formule donnant le nombre de graphes de k-arches étiquetés, pour $k \geq 1$. Nous établissons cette formule en utilisant un argument bijectif basé sur une généralisation du code de Prüfer pour les arbres étiquetés aux sommets ([PRU 18]), qui a été généralisé aux k-arbres ([RÉN 69]).

Une *feuille* d'un graphe de k-arches est un sommet de degré k. Par exemple, le graphe de k-arches de la figure 1 b) possède quatre feuilles, respectivement étiquetée 2, 3, 4 et 9. Cette définition est légitime puisque, lorsque $n = 1$, un sommet de degré un dans un arbre est une feuille, au sens commun de la théorie des graphes.

Proposition 1 *Soit $k \geq 1$ un entier fixé. Alors, le nombre $\mathcal{G}_n^k$ de graphes de k-arches sur n sommets étiquetés, pour $n > k$, est donné par*

$$\mathcal{G}_n^k = \binom{n}{k}^{n-k-1} \qquad \text{et} \qquad \mathcal{G}_k^k = 1. \tag{2}$$

Preuve. Nous construisons une bijection entre les graphes de k-arches à n sommets étiquetés et les mots $w = w_1 w_2 \dots w_{n-k-1}$ de longueur $n-k-1$, où chaque w_i, $i = 1, 2, \dots, n-k-1$, est une k-lettre (valide) de la forme suivante (vecteur colonne) :

$$(v_{i_1}, v_{i_2}, \dots, v_{i_k})^T \tag{3}$$

telle que

1. pour tout $1 \leq j \leq k$, $1 \leq i_j \leq n$;
2. pour tout $1 \leq j \leq n$, v_j est un sommet du graphe de k-arches ;
3. $i_1 < i_2 < \dots < i_k$.

De tels mots sont appelés *valides*. Nous supposons implicitement que les étiquettes sont ordonnées, *i.e.*, $v_1 < v_2 < \ldots < v_n$. La bijection fonctionne par effeuillage.

Soit g un graphe de k-arches à n sommets dont l'ensemble des sommets est $V = \{v_1, v_2, \dots, v_n\}$. A la première étape, on enlève la feuille de g ayant la plus petite étiquette ainsi que ses arêtes incidentes et on forme une k-lettre avec ses sommets adjacents en les ordonnant en ordre croissant. Le fait que le degré de chaque feuille soit k assure que l'on forme une unique k-lettre valide. Après la première étape, on obtient un graphe de k-arches à $k-1$ sommets et on répète la première étape. En répétant $n-k-1$ fois cette étape, on obtient un mot valide de longueur $n-k-1$. Notons qu'après la dernière étape, le graphe de k-arches g devient un seul $k-1$-simplexe, *i.e.*, un graphe complet à k sommets. Par exemple, si on applique cette construction au graphe de k-arches de la figure 1 b), il vient le mot valide de longueur 9 suivant :

$$\begin{pmatrix} 1 & 7 & 1 & 8 & 7 & 5 & 8 & 8 & 10 \\ 6 & 11 & 6 & 11 & 11 & 12 & 10 & 7 & 11 \end{pmatrix}. \tag{4}$$

Réciproquement, étant donné un mot valide $w = w_1 w_2 \dots w_{n-k-1}$ de longueur $n-k-1$, on désire construire un graphe de k-arches à n sommets étiquetés. Outre le mot w, on utilise un sous-ensemble dynamique L de l'ensemble des sommets V, c'est-à-dire, un sous-ensemble dont les éléments et la taille évoluent. Au départ, les éléments de L sont les sommets n'apparaissant pas dans w et on étend w en accolant une copie de la dernière k-lettre, w_{n-k-1}, à la fin de w. Ce mot étendu est à nouveau noté w

$$w := w || w_{n-k-1} = w_1 w_2 \dots w_{n-k-1} w_{n-k-1}.$$

Dans le cas du graphe de k-arches de la figure 1 b), on obtient $L = \{2, 3, 4, 9\}$ et w devient

$$\begin{pmatrix} 1 & 7 & 1 & 8 & 7 & 5 & 8 & 8 & 10 & 10 \\ 6 & 11 & 6 & 11 & 11 & 12 & 10 & 7 & 11 & 11 \end{pmatrix}.$$

On enlève le plus petit élément de L et on le connecte à tous les éléments de la première k-lettre (w_1) de w. On supprime alors la lettre w_1 de w et on appelle toujours w ce mot réduit. On met à jour L en ajoutant chaque sommet $v_i \in w_1$ n'apparaissant pas dans les lettres restantes de w. On répète alors l'étape de récurrence précédente avec le mot $w = w_2 \dots w_{n-k-1} w_{n-k-1}$, et ainsi de suite jusqu'à ce que l'on atteigne le mot vide, en gardant à chaque étape les composantes connexes créées. Pour terminer la transformation réciproque, il suffit de connecter les k sommets de la dernière k-lettre w_{n-k-1} de w.

Il nous faut vérifier que nous obtenons un graphe connexe qui est bien un graphe de k-arches correspondant au mot w lorsque l'algorithme d'effeuillage est appliqué. Ceci est assez direct en utilisant un argument récursif sur la longueur du mot. La preuve est identique au cas des arbres (voir [MOO 70, SW 86]) et les détails sont laissés au lecteur. ■

La figure 2 montre un exemple complet de la transformation réciproque pour un graphe d'arches comportant six sommets.

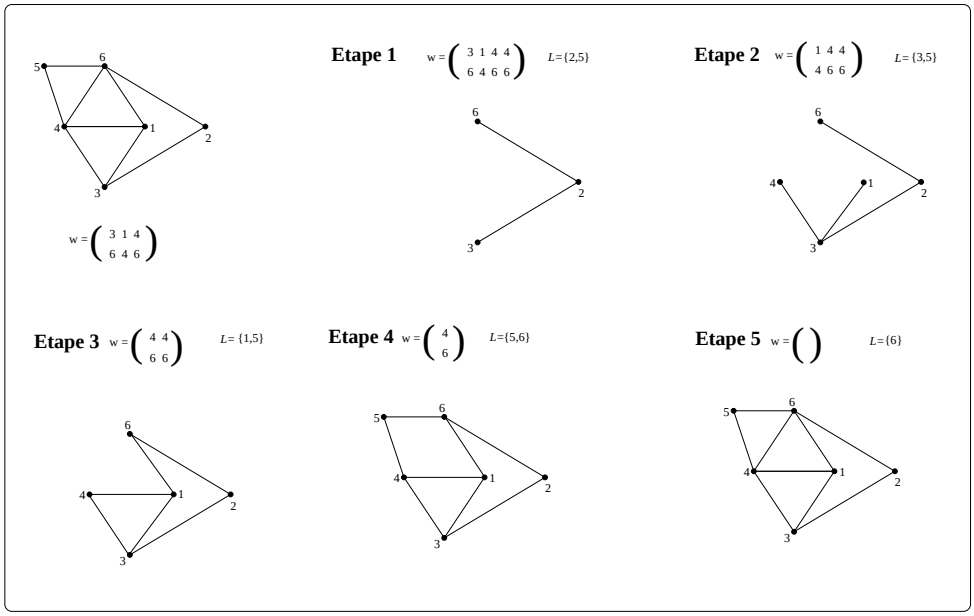


FIG. 2. Illustration de la transformation réciproque.

Remarque 1 *Il est intéressant de noter que, pour $k = 1$, la formule* [2] *reste valable. En effet, cette formule coïncide alors avec la formule de Cayley ($a_n = n^{n-2}$) énumérant les arbres sur n sommets étiquetés.*

Remarque 2 *La transformation réciproque de la preuve de la proposition 1 induit naturellement un algorithme de génération aléatoire des graphes de k-arches au moyen des mots valides.*

3. Remerciements

Nous remercions grandement Bruno Leclerc pour sa présentation des graphes de k-arches lors d'un séminaire à Montréal (Séminaire du LaCIM, Université du Québec à Montréal) et Pierre Leroux pour des discussions très utiles.

4. Bibliographie

[BAR 88] J. P. BARTHÉLEMY et A. GUÉNOCHE, *Les Arbres et les Représentations des Proximités*, Masson, 1988.

[BAR 91] J. P. BARTHÉLEMY and A. GUÉNOCHE, *Trees and Proximity Representations*, Wiley-Interscience Series in Discrete Mathematics and Optimization, 1991.

[BEI 69a] L. W. BEINEKE and J. W. MOON, "Several proofs of the number of labeled 2-dimensional trees", dans F. Harary ed., *Proof Techniques in Graph Theory*, Academic Press, 1969, p. 11–20.

[BEI 69b] L. W. BEINEKE and R. PIPPERT, "The number of labeled k-dimensional trees", *J. Combin. Theory* **6**, 1969, p. 200–205.

[BOU] M. BOUSQUET and C. LAMATHE, "Enumeration of solid 2-trees according to edge number and edge degree distribution", à paraître dans *Discrete Math.*.

[FOW 02] T. FOWLER, I. GESSEL, G. LABELLE, P. LEROUX, "The specification of 2-trees", *Adv. in Appl. Math.* **28**, 2002, p. 145–168.

[HAR 73] F. HARARY and E. PALMER, *Graphical Enumeration*, Academic Press, 1973.

[LAB 03] G. LABELLE, C. LAMATHE and P. LEROUX, "A classification of plane and planar 2-trees", *Theoret. Comput. Sci.* **307**, 2003, p. 337–363.

[LAB 04] G. LABELLE, C. LAMATHE and P. LEROUX, "Labelled and unlabelled enumeration of k-gonal 2-trees", *J. Combin. Theory Ser. A* **106**, 2004, p. 193–219.

[LAM 04] C. LAMATHE, *Spécification de classes de structures arborescentes*, thèse de doctorat, Publications du LaCIM, vol. **33**, 2004.

[LEC 02] B. LECLERC, "Graphes d'arches", *Math. & Sci. humaines* **157**, 2002, p. 27–48.

[LEC 98] B. LECLERC and V. MAKARENKOV, "On some relations between 2-trees and tree metrics", *Discrete Math.* **192**, 1998, p. 223–249.

[MOO 69] J. W. MOON, "The number of labeled k-trees", *J. Combin. Theory* **6**, 1969, p. 196–199.

[MOO 70] J. W. MOON, *Counting labelled trees*, Canadian mathematical monographs, 1970.

[PAL 69] E. PALMER, "On the number of labeled 2-trees", *J. Combin. Theory* **6**, 1969, p. 206–207.

[PRU 18] H. PRÜFER, "Neuer Beweis eines Satzes über Permutationen", *Arch. Math. Phys.* **27**, 1918, p. 742–744.

[RÉN 69] C. RÉNYI and A. RÉNYI, "The Prüfer code for k-trees", dans *Combinatoral theory and it's applications* (Proc. Colloq. Balatonfüred, 1969), North-Holland, 1970, p. 945–971.

[SW 86] D. STANTON and D. WHITE, "Constructive combinatorics", *Springer-Verlag*, 1986.

[TOD 89] P. TODD, "A k-tree generalization that characterizes consistency of dimensioned engineering drawings", *SIAM J. Disc. Math.* **2**, 1989, p. 255–261.

Analyse textuelle des éléments plaisants et déplaisants de visages de femmes caucasiennes cités par un groupe de juges naïfs

J. Latreille[1], L. Ambroisine[1], S. Guéhenneux[1], G. Coudin[2], R. Jdid[1,3], S. Gardinier[1], E. Mauger[1], F. Morizot[1], E. Tschachler[1,4], C. Guinot[1]

[1] *CE.R.I.E.S, Neuilly sur Seine, France*
[2] *Institut de Psychologie, Laboratoire de Psychologie Sociale, Université Paris V, Boulogne, France*
[3] *Département de Dermatologie, Hôpital Tarnier, Paris, France*
[4] *Département de Dermatologie, Université médicale de Vienne, Vienne, Autriche*

RÉSUMÉ. *Lors d'une étude sur le vieillissement de la peau, il a été demandé à 194 juges, 100 hommes et 94 femmes, de relever des éléments plaisants et déplaisants à partir de 60 photographies de visages de femmes âgées de 30 à 65 ans. L'analyse textuelle des éléments cités par l'ensemble des juges a mis en évidence trois groupes de juges utilisant un vocabulaire identique pour décrire les visages. Une recherche de typologie des photographies de visages réalisée à partir des éléments cités a conduit à identifier quatre groupes de photographies de visages présentant des similitudes en terme d'éléments cités.*

MOTS-CLÉS : algorithme K-means, analyse factorielle des correspondances, analyse textuelle, classification ascendante hiérarchique.

1 Introduction

Différentes études ont exploré les aspects du visage pouvant être liés à la perception de l'âge (morphologie, symétrie, aspect global de la peau), mais peu de recherches se sont intéressées aux liens entre des signes définis du vieillissement cutané et la perception de l'âge [MAR 80, BUR 95]. Dans ce contexte, une étude a été menée en 2003 pour analyser l'impact de la présence et de l'intensité de signes de vieillissement cutané sur l'âge perçu et l'attirance. Il a été demandé à un panel de juges naïfs, composé d'hommes et de femmes, de donner un âge et une note d'attirance, et de relever les éléments plaisants et déplaisants de 60 photographies de visages de femmes caucasiennes. Les objectifs de cette analyse préliminaire sont, d'une part, d'identifier des groupes de juges utilisant un même vocabulaire pour décrire les éléments plaisants et déplaisants, et d'autre part, d'identifier des groupes de photographies de visages pour lesquels les mêmes éléments ont été cités par les juges. Le genre des juges pouvant éventuellement influencer les éléments relevés, la seconde problématique a été réalisée pour chaque genre séparément.

2 Matériel

L'âge des femmes photographiées était compris entre 30 et 65 ans. Les femmes étaient non maquillées, les yeux clos, avec un bandeau sur les cheveux pour dégager leur visage. De plus, tous les accessoires pouvant influencer le jugement tels que foulards et bijoux, avaient été effacés des photographies au préalable (Adobe Photoshop® version 7.0.1). Par ailleurs, nous avions à notre disposition différentes informations concernant ces femmes dont leur âge chronologique, dix paramètres décrivant la

morphologie de leur visage, et la présence et l'intensité de vingt-deux signes de vieillissement de la peau de leur visage appréciés par un dermatologue grâce à des échelles ordinales de gravité.

L'étude s'est déroulée en une série de sessions, pour des raisons pratiques un maximum de huit juges participaient à chaque session. Le panel de juges était composé de 100 hommes et 94 femmes, âgés entre 30 et 65 ans. Chaque juge était isolé dans un box où les photographies numériques défilaient les unes après les autres sur l'écran d'un moniteur. Les juges avaient environ une minute par photographie pour évaluer l'âge, donner une note d'attirance et relever des éléments plaisants et déplaisants du visage sur un questionnaire papier. Les juges devaient également indiquer dans le questionnaire papier leur genre, leur âge, leur catégorie socio-professionnelle, ainsi que leur niveau d'études.

3 Méthodes

3.1 Génération du vocabulaire

La plupart du vocabulaire utilisé pour décrire les éléments plaisants a été également utilisé pour décrire les éléments déplaisants, par exemple « la peau », « le teint »… Pour certaines photographies certains juges n'ont noté aucun élément plaisant et/ou déplaisant. De ce fait, 8243 réponses pour les éléments plaisants (définissant la variable textuelle **plait)** et 9558 réponses pour les éléments déplaisants (variable textuelle **déplait)** ont servi à générer le vocabulaire initial (logiciel SPAD® version 5.6). La procédure MOTS a permis de créer le vocabulaire initial de « mots », un « mot » étant défini comme une suite de caractères délimitée par un ensemble de caractères séparateurs (en analyse textuelle des données, on parlerait plutôt de « forme graphique » [LEB 88]). L'outil interactif de modification du vocabulaire a ensuite été utilisé afin de supprimer, corriger et mettre en équivalence certains « mots ». Après élimination, correction et regroupement des « mots », les modifications ont été enregistrées dans un dictionnaire qui a ensuite été appliqué sur les variables textuelles **plait** et **déplait** traitées séparément. La procédure SEGME a ensuite permis de rechercher les « segments répétés » sur chaque variable textuelle, un « segment répété » étant défini comme une suite de séquences de « mots » non séparée par un séparateur de séquence et dont la fréquence est supérieure ou égale à deux. Enfin, un tableau de contingence, avec en ligne l'identifiant « juges_photographies » (194 x 60 = 11640 lignes) et en colonne les « mots et segments répétés » a été créé pour chaque variable textuelle (procédure TEXNU).

3.2 Typologie de juges

Pour rechercher d'éventuels groupes de juges ayant utilisé le même vocabulaire pour décrire les photographies, une analyse factorielle des correspondances (AFC) a été d'abord réalisée à partir d'un tableau de contingence « juges » x « mots et segments répétés », les mots et segments répétés résultant des variables **plait** et **déplait** (procédure CORBI). Une classification ascendante hiérarchique (méthode de Ward) [EVE 93] a ensuite été effectuée sur les axes factoriels conservés lors de l'AFC (procédure CLUSTER de SAS®, version 8.2). Le nombre de classes le plus plausible a été déterminé grâce aux critères du pseudo F et du pseudo t^2, puis les juges ont été affectés dans les classes selon leur proximité (algorithme K-means de McQueen, procédure FASTCLUS de SAS®). Les classes obtenues ont finalement été décrites à l'aide des informations disponibles sur les juges (procédures FREQ et GLM de SAS®).

3.3 Typologie de visages

La recherche de groupes de photographies de visages présentant des caractéristiques communes d'après les éléments plaisants et déplaisants relevés par les juges a été réalisée pour les hommes-juges et les femmes-juges séparément. Deux tableaux de contingence ont ainsi été créés, un premier tableau « photographies » x « mots et segments répétés » cités par les femmes-juges, et un second tableau « photographies » x « mots et segments répétés » cités par les hommes-juges. Une AFC a ensuite été réalisée sur chacun des tableaux. Les mêmes méthodes de classification que celles précédemment décrites

ont été utilisées pour obtenir une typologie de photographies de visages à partir des éléments cités. Les classes obtenues ont été ensuite décrites à l'aide des informations disponibles sur les visages des femmes photographiées (procédures FREQ et GLM de SAS®).

4 Résultats

4.1 Génération du vocabulaire

Le vocabulaire initial a permis l'identification de 3233 « mots », puis la génération du dictionnaire. Ce dictionnaire a ensuite été appliqué sur les variables textuelles **plait** et **déplait** traitées séparément. Cent dix-sept « mots et segments répétés » ont été conservés pour la variable **plait,** et 131 « mots et segments répétés » pour la variable **déplait**. Dans la suite du papier, les « mots » et « segments répétés » utilisés pour décrire les éléments plaisants seront indiqués en *italique* et les éléments déplaisants en souligné.

4.2 Typologie des juges

Une AFC a été réalisée sur le tableau de contingence « juges » x « mots et segments répétés » issus de la variable **plait** et de la variable **déplait**, 194 lignes x (117+131) colonnes. Les vingt-cinq « mots et segments répétés » qui avaient été cités par moins de cinq juges ont été utilisés en variables illustratives dans l'analyse. Deux axes factoriels ont été conservés. Le premier axe oppose les juges ayant noté des éléments liés à la morphologie du visage (nez, lèvres-bouche, menton, *nez*, *lèvres-bouche*…) aux juges ayant noté des éléments liés à l'aspect de la peau du visage (*peau lisse*, peau grasse, *peau nette*, peau rouge, affaissée…). Le deuxième axe oppose les juges ayant noté des éléments liés à l'interprétation de l'expression du visage (*aimable*, *apaisé*, pas aimable, triste, visage agressif…) aux juges ayant noté des éléments liés à l'aspect de la peau du visage (*peau lisse*, tache brune, ridée, peau grasse…). Trois groupes de juges ont été ensuite identifiés grâce à la classification effectuée sur les deux premiers axes factoriels. Le premier groupe (n=23) a tendance à utiliser un vocabulaire lié à la morphologie du visage, le deuxième groupe (n=76) un vocabulaire lié à la fois à l'expression du visage et à l'aspect de la peau du visage, et le troisième groupe (n=95) un vocabulaire lié à l'aspect de la peau du visage. Un lien significatif du vocabulaire utilisé avec le genre des juges a été trouvé ($p < 0{,}0001$) ; en effet le premier et le deuxième groupe sont composés majoritairement d'hommes (78% et 62%, respectivement) alors que le troisième groupe est composé majoritairement de femmes (63%). Une tendance a également été trouvée avec l'âge, le premier groupe ayant tendance à être plus jeune que le deuxième groupe (moyenne ± écart à la moyenne : 44±2 ans versus 50±1 ans, $p < 0{,}0367$).

4.3 Typologie de visages

La recherche de typologie de photographies de visages réalisée à partir des éléments cités par les hommes-juges d'une part, et les femmes-juges d'autre part, a conduit à des résultats similaires. De ce fait, une nouvelle recherche de typologie de visages a été effectuée à partir des éléments cités par l'ensemble des juges. L'AFC a été réalisée sur le tableau de contingence « photographies » x « mots et segments répétés » cités par les juges, 60 lignes x (117+131) colonnes. Les vingt-cinq « mots et segments répétés » qui avaient été cités par moins de cinq juges ont été utilisés en variables illustratives dans l'analyse. Trois axes factoriels ont été conservés. Les « mots et segments » ont été représentés graphiquement sur le premier plan factoriel engendré par les axes factoriels 1 et 2 (cf. figure 1.a) et sur le deuxième plan factoriel engendré par les axes factoriels 1 et 3 (cf. figure 1.b). Les « mots et segments » les moins contributifs à la construction des axes ont été masqués. Le premier axe oppose les photographies de visages décrites par des mots et segments liés au vieillissement de la peau (rides, affaissement, taches brunes, taches de vieillesse…), aux photographies décrites par des mots et segments liés à des problèmes de peau (peau grasse, pores dilatés, bouton, peau rouge…). Le deuxième axe oppose les photographies de visages décrites par des mots et segments liés à des descripteurs d'une « belle » peau mais également à la pilosité (*peau*, *teint claire*, *peau lisse*, *peau nette*, duvet lèvres, sourcils fournis…), aux photographies décrites par des mots et segments liés, là encore, à des problèmes de peau (bouton, peau rouge, peau grasse,

rougeurs…). Enfin, le troisième axe est caractérisé par la présence d'irrégularités pigmentaires (taches brunes, taches, taches de rousseur, grain de beauté…). Quatre groupes de photographies de visages ont été ensuite identifiés grâce à la classification effectuée sur les trois premiers axes factoriels (cf. figures 1.a et 1.b). Le premier groupe (n=16) est caractérisé par des descripteurs associés à une belle peau et à la pilosité, le deuxième (n=13) par des problèmes de peau, le troisième (n=14) par des irrégularités pigmentaires, et enfin le quatrième (n=17) par des descripteurs du vieillissement cutané.

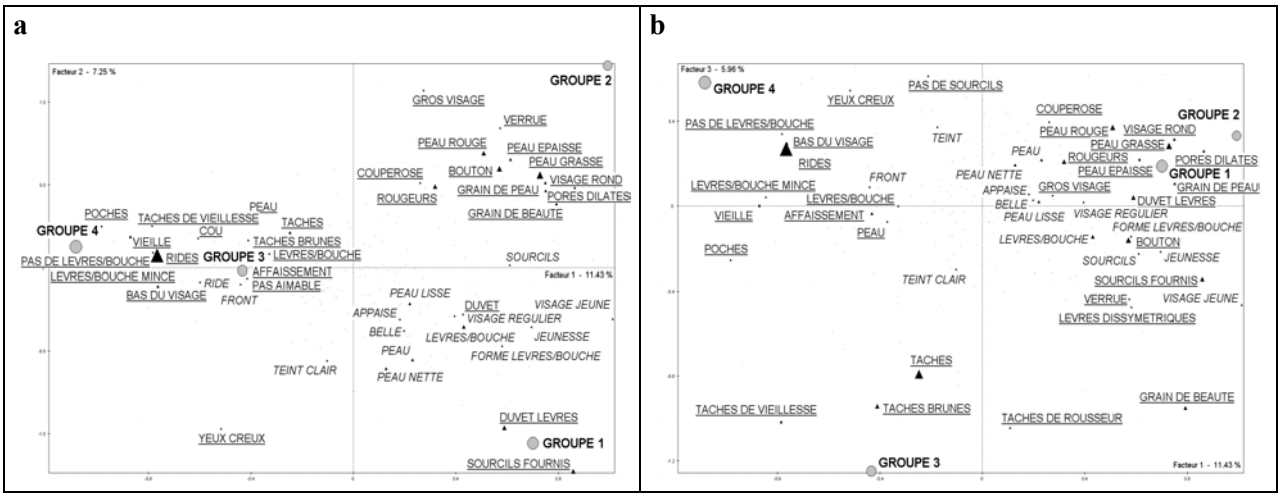


Figure 1 : représentation des éléments qui *PLAISENT* et qui DEPLAISENT et des centres de gravités ◯ des groupes de photographies de visages a) sur le plan factoriel 1-2 et b) sur le plan factoriel 1-3

5 Discussion

Après une étape méticuleuse de traitement de l'information textuelle, qui a été toutefois facilitée grâce aux connaissances des experts, cette recherche a mis en évidence des groupes de juges utilisant un vocabulaire identique pour décrire les éléments plaisants et déplaisants de visages de femmes. D'autre part, un consensus entre le jugement des hommes et des femmes a été également mis en évidence. Puis, la recherche de typologie de visages effectuée à partir des éléments cités par l'ensemble des juges a permis d'obtenir quatre groupes de photographies de visages. Les deux typologies seront utilisées dans la suite des travaux pour étudier les liens entre les caractéristiques des juges, celles des visages des femmes, la perception de l'âge des femmes, ainsi que la note d'attirance donnée par les juges.

6 Bibliographie

[BUR 95] BURT D.M., PERRETT D.I., " Perception of age in adult Caucasian male faces: computer graphic manipulation of shape and colour information ", *Proceedings of the Royal Society of London, Series B, Biological Sciences,* 1995, 259, 137-143.

[EVE 93] EVERITT B.S., *Cluster analysis,* Arnold, 1993.

[LEB 95] LEBART L., MORINEAU A., PIRON M., *Statistique exploratoire multidimensionnelle*, DUNOD, 1995.

[LEB 88] LEBART L., SALEM A., *Analyse statistique des données textuelles*, DUNOD, 1988.

[MAR 80] MARK L.S., PITTENGER J.B., HINES H., CARELLO C., SHAW R.E., TODD J.T., " Wrinkling and head shape as coordinated sources of age-level information ", *Perception & Psychophysics*, 1980, 27, 117-124.

Clustering via la programmation DC pour la détermination d'arbre hiérarchique de multidiffusion

LE THI Hoai An[1], **LE Hoai Minh**[2], **PHAM DINH Tao**[3]

[1,2]*Laboratory of Theoretical and Applied Computer Science (LITA).*
UFR MIM, Metz University, Ile de Saulcy, 57045 Metz, France.

[3]*Laboratory of Mathematics (LMI), National Institute for Applied Sciences - Rouen*
BP 08, Place Emile Blondel F 76131 Mont Saint Aignan Cedex, France.

RESUME *La multidiffusion(multicast) offre à la communication de groupe un gain considérable d'efficacité, en particulier pour des grands groupes. Les hiérarchies aident à réduire la complexité, le nombre de messages échangés entre les participants. Une des approches est d'utiliser les algorithmes de clustering pour déterminer l'arbre hiérarchique de multicast. Les algorithmes clustering aident à diviser des populations d'utilisateur selon une variété de critères. Ce problème peut se formuler sous la forme d'un problème d'optimisation non convexe, non différentiable pour lequel nous introduisons, dans cet article, une nouvelle approche basée sur la programmation DC (Difference of Convex functions) et DCA (DC Algorithms).*

MOTS CLES *Programmation DC, DCA, Multidiffusion, Arbre hiérarchique, Classification, K-means.*

1 Introduction

La communication multicast joue de plus en plus un rôle important de nos jours. Le but principal de la communication multicast est de fournir un service de communication entre les participants du groupe tout en assurant une bonne qualité de service et minimisant le coût total de communication. Citons ici quelques applications de la communication multicast : Systèmes de Groupware pour faciliter la conception et l'évaluation de collaboration, Vidéo conférence, Diffusion en temps réel des événements internationaux, Peer-to-peer applications pour le partage des données ou traitement. Une des approches de communication multicast est d'utiliser *Steiner Trees*. Cette approche a été introduite dans les années 80, plusieurs discussions sur ce sujet se trouvent dans [5], [8]. Dans ce papier nous traitons ce problème par une technique d'optimisation non convexe, non différentiable basée sur la programmation DC (Difference of Convex functions) et DCA (DC optimisation Algorithms). La programmation DC et DCA ont été introduits par T. PHAM DINH en 1986 et intensivement développés par H.A. LE THI et T. PHAM DINH depuis 1993 ([3], [4], [6], [7] et références incluses) pour devenir maintenant classiques et de plus en plus populaires. Ils ont été appliqués avec grand succès à nombreux problèmes d'optimisation non convexe différentiable ou non de grande dimension dans différents domaines des sciences appliquées, en particulier aux problèmes de classification en analyse de données et data mining ([1], [3], [9], [10]) auxquels ils fournissent souvent des solutions globales et sont plus performants que des approches standard (voir [3], [4], [6], [7] et références incluses).

2 Formulation du problème

Un arbre hiéararchique de multidiffusion contient une source (le centre total) et plusieur niveaux de hiérarchie. Un noeud est relié à un noeud de niveau hiérarchique supérieur (sauf la source) et ses noeuds de niveau hiérarchique inférieur (s'il en existe).

La communication multidiffusion fonctionne de façon suivante : le noeud central total envoie le message aux autres centres de chaque sous-ensemble (centres de niveau 1), et à son tour, chaque centre retransmet le message reçu aux autres noeuds de niveau hiérarchique inférieur dans son sous-ensemble. Ce processus

se répète jusqu'à ce que le message arrive aux noeuds termimaux (les noeuds les plus bas dans l'arbre hiérarchique).
Dans cet article, nous considérons seulement un arbre hiérarchique de communication de deux niveaux.

Considérons un ensemble de points $\mathcal{A} := \{a_j \in \mathbb{R}^n : j = 1, ..., p\}$. Nous devons regrouper ces points en **m** sous-ensembles (cluster). Pour chaque sous-ensemble $\mathcal{A}_i$, nous choisissons un centre (centre de niveau 1) y_i ($y_i \in \mathcal{A}$) qui est relie à tous les autres noeuds dans ce sous-ensemble. Le centre total y^* ($y^* \in \mathcal{A}$) de l'arbre hiérarchique (celui qui est relié aux noeuds y_i) est choisi en fonction de centres y_i. Le coût total de l'arbre est défini par :

$$C(\mathcal{A}_1, ..., \mathcal{A}_m, y_1, ..., y_m, y^*) = \sum_{i=1}^{m} \sum_{a \in \mathcal{A}_i, a \neq y^*} \|a - y_i\|^2 + \sum_{i=1}^{m} \|y_i - y^*\|^2.$$

Le premier terme définit la somme des coûts des sous-ensembles (le coût d'envoi de chaque centre de niveau 1 aux noeuds dans son sous-ensemble). Nous remarquons que le noeud centre total y^* appartient à un des sous-ensembles donc nous ajoutons la condition $a \neq y^*$ pour éviter de compter deux fois le coût d'envoi du noeud central total y^* aux centres de niveau 1 du sous-ensemble qui contient y^*.
Le deuxième terme définit le coût d'envoi du noeud central total aux centres de niveau 1.
Le but est de trouver un arbre hiérarchique qui minimise le coût total.
Cette formulation n'est pas convenable pour appliquer directement les techniques d'optimisation. Nous allons reformuler le problème grâce à l'utilisation des centres artificiels [5]. Considérons un ensemble des centres artificiels $\{x_i : i = 1..m\}$ dans $\mathbb{R}^n$. Le centre total x^* est calculé à partir des centres x_i par la formulation suivante : $x^* = \frac{1}{m} \sum_{i=1}^{m} x_i$. Les clusters $\mathcal{A}_i$ sont définis par : $\mathcal{A}_i = \{a \in \mathcal{A} : \|x_i - a\| < \min_{j=1..m, j \neq i} \|x_j - a\|\}$.
Le coût total de l'arbre, avec les centres aritificiels, s'écrit comme :

$$\tilde{C}(x_1, ..., x_m) = \sum_{j=1}^{p} \min_{i=1..m} \|x_i - a_j\|^2 + \sum_{i=1}^{m} \|x_i - \frac{1}{m} \sum_{l=1}^{m} x_l\|^2.$$

Pour retrouver les centres réels, la façon la plus naturelle est de prendre les noeuds qui sont les plus proches des centres artificiels. Nous imposons donc les contraintes $\sum_{i=1}^{m} \min_{j=1..p} \|x_i - a_j\|^2 = 0$ (si cette condition vérifie, les noeuds artificiels deviennent les noeuds réels) et obtenons finalement la formulation suivante :

$$(P) \quad \min \left\{ \frac{1}{2} \sum_{j=1}^{p} \min_{i=1..m} \|x_i - a_j\|^2 + \frac{1}{2} \sum_{i=1}^{m} \|x_i - \frac{1}{m} \sum_{l=1}^{m} x_l\|^2 : \sum_{i=1}^{m} \min_{j=1..p} \|x_i - a_j\|^2 = 0 \right\}.$$

C'est un problème d'optimisation non convexe et non différentiable très difficile à traiter efficacement par les approches standard. Il est doublement non convexe en sa fonction objectif et sa contrainte. En pénalisant la dernière, on obtient le problème (P_t) suivant : ($\tau > 0$)

$$(\mathrm{P}_\tau) \quad \min \left\{ \tfrac{1}{2} \sum_{j=1}^{p} \min_{i=1..m} \|x_i - a_j\|^2 + \tfrac{1}{2} \sum_{i=1}^{m} \|x_i - \tfrac{1}{m} \sum_{l=1}^{m} x_l\|^2 + \tfrac{\tau}{2} \sum_{i=1}^{m} \min_{j=1..p} \|x_i - a_j\|^2 : x_i \in \mathbb{R}^n \right\}.$$

On va démontrer que (P_t) est un programme DC en mettant en évidence une décomposition DC de sa fonction objectif qui semble bien être adaptée au problème. Mais avant cela, on présente brièvement ci-dessous les grandes lignes de la programmation DC et DCA.

3 Résolution du problème (P_t) par DCA

La programmation DC joue un rôle central en optimisation non convexe et optimisation globale car la quasi totalité des problèmes d'optimisation de la vie courante est de nature DC. Elle connaît des développements spectaculaires au cours de cette dernière décennie. DCA est une méthode de descente (primale-duale) pour la résolution d'un programme DC, qui est la minimisation d'une fonction DC de la forme (les contraintes convexes peuvent être incorporées à la fonction objectif à l'aide de la fonction indicatrice) :

$$\alpha := \inf\{f(x) := g(x) - h(x) \ : \ x \in \mathbb{R}^n\}, \tag{1}$$

où g, h sont les fonctions convexes semi-continues et propres sur $\mathbb{R}^n$. Une telle fonction f est appelée fonction DC et les fonctions convexes g et h, composantes DC de f. Une fonction DC admet une infinité de décomposition DC. La dualité DC est définie via la conjugaison de fonction convexe (la conjuguée de g, notée g^* : $g^*(y) := \sup\{\langle x, y\rangle - g(x) \ : \ x \in \mathbb{R}^n\}$ et le programme dual de (1) est donné par

$$\alpha := \inf\{h^*(y) - g^*(y) \ : \ y \in \mathbb{R}^n\}, \tag{2}$$

(l'espace dual de $\mathbb{R}^n$ est identifié à lui-même).

Basé sur les conditions d'optimalité locale et la dualité DC, DCA consiste en la construction de deux suites $\{x^k\}$ et $\{y^k\}$, candidats respectifs aux solutions des problèmes primal et dual que l'on améliore à chaque itération (les deux suites $\{g(x^k) - h(x^k)\}$ et $\{h^*(y^k) - g^*(y^k)\}$ sont décroissantes) et qui convergent vers des solutions primale et duale x^* et y^*verifiant des conditions d'optimalité locale et

$$x^* \in \partial g^*(y^*), \ y^* \in \partial h(x^*). \tag{3}$$

Cette relation (3) implique que x^* est une solution optimale du programme convexe

$$\inf\{f(x) + h(x) - [h(x^*) + \langle x - x^*\rangle] : x \in \mathbb{R}^n\}. \tag{4}$$

Le schéma général de DCA prend la forme :

$$y^k \in \partial h(x^k); \ x^{k+1} \in \partial g^*(y^k). \tag{5}$$

La première interprétation de DCA est simple : à chaque itération on remplace dans le programme DC primal la deuxième composante DC h par sa minorante affine $h_k(x) := h(x^k) + \langle x - x^k, y^k\rangle$ au voisinage de x^k pour obtenir le programme convexe suivant

$$\inf\{g(x) - h_k(x) \ : \ x \in \mathbb{R}^n\} \tag{6}$$

dont l'ensemble des solutions optimales n'est autre que $\partial g^*(y^k)$.
De manière analogue, la deuxième composante DC g^* du programme DC dual (2) est remplacée par sa minorante affine $(g^*)_k(y) := g^*(y^k)+\langle y-y^k, x^{k+1}\rangle$ au voisinage de y^k pour donner naissance au programme convexe

$$\inf\{h^*(y) - (g^*)_k(y) \ : \ y \in \mathbb{R}^n\} \tag{7}$$

dont $\partial h(x^{k+1})$ est l'ensemble des solutions optimales. DCA opère ainsi une double linéarization à l'aide des sous-gradients de h et g^*. Il est à noter que DCA travaille avec les composantes DC g et h et non pas avec la fonction f elle-même. Chaque décomposition DC de f donne naissance à un DCA. Pour un programme DC donné, la question de décomposition DC optimale reste ouverte, en pratique on cherche des décompositions DC bien adaptées à la structure spécifiques du programme DC étudié pour lesquelles les suites $\{x^k\}$ et $\{y^k\}$ sont faciles à calculer, si possible explicites pour que les DCA correspondants soient moins coûteux en temps et par conséquent capables de supporter de très grandes dimensions.

Pour une étude complète de la programmation DC et DCA, se reporter aux (voir [3], [4], [6], [7] et références incluses). Le traitetement d'un programme non convexe par une approche DC et DCA devrait comporter donc deux tâches : la recherche d'une décomposition DC adéquate et celle d'un bon point initial.

3.1 Décomposition DC pour (P_t)

Pour la simplicité des calculs, on va reformuler le problème (P_τ) comme un programme DC dans un espace matriciel approprié. La variable sera une matrice X de type $m \times n$ dont la $i^{\text{ème}}$ ligne X_i est égale à x_i. L'espace de travail est donc l'espace $\mathcal{M}_{m,n}(\mathbb{R})$ de matrices à coefficients réels ayant m lignes et n colonnes. Sa structure hilbertienne est défine à l'aide du produit scalaire usuel

$$\mathcal{M}_{m,n}(\mathbb{R}) \ni X \leftrightarrow (X_1, X_2, \ldots, X_m) \in (\mathbb{R}^n)^m, \ X_i \in \mathbb{R}^n, (i = 1, .., m), \ \langle X, Y\rangle := \sum_{i=1}^{m} \langle X_i, Y_i\rangle = \sum_{i=1}^{m} \|X_i\|^2$$

et sa norme hilbertienne $\|X\|^2 := \sum_{i=1}^m \langle X_i, X_i\rangle = \sum_{i=1}^m \|X_i\|^2$.

A l'aide de la propriété $\min\limits_{i=1..m} \|x_i - a_j\|^2 := \sum\limits_{i=1}^m \|x_i - a_j\|^2 - \max\limits_{i=1..m} \sum\limits_{r=1,r\neq i}^m \|x_r - a_j\|^2$ la fonction objectif de (P_t) peut s'écrire comme

$$\begin{aligned} F(X) := \; & \tfrac{1}{2}\sum_{j=1}^p\sum_{i=1}^m \|X_i - a_j\|^2 - \tfrac{1}{2}\sum_{j=1}^p \max_{i=1..m} \sum_{r=1,r\neq i}^m \|X_r - a_j\|^2 + \tfrac{1}{2}\sum_{i=1}^m \|X_i - \tfrac{1}{m}\sum_{l=1}^m X_l\|^2 \\ & + \tfrac{\tau}{2}\sum_{j=1}^p\sum_{i=1}^m \|X_i - a_j\|^2 - \tfrac{\tau}{2}\sum_{j=1}^p \max_{j=1..p} \sum_{s=1,s\neq j}^p \|X_i - a_s\|^2 \end{aligned}$$

qui est une fonction DC sur l'espace matriciel $\mathcal{M}_{m,n}(\mathbb{R})$ avec la décomposition DC suivante :

$$\begin{aligned} G(X) := \; & \tfrac{(\tau+1)}{2}\sum_{j=1}^p\sum_{i=1}^m \|X_i - a_j\|^2 + \tfrac{1}{2}\sum_{i=1}^m \|X_i - \tfrac{1}{m}\sum_{l=1}^m X_l\|^2, \\ H(X) := \; & \tfrac{\tau}{2}\sum_{i=1}^m \max_{j=1..p} \sum_{s=1,s\neq j}^p \|X_i - a_s\|^2 + \sum_{j=1}^p \max_{i=1..m} \sum_{r=1,r\neq i}^m \tfrac{1}{2}, \|X_r - a_j\|^2. \end{aligned} \tag{8}$$

Par suite (P_t) est un programme DC dans l'espace matriciel

$$(P_t) \qquad \min\{G(X) - H(X) : X \in \mathcal{M}_{m,n}(\mathbb{R})\}.$$

Pour construire les suites matricielles $\{X^{(k)}\}$ et $\{Y^{(k)}\}$ générées par DCA appliqué à (P_t), on va calculer les sous-différentiels de H et de G^*.

3.2 Calcul de $\partial H(X)$

On a $H(X) := \frac{\tau}{2}\sum\limits_{i=1}^m \max\limits_{j=1..p} \sum\limits_{r=1,r\neq j}^p \|X_i - a_r\|^2 + \sum\limits_{j=1}^p \max\limits_{i=1..m} \sum\limits_{r=1,r\neq i}^m \frac{1}{2}\|X_r - a_j\|^2 := \tau H_1(X) + H_2(X)$.

Les règles usuelles de calcul de sous-différentiels de fonctions convexes permettent d'obtenir les résultats suivants :

$$\partial H(X) = \tau\partial H_1(X) + \partial H_2(X) \tag{9}$$

où les sous-différentiels $\partial H_1(X)$ et $\partial H_2(X)$ s'obtiennent de manière explicite (co signifie l'enveloppe convexe).

$$\partial H_1(X) = \sum_{i=1}^m \partial h_i^1(X), \quad h_i^1(X) := \max_{j=1,\ldots,p} h_{i,j}^1(X), \quad h_{i,j}^1(X) := \frac{1}{2}\sum_{r=1,r\neq j}^p \|X_i - a_r\|^2; \tag{10}$$

$$\partial h_i^1(X) = co\{\partial h_{i,j}^1(X) : h_{i,j}^1(X) = h_i^1(X)\}, \quad [\nabla h_{i,j}^1(X)]_k = 0 \text{ si } k \neq i, (p-1)X_i - \sum_{r=1,r\neq j}^p a_r \text{ sinon.} \tag{11}$$

De manière analogue pour le sous-différentiel de $\partial H_2(X)$

$$\partial H_2(X) = \sum_{j=1}^p \partial h_j^2(X), \quad h_j^2(X) := \max_{i=1,\ldots,m} h_{j,i}^2(X), \quad h_{j,i}^2(X) := \frac{1}{2}\sum_{s=1,s\neq i}^m \|X_s - a_j\|^2; \tag{12}$$

$$\partial h_j^2(X) = co\{\partial h_{j,i}^1(X) : h_{j,i}^2(X) = h_j^2(X)\}, \quad [\nabla h_{j,i}^2(X)]_k = 0 \text{ si } k = i,\ X_k - a_j \text{ sinon.} \tag{13}$$

3.3 Calcul de ∂G^*

Après les calculs nous constatons que la fonction convexe G est une forme quadratique définie positive sur l'espace matriciel $\mathcal{M}_{m,n}(\mathbb{R})$. Sa dérivée est donnée par

$$\nabla G(X) = [(1 + (\tau+1)p)I - T]X - A \tag{14}$$

où $A := \sum_{j=1}^p A^{(j)}$, i.e., $A_i = \sum_{j=1}^p a_j$ pour $i = 1, ..., m$ et $T := \frac{1}{m}ee^T$ avec $T^2 = T$. Puisque $X = \nabla G^*(Y) \Longleftrightarrow Y = \nabla G(X)$ et $[(1 + (\tau+1)p)I - T]^{-1} = \frac{1}{1+(\tau+1)p}I + \frac{1}{(\tau+1)p}T$ on obtient

$$\nabla G^*(Y) = \frac{1}{1 + (\tau+1)p}[I + \frac{1}{(\tau+1)p}T](Y + A). \tag{15}$$

Finalement (P_t) est un programme DC particulièrement intéressant (la fonction objectif est la différence entre une forme quadratique définie positive et une fonction convexe non différentiable) dont les programmes linéarisés sont des programmes quadratiques convexes ayant des solutions explicites.

3.4 Schéma de l'algorithme DC pour résoudre le problème (P_t)

Initialisation

– Choisir m points $x_1^0, ..., x_m^0$ dans $\mathbb{R}^n$, une tolérance $\epsilon > 0$. Soit $X^{(0)} \in \mathcal{M}_{m,n}(\mathbb{R})$ telle que $(X^{(0)})_i := x_i^0$ for $i = 1, ..., m$.

Repeat $k = 1, 2, ...$

– Calculer $Y^{(k)} \in \partial H(X(^{(k)})$ à l'aide des formulations ((9) - (13)) permettant le calcul explicite de ∂H.

– Calculer $X^{(k+1)} \in \partial G^*(Y^{(k)})$ à l'aide (15)

Until $\left\|X^{(k+1)} - X^{(k)}\right\| \leq \epsilon(\left|X^{(k)}\right| + 1)$ **or** $\left|F(X^{(k+1)} - F(X^{(k)}\right| \leq \epsilon(\left|F(X^{(k)}\right| + 1)$.

Soient X^* la solution calculée par DCA et les centres artificiels correspondants $x_i^* = (X^*)_i$, $i = 1, ..., m$

Retrouver la solution

– Trouver les centres réels $y_1, ..., y_m$ (les noeuds les plus proches des centres aritificels $x_1^*, ..., x_m^*$)

– Trouver le centre total $y^* : \min_{a \in A} \sum_{i=1}^{m} \|y_i - a\|^2$.

Conclusion : nous avons présenté la reformulation, à l'aide des centres artificiels ([5]) du problème de détermination d'arbre hiérarchique (à deux niveaux) de multidiffusion dans le cadre de la programmation DC et le DCA pour sa résultion. Il s'avère que le programme DC est la minimisation de la différence d'une forme quadratique définie positive et d'une fonction convexe non différentiable, chose fort intéressante pour l'application de DCA qui s'interprète alors comme une technique qui consiste à résoudre une suite de programmes quadratiques convexes approximants dont les solutions se calculent de manière explicite. Les solutions (centres artificiels) données par DCA permettent alors de retrouver des centres réels et le centre total grâce à des procédures simples de tri.

Références

[1] Julia Neumann, Christoph Schnörr, Gabriele Steidl, SVM-based Feature Selection by Direct Objective Minimisation, Pattern Recognition, Proc. of 26th DAGM Symposium, LNCS, Springer, August 2004.

[2] Le Thi Hoai An and Pham Dinh Tao, *DC Programming : Theory, Algorithms and Applications. The State of the Art.* Proceedings of The First International Workshop on Global Constrained Optimization and Constraint Satisfaction (Cocos' 02), 28 pages, Valbonne-Sophia Antipolis, France, October 2-4, 2002.

[3] Le Thi Hoai An and Pham Dinh Tao, *Large Scale Molecular Optimization from distances matrices by a DC optimization approach,* SIAM Journal of Optimization,Volume 14, Number 1, 2003, pp.77-116.

[4] Le Thi Hoai An and Pham Dinh Tao, *The DC (difference of convex functions) Programming and DCA revisited with DC models of real world nonconvex optimization problems,* Annals of Operations Research 2005, Vol 133, pp. 23-46.

[5] Long Jia, A. Bagirov, I. Ouveysi, A.M. Rubinov, *Optimization based clustering algorithms in Multicast group hierarchies*, Proceedings of the Australian Telecommunications, Networks and Applications Conference (ATNAC), 2003, Melbourne Australia, (published on CD, ISNB 0-646-42229-4).

[6] Pham Dinh Tao and Le Thi Hoai An, *Convex analysis approach to d.c. programming : Theory, Algorithms and Applications,* Acta Mathematica Vietnamica, dedicated to Professor Hoang Tuy on the occasion of his 70th birthday, Vol.22, Number 1 (1997), pp. 289-355.

[7] Pham Dinh Tao and Le Thi Hoai An, *DC optimization algorithms for solving the trust region subproblem,* SIAM J. Optimization, Vol. 8, pp. 476-505 (1998).

[8] Tina Wong, Randy Katz, Steven McCanne, *A Preference Clustering Protocol for Large-Scale Multicast Applications*, Proceedings of the First International COST264 Workshop on Networked Group Communication, 1999, pp 1-18.

[9] Stefan Weber, Thomas Schüle, Joachim Hornegger, Christoph Schnörr, Binary Tomography by Iterating Linear Programs from Noisy Projections, Proceedings of International Workshop on Combinatorial Image Analysis (IWCIA), 2004. Auckland, New Zealand, Dec. 1.-3./2004, Lecture Notes in Computer Science, Springer Verlag.

[10] Stefan Weber, Christoph Schnörr, Thomas Schüle, Joachim Hornegger, Discrete Tomography by Convex-Concave Regularization and D.C. Programming, Technical Report 15/2003, Computer Science Series, December 2003.

Congruence entre des matrices de distance

Pierre Legendre, François-Joseph Lapointe

Département de sciences biologiques,
Université de Montréal,
Case postale 6128, succursale Centre-ville,
Montréal (Québec) Canada, H3C 3J7

RÉSUMÉ. Cet article décrit un test de congruence entre plusieurs matrices de distance (CEMD) provenant de tableaux de données destinés à être utilisés ensemble dans des analyses de données subséquentes. Ce test, qui utilise la statistique W *de Kendall, constitue une généralisation du test de Mantel au cas de plusieurs matrices. Il est appliqué ici pour la première fois pour étudier la congruence de plusieurs gènes. Le résultat permet de décider s'il convient, ou non, de les utiliser ensemble dans une même analyse phylogénétique.*

MOTS-CLÉS : analyse phylogénétique, coefficient de concordance de Kendall, combinaison de données, congruence, gènes, matrices de distance.

1 Introduction

Cet article décrit un test de congruence entre plusieurs matrices de distance provenant de tableaux de données destinés à être utilisés ensemble dans d'autres analyses de données comme des analyses de classification, des analyses phylogénétiques, ou encore des ordinations simples ou canoniques. Ce test, décrit dans [LEG 04], constitue une généralisation du test de Mantel [MAN 67] au cas de plusieurs matrices de distance. Il sera appliqué ici pour la première fois pour étudier la congruence de plusieurs gènes, afin de décider s'il convient de les utiliser ensemble pour réaliser une analyse phylogénétique.

En analyse phylogénétique, l'incongruence entre des tableaux de données portant sur les mêmes taxa peut avoir des origines diverses [WEN 98]. Même si on admet que les organismes ont une seule histoire évolutive, les données morphologiques, sérologiques et moléculaires peuvent conduire à des reconstructions phylogénétiques différentes, à cause par exemple de la convergence des caractères morphologiques. De même, les gènes nucléaires, mitochondriaux et chloroplastiques peuvent avoir des histoires évolutives différentes. Certains gènes peuvent résulter de transferts latéraux entre des branches distinctes de l'Arbre de la Vie. L'incongruence est une réalité quotidienne dans le travail des phylogénéticiens. La méthode décrite dans cet article permet de détecter la congruence et de décider s'il convient d'utiliser toute l'information dans une seule reconstruction d'arbre phylogénétique ou de réaliser des analyses séparées sur les différents jeux de données.

Il existe plusieurs méthodes statistiques permettant de tester l'incongruence des données en analyse phylogénétique (par exemple [FAR 95]). Ces méthodes se limitent cependant à la comparaison de deux jeux de données à la fois, ne s'appliquent pas aux matrices de distance et reposent sur le principe de la parcimonie. Notre approche permet de comparer plusieurs jeux de données, présentés sous la forme de distances ou non, ainsi que des matrices d'arbres reconstruits avec d'autres approche que la parcimonie.

2 Le test CEMD

L'analyse de la *congruence entre des matrices de distance* (CEMD) prend pour point de départ p matrices de distance, ou encore p tableaux de données décrivant n objets à l'aide de m_1, m_2, …, m_p variables (gènes, nucléotides ou bases en génétique, ou autres types de variables dans d'autres domaines d'application). Le test procède comme suit. Pour tester l'hypothèse nulle (H_0 : incongruence de toutes les matrices de distance) soumises au test contre l'hypothèse contraire (H_1 : au moins deux de ces matrices sont congruentes).

1. Si on désire étudier des tableaux de données brutes, on calcule, pour chaque tableau, une matrice de distance appropriée aux données qu'il contient. La mesure de distance peut varier d'un tableau à l'autre.
2. On déplie la portion triangulaire (supérieure ou inférieure) de chaque matrice de distance en un vecteur. On écrit ces vecteurs dans les lignes successives d'une matrice de travail qui comporte p lignes et $n(n–1)/2$ colonnes. Dans le cas des matrices de distance asymétriques, on écrit les portions triangulaires supérieure et inférieure de la matrice de distance dans la matrice de travail.
3. On transforme les distances en rangs, ligne par ligne.
4. On calcule le coefficient de concordance W de Kendall entre les lignes de la matrice de travail. On transforme W en une statistique χ^2 de Friedman qui fournit la valeur de référence (χ^2_{ref}) du test statistique.
5. La statistique est testée par permutations. Pour ce faire, on permute chaque matrice de distance comme dans le test de Mantel [MAN 67, LEG 00], et ce indépendamment d'une matrice à l'autre. On permute les objets de la matrice au hasard et on récrit les distances dans la matrice de travail dans l'ordre correspondant aux objets permutés. On calcule la statistique χ^2 pour les données permutées.
6. On répète l'étape 5 un grand nombre de fois (par ex. 999 fois). On assemble toutes les valeurs χ^2 ainsi obtenues, y compris la valeur χ^2_{ref}, dans une distribution, et on détermine combien de ces valeurs sont supérieures ou égales à la valeur χ^2_{ref}. La probabilité unilatérale des données sous l'hypothèse nulle du test est obtenue en divisant cette valeur par le nombre de permutations plus une (par ex. 1000).

Lorsque le test global est significatif, cela indique qu'il y a au moins deux matrices qui sont congruentes. On peut réaliser des tests *a posteriori* de la contribution de chaque matrice à la statistique χ^2 en permutant une seule matrice à la fois. Une matrice qui n'est pas congruente avec d'autres n'aura, une fois permutée, que peu d'effet sur la statistique globale. L'hypothèse nulle de ce test est H_0 : incongruence de cette matrice de distance par rapport à toutes les autres. La méthode donne un poids égal à toutes les matrices de distance dans la procédure de test. Une version pondérée du test permet de donner des poids inégaux aux matrices, par exemple en fonction du nombre de gènes ou de bases que contient chaque tableau de données [LEG 04]. Le test de Mantel entre les matrices de distance transformées en rangs fournit de l'information complémentaire, cette fois au niveau de chaque paire de tableaux de données.

3 Exemple

Mark Springer a obtenu 8 matrices de distances patristiques représentant des arbres phylogénétiques obtenus par analyse phylogénétique de gènes séquencés sur des animaux représentant 11 ordres de mammifères [SPR 99]. Les gènes sont les suivants : (1) 12S, (2) valine, (3) 16S, (4) vwf, (5) irbp, (6) a2ab, (7) cytb et (8) aquaporine. Les gènes 1, 2, 3 et 7 font partie de l'ADN mitochondrial alors que les gènes 4, 5, 6 et 8 proviennent de l'ADN nucléaire. Nous allons calculer des statistiques de Mantel et réaliser des analyses de congruence entre ces matrices de distance afin de déterminer si elles pourraient être utilisés ensemble dans une méthode de consensus pour reconstruire la phylogénie des ordres de mammifères.

Une première analyse CEMD a permis de rejeter l'hypothèse d'indépendance des 8 matrices de distance ($P = 0.0001$ après 9999 permutations des distances à l'intérieur de chaque matrice). Les tests *a posteriori* rejettent l'hypothèse d'incongruence des matrices individuelles, à l'exception des matrices 2 et 7 ($P = 0.0994$ après correction de Holm pour 8 tests simultanés). Les corrélations de Mantel sont faibles entre ces deux matrices et les matrices 4, 5, 6 et 8. Cela indique que le groupe de 8 matrices n'est pas homogène.

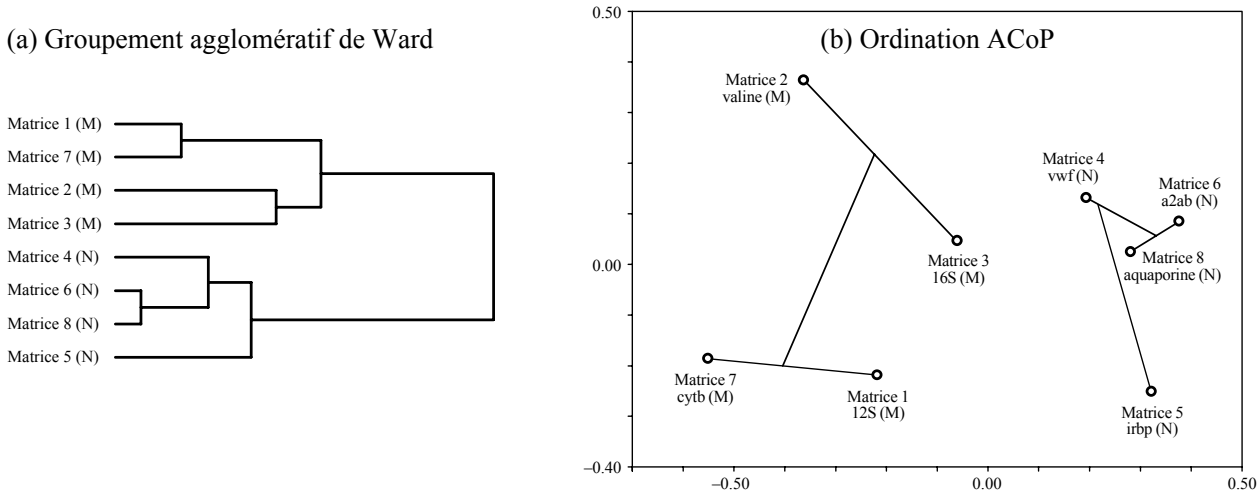


Figure 1. Groupement et ordination des matrices de distance représentant les 8 arbres phylogénétiques, basés sur les corrélations de Spearman entre matrices de distance. (a) Groupement agglomératif de Ward. (b) Ordination par analyse en coordonnées principales (ACoP). La topologie du groupement est dessinée sur l'ordination, à l'exclusion de la fusion finale. N : gène nucléaire. M : gène mitochondrial.

Un groupement agglomératif de Ward a permis d'identifier deux groupes (Fig. 1a) : les matrices 1, 2, 3 et 7 d'une part qui représentent les arbres dérivés des gènes mitochondriaux et les matrices 4, 5, 6 et 8 d'autre part qui représentent les arbres obtenus des gènes nucléaires. Puisque les statistiques de Mantel sont toutes positives, elles se comportent comme des similarités. Après transformation des similarités en distances, une ordination par analyse en coordonnées principales du tableau des statistiques de Mantel montre clairement la séparation des matrices en deux groupes (Fig. 1b).

Tableau 1. Résultats des tests CEMD globaux ainsi que des tests *a posteriori* portant sur les matrices de distance individuelles. P = probabilité (après 9999 permutations) sans correction, P_H = après correction de Holm pour 4 tests simultanés.

(a) Matrices correspondant aux gènes 1, 2, 3 et 7	**(b) Matrices correspondant aux gènes 4, 5, 6 et 8**
Test CEMD global. H_0 : les 4 mat. sont incongruentes	Test CEMD global
Statistique W de Kendall = 0.64755	Statistique W de Kendall = 0.78021
χ^2 de Friedman = 139.87013, P=0.0001 (rejet H_0)	χ^2 de Friedman = 168.52597, P=0.0001 (rejet H_0)
Tests *a posteriori*. H_0 : incongruence de cette matrice	Tests *a posteriori*
Matrice 1 : P = 0.0003, P_H = 0.0012 (rejet de H_0)	Matrice 4 : P = 0.0001, P_H = 0.0004 (rejet de H_0)
Matrice 2 : P = 0.0385, P_H = 0.0385 (rejet de H_0)	Matrice 5 : P = 0.0001, P_H = 0.0004 (rejet de H_0)
Matrice 3 : P = 0.0161, P_H = 0.0322 (rejet de H_0)	Matrice 6 : P = 0.0001, P_H = 0.0004 (rejet de H_0)
Matrice 7 : P = 0.0084, P_H = 0.0252 (rejet de H_0)	Matrice 8 : P = 0.0001, P_H = 0.0004 (rejet de H_0)
Statistiques de Mantel calculées sur les rangs	Statistiques de Mantel calculées sur les rangs
Matrice 1 1.0000 0.3976 0.6193 0.7793	Matrice 4 1.0000 0.6303 0.7670 0.7061
Matrice 2 0.3976 1.0000 0.5002 0.4672	Matrice 5 0.6303 1.0000 0.6763 0.5636
Matrice 3 0.6193 0.5002 1.0000 0.4167	Matrice 6 0.7670 0.6763 1.0000 0.8984
Matrice 7 0.7793 0.4672 0.4167 1.0000	Matrice 8 0.7061 0.5636 0.8984 1.0000

Les test CEMD ont été répétés pour les deux groupes séparément (Tableau 1). L'hypothèse globale d'indépendance des 4 membres de chaque groupe fut rejetée, de même que celle d'indépendance des matrices individuelles lors des tests *a posteriori*. Les statistiques de Mantel sont élevées ; elles montrent le degré de corrélation de chaque paire de matrices de distance.

L'analyse CEMD a montré l'existence de deux groupes distincts de gènes parmi les 8 matrices de distance et a confirmé la congruence interne de chacun de ces groupes. Dans ces conditions, il sera préférable de réaliser une reconstruction phylogénétique pour les gènes nucléaires et une autre pour les gènes mitochondriaux. Lapointe et Cucumel [LAP 02] avaient cherché comment diviser ces mêmes arbres phylogénétiques en groupes avant des analyses de consensus. La méthode CEMD nous a permis de déterminer statistiquement quels sont les arbres dont il est intéressant de rechercher le consensus.

4 Discussion

Des simulations numériques rapportées dans [LEG 04] ont montré que le test global de même que les tests *a posteriori* ont une erreur de type I correcte et une bonne puissance. La puissance des tests augmente en fonction du nombre d'objets (n) et du nombre de matrices congruentes dans l'étude. Pour un nombre donné de matrices congruentes, la puissance diminue lorsqu'on augmente le nombre de matrices incongruentes dans l'analyse. De nouvelles simulations sont en cours pour établir la puissance de ce test pour des données phylogénétiques et la comparer à celle d'autres tests qui ont été proposés dans ce domaine [FAR 95].

Un programme d'ordinateur permettant de réaliser le test CEMD (le sigle en anglais est CADM) est disponible sur le site WWWeb <http://www.bio.umontreal.ca/legendre/>.

5 Bibliographie

[FAR 95] FARRIS J.S., KÄLLERSJÖ M., KLUGE A.G., BULT C., "Testing significance of incongruence", *Cladistics*, vol. 10, 1995, p. 315-319.

[LAP 02] LAPOINTE F.J., CUCUMEL G., "Multiple consensus trees", *in: Classification, Clustering and Data Analysis - Recent Advances and Applications*, Jajuga K., A. Sokolowski A., Bock H.-H., editors, Springer-Verlag, Berlin, 2002, p. 359-364.

[LEG 00] LEGENDRE P., "Comparison of permutation methods for the partial correlation and partial Mantel tests", *Journal of Statistical Computation and Simulation*, vol. 67, 2000, 37-73.

[LEG 04] LEGENDRE P., LAPOINTE F.-J., "Assessing congruence among distance matrices: single malt Scotch whiskies revisited", *Australian and New Zealand Journal of Statistics*, vol. 46, 2004, p. 615-629.

[MAN 67] MANTEL N., "The detection of disease clustering and a generalized regression approach", *Cancer Research*, vol. 27, 1967, p. 209-220.

[SPR 99] SPRINGER M.S., AMRINE H.M., BURK A., STANHOPE M.J., "Additional support for Afrotheria and Paenungalata, the performance of mitochondrial versus nuclear genes, and the impact of data partitions with heterogeneous base composition", *Systematic Biology*, vol. 48, 1999, p. 65-75.

[WEN 98] WENDEL J.F., DOYLE J.J., "Phylogenetic incongruence : window into genome history and molecular evolution", *in* : *Molecular systematics of plants, II. DNA sequencing*, Kluwer Academic Publishers, Boston, 1998, p. 265-296.

Etat de l'art de la construction de variables

Gaëlle Legrand, Nicolas Nicoloyannis

Laboratoire ERIC,
Université Lumière Lyon 2,
5 avenue Pierre Mendès-France
69676 Bron Cedex, France

RÉSUMÉ. La qualité d'apprentissage est fortement liée à la présence de variables discriminantes. Les variables composant l'espace de représentation des données ne sont pas forcément les mieux adaptées pour décrire le problème. Or, en l'absence de nouvelles informations disponibles, il convient de créer de nouvelles variables qui permettront d'expliciter l'espace de représentation. La construction de variables permet de résoudre ce problème.

MOTS-CLÉS : Construction de variables, espace de représentation, qualité d'apprentissage.

1 Introduction

La qualité d'un apprentissage est entre autres choses liée à la présence de variables discriminantes. Or, dans le cas d'une qualité d'apprentissage médiocre et en l'absence de nouvelles informations disponibles, il est nécessaire de trouver un moyen qui, à partir de l'information disponible, nous permettra de re-décrire les données d'entrée du problème d'apprentissage considéré et éventuellement d'obtenir de nouvelles variables discriminantes. Les méthodes de construction de variables résolvent ce problème. En effet, la construction de variables permet, lors de la phase de prétraitement des données, la création de nouvelles variables synthétiques. Ces variables synthétiques sont issues de la découverte de relations entre variables initiales. Les méthodes de construction de variables entraînent une augmentation de l'espace de représentation des données, dans la mesure où de nouvelles variables sont construites. Cependant, aucune information extérieure aux données initiales n'est ajoutée lors du processus de construction. La construction de variables doit permettre l'accroissement de la qualité prédictive.

2 Classification des méthodes de construction

Il existe différentes taxonomies des méthodes de construction de variables : celle de Bloedorn, [BLO 98] et celle de Fawcett, [FAW 93]. Cependant, elles ne permettent pas de classifier la totalité des méthodes de construction. Nous proposons de classifier les méthodes en quatre catégories, inspirées fortement de la taxonomie de Bloedorn, de la manière suivante : les méthodes de construction par analyse topologique des arbres, les méthodes de construction par analyse et exploration des données, les méthodes de construction basée sur l'utilisation des connaissances du domaine ou d'un expert et les méthodes multi-stratégiques.

3 Construction par analyse et exploration des données

Ces méthodes analysent et explorent les données, les relations existantes entre variables exogènes, individus, et variable endogène afin de déterminer de nouvelles variables.

BACON, [LAN 83], base sa procédure de construction sur les interdépendances existantes entre des variables numériques.
STAGGER, [SCH 87]et [SCH 86], génère de nouvelles variables à l'aide de combinaisons booléennes de variables numériques. STAGGER associe l'apprentissage par pondération et l'apprentissage de fonctions booléennes.
FCE, [CAR 92], prend comme point de départ un ensemble d'espaces de représentation. Après avoir détecté les parties inconsistantes de chaque espace, un nouvel espace de représentation est construit. Cet espace est un produit des espaces déjà existants : sa taille est donc supérieure à celle des espaces initiaux.
AQ17-DCI, [BLO 91] est basé sur l'algorithme AQ classique [MIC 69], mais il inclut également un algorithme d'induction qui génère de nouvelles variables. La qualité de chaque nouvelle variable générée est évaluée selon une fonction de qualité. Si cette fonction est supérieure à un certain seuil alors la variable est sélectionnée. Cet algorithme fonctionne en deux parties. La première partie correspond au traitement des variables numériques et la seconde partie au traitement des variables binaires.
Méthode utilisant la théorie des treillis : L'algorithme IGLUE, développé par [NGU 98] procède en deux étapes : la construction du treillis pour les exemples positifs de l'ensemble d'apprentissage et la construction de nouvelles variables.
Le système Genetic Constructive Induction (GCI) [BEN 96] utilise une approche par programmation génétique pour construire de nouvelles variables. Un algorithme d'apprentissage par rétropropagation réalise la tâche d'apprentissage puis utilise l'espace de représentation modifié. L'espace de recherche du système de programmation génétique consiste en l'ensemble des combinaisons possibles des opérateurs de construction. De nombreux autres algorithmes utilisent la programmation génétique tels que GABIL [DEJ 93], GA-SMART [KIR 92], ou [VAF 95].
Les méthodes d'analyse de données analysent les données dans leur ensemble, en prenant en compte toutes les variables.
MIRO, [DRA 89], utilise les connaissances du domaine sous la forme d'un ensemble de règles spécifiées par un expert afin de construire un nouvel espace de représentation sur lequel sera appliqué un processus d'induction afin de construire de nouvelles variables.
IB3-CI, [AHA 91A], génère des variables booléennes à partir d'opérateurs de conjonction et combine l'apprentissage à partir d'instance de [AHA 91A] avec la construction de variables incrémentale de [SCH 87].
AQ15, [MIC 86] : Un expert du domaine définit un ensemble de règles qui sont fournies à l'algorithme AQ15. Ces règles sont soit sous une forme arithmétique soit sous une forme logique. AQ15 utilise cet ensemble des règles pour la construction de nouvelles variables.
RINCON a pour but explicite d'accroître la concision de la théorie du domaine, et est incrémental. La théorie du domaine est représentée par un graphe acyclique de concepts, organisés du général au spécifique.

4 Construction de variables par analyse topologique des arbres

Les méthodes de ce type déterminent de nouvelles variables par l'analyse des règles issues d'arbres d'induction.
CITRE, [MAT 90], est un système basé sur les arbres de décision. Il effectue de la construction de variables en sélectionnant les relations des nouvelles variables dans les branches positives de l'arbre.
AQ17-HCI, [WNE 94], génère des nouvelles variables à partir des règles obtenues à partir de l'algorithme AQ. AQ est appliqué sur l'ensemble d'apprentissage et génère un ensemble de règles. Ces règles sont ensuite évaluées par l'intermédiaire d'un critère d'évaluation, et les meilleures règles sont combinées en de nouvelles variables. Ces variables sont incorporées dans l'ensemble d'apprentissage et le processus d'apprentissage est répété. Le processus continue jusqu'à satisfaction d'un critère d'arrêt lié aux performances des nouvelles variables.
Les algorithmes STRUCT, [WAT 91], **et PRAX**, [BAL 92], fonctionnent de la même manière que AQ17-HCI : ils utilisent les règles d'apprentissage pour représenter et créer de nouvelles variables.

FRINGE, [PAG 89] et [PAG 90], s'applique initialement sur des problèmes à deux classes avec des variables exogènes booléennes. C'est un processus itératif qui se déroule en trois étapes : La construction d'un arbre de décision, l'analyse de cet arbre et la détermination des variables synthétiques candidates qui seront introduites dans la liste des variables prédictives.
LFC associe la construction de variables intermédiaires et de la forward selection. LFC construit un arbre de décision en appliquant une recherche de type forward selection contrainte, combinée à la construction de variables synthétiques sur les nœuds.
GALA, [HU 96], est un algorithme de construction qui ressemble à LFC de par son ensemble d'opérateurs.

5 Construction de variables Multi-stratégique

L'une des méthodologies les plus importantes en construction de variables est l'intégration de multiples stratégies d'apprentissage coopérant afin d'obtenir des résultats de bonne qualité.
AQ-BC combine induction supervisée et classification bayésienne non supervisée. L'utilisation de la classification bayésienne par l'intermédiaire du système AUTOCLASS, [CHE 96], a pour but la création d'un espace de représentation plus adapté à l'apprentissage.
INDUCE-1, [MIC 77], est une méthode dirigée à la fois par les données et par les connaissances du domaine. Elle utilise une variété de règles et de procédures pour générer de nouvelles variables, nommées méta-variables. Ceci s'effectue grâce à une description structurelle des exemples d'apprentissage, qui correspond à la partie dirigée par les connaissances du domaine, associée à la détermination des dépendances qualitative entre les variables, qui correspond à la partie dirigée par les données.
AQ17, [BLO 93], est une méthode à la fois dirigée par les données, les hypothèses et les connaissances du domaine. Elle intègre de manière synergique les systèmes INDUCE-1, AQ15, AQ17-DCI et AQ17-HCI.

6 Conclusions

Il existe un grand nombre de méthodes de construction de variables. Cependant, les méthodes ne traitant qu'un seul type de variables sont difficilement applicables sur des problèmes réels. C'est le cas des méthodes BACON et STAGGER qui ne traitent que des variables numériques. C'est également le cas de la méthode de FRINGE qui ne travaillent que sur des variables booléennes. L'utilisation de cette méthode est également restreinte par le fait qu'elle ne traite que des problèmes à deux classes. La méthode LFC, quant à elle, est trop difficile à paramétrer. Le type de méthodes qui nous paraît le plus attrayant sont les méthodes utilisant l'analyse topologique des arbres, telles que FRINGE. En effet, ces méthodes, grâce à l'utilisation d'un arbre d'induction pour générer les règles qui serviront à la construction de nouvelles variables, tiennent compte des relations existantes entre les variables exogènes et plus particulièrement entre certaines modalités de différentes variables exogènes. Ces méthodes sont, de plus, intéressantes car elles permettent de prendre en compte les liens entre les règles générées et les classes de la variable endogène.

7 Bibliographie

[AHA 91A] Aha, D. *Incremental constructive induction: An instance-based approach*. in *Proc. of the Eighth International Workshop on Machine Learning*. 1991. Evanston.
[AHA 91B] Aha, D.W., D. Kibler, and M.K. Albert, *Instance-Based Learning Algorithms.* Machine Learning, 1991. **6**: p. 37-66.
[BAL 92] Bala, J., R.S. Michalski, and J. Wnek. *The principal axes method for constructive induction*. in *Proc. International Conference on Machine Learning*. 1992. Aberdeen, Scotland.
[BEN 96] Bensusan, H. and I. Kuscu. *Constructive Induction using Genetic Programming*. in *ICML'96 Evolutionary computing and Machine Learning Workshop*. 1996.
[BLO 91] Bloedorn, E. and R.S. Michalski, *Constructive Induction from Data in AQ17-DCI: Further Experiments*. 1991, George Mason University: Fairfax, VA.

[BLO 93] Bloedorn, E., J. Wnek, and R.S. Michalski. *Multistrategy Constructive Induction: AQ17-MCI.* in *Proc. of the Second International Workshop on Multistrategy Learning (MSL-93)*. 1993: Harpers Ferry.

[BLO 98] Bloedorn, E. and R. Michalski, Data-driven constructive induction. IEEE Trans. on Intelligent Systems, 1998. 13(2): p. 30-37.

[CAR 92] Carpineto, C. Trading off consistency and efficiency in version-space induction. in Proc. of Ninth International Machine Learning Conference. 1992. Aberdeen, Scotland.

[CHE 96] Cheeseman, P. and P. Stutz, Bayesian Classification (Autoclass): Theory and Results. Advances in Knowledge Discovery and Data Mining, ed. U.M.F.A. Press. 1996.

[DEJ 93] DeJong, K.A., W.M. Spears, and F.D. Gordon, Using genetic algorithms for concept learning. Machine Learning, 1993(13): p. 161-188.

[DRA 89] Drastal, G., S. Raatz, and G. Czako. Induction in an abstract space. in Proc. of the Eleventh International Joint Conference on Artificial Intelligence. 1989. Detroit: MI.

[FAW 93] Fawcett, T., Feature Discovery for Inductive Learning, in Department of Computer and Information Science Univ. Of Massachusetts. 1993.

[HU 96] Hu, Y. and D. Kibler. Generation of Attributes for Learning Algorithms. in Proc. of the Thirteenth National Conference on Artificial Intelligence (AAAI96). 1996. Portland.

[KIR 92] Kira, K. and L.A. Rendell. A practical approach to feature selection. in Proc. of the Ninth International Conference on Machine Learning. 1992.

[LAN 83] Langley, P., G.L. Bradshaw, and H. Simon, Rediscovering chemistry with the Bacon system, in Machine Learning: An Artificial Intelligence Approach, J.C. R. Michalski, and T. Mitchell, Editor. 1983: Tioga, Palo Alto, CA. p. 307-330.

[MAT 90] Matheus, C.J. Adding Domain Knowledge to SBL thorough Feature Construction. in Proc.of the Eighth National Conference on Articial Intelligence. 1990.

[MIC 69] Michalski, R.S. On the Quasi-Minimal Solution of the Covering Problem. in Proceedings of the V International Symposium on Information Processing. 1969. Yugoslavia.

[MIC 77] Michalski, R.S. and J.B. Larson, Inductive Inference of VL Decision Rules. ACM SIGART Newsletter, 1977. 63: p. 38-44.

[MIC 86] Michalski, R.S., et al. The multipurpose incremental learning system AQ15 and its testing application to three medical domains. in AAAI-86. 1986.

[NGU 98] Nguifo, E.M. and P. Njiwoua. Using Lattice-based Framework as a Tool for Feature Extraction. in Proceedings of the 10 th European Conference on Machine Learning. 1998.

[PAG 89] Pagallo, G. Learning DNF by decision trees. in Proc. of the eleventh International Joint Conference on Artificial Intelligence. 1989.

[PAG 90] Pagallo, G. and D. Haussler, Boolean Feature discovery in empirical learning. Machine Learning, 1990: p. 71-99.

[SCH 87] Schlimmer, J.S. Incremental adjustment of representations in learning. in Proc. of the 4 th International Conference on Machine Learning. 1987.

[SCH 86] Schlimmer, J.C. and R.H. Granger, Incremental learning from noisy data. Machine Learning, 1986. 1: p. 317-354.

[VAF 95] Vafaie, H. and K. DeJong. Genetic algorithms as a tool for restructuring feature space representation. in Proc. of the International Conference on Tools with Artificial Intelligence. 1995.

[WNE 94] Wnek, J. and R.S. Michalski, Hypothesis-driven constructive induction in AQ-HCI: A method and experiments. Machine Learning, An Artificial Intelligence Approach, 1994. 14: p. 139-168.

[WAT 91] Watanabe, L. and L. Rendell. Learning structural decision trees from examples. in Proc. of the 12th International Joint Conference on Artificial Intelligence. 1991.

Une forme unifiée pour les indices de discrimination de classes. Application en cas de données génotypiques

Israël César Lerman

IRISA Campus Universitaire de Beaulieu
35042 RENNES Cedex
FRANCE
Email : lerman@irisa.fr

RÉSUMÉ. Nous proposons une forme unique pour un indice numérique de discrimination ou d'explication de classes homogènes d'individus ou de catégories par des variables descriptives. Cette forme embrasse les structures de données les plus diverses : numériques, booléennes, de contingence et surtout, relationnelles de différents types. Cette forme est issue des cas de données classiques (e.g. numériques) pour recouvrir le cas des données relationnelles. Cette extension est fondée sur un développement spécifique faisant appel à une famille de coefficients d'association entre variables relationnelles. Une application en cas de données génotypiques liées à l'hémochromatose (surcharge en fer) est considérée.

MOTS-CLÉS : Discrimination, classification, rapport de corrélation, variables qualitatives relationnelles, données génotypiques.

1 Introduction

La donnée est une partition π sur un ensemble **O** d'objets en classes homogènes relativement à une description par un ensemble **A** d'attributs (on dit encore variables ou caractères). π est généralement obtenue au moyen d'un algorithme de classification automatique. Un problème fondamental de la classification des données consiste à « comprendre » la partition π et à interpréter chacune de ses classes dans les termes de la description initiale [CEL 89] [YOU 04]. Nous considérons ici l'importance des rôles de chacune des variables prises séparément. Le cas de référence le plus basique est celui où l'attribut est numérique. C'est alors un rapport de corrélation qui permet de mesurer le degré de discrimination de la partition π par l'attribut. Le morcellement additif de ce dernier selon les différentes classes permet d'évaluer le rôle discriminant de l'attribut pour chacune des classes. Nous rappellerons au paragraphe 2 l'expression très classique d'un tel indice et nous évoquerons les extensions naturelles en cas de données booléennes ou de contingence (explication d'une classification des lignes à travers les colonnes). Ce type d'indice a deux formes ; la première est définie comme un rapport de variances et la seconde, comme le carré d'un coefficient de corrélation. C'est l'interprétation corrélative de l'indice qui nous permettra de traiter, *en préservant toute leur globalité,* le cas des *variables qualitatives de toutes sortes*. En effet, ce qui est généralement proposé dans la littérature pour ce problème de discrimination, passe par la décomposition de la variable qualitative en autant d'attributs booléens qu'elle a de modalités [YOU 04]. De la sorte, il n'est tenu aucun compte de la sémantique sous-jacente à l'ensemble des valeurs de la variable. Pour résoudre ce problème que nous considérons au paragraphe 3, nous faisons appel au développement d'une famille spécifique de coefficients d'association entre variables relationnelles. L'expérimentation des coefficients pour des données génotypiques touchant le métabolisme du fer sera présentée au paragraphe 4. Enfin, le paragraphe 5 définit la conclusion.

2 Les cas les plus classiques

Le cas de référence et le plus basique est celui où la variable que nous notons *v* est numérique. Désignons la partition π par $\{\mathbf{O}_c \mid 1 \le c \le C\}$. À la variable *v* nous associons celle v_π qui est constante sur chacune des classes $\mathbf{O}_c$, elle est égale à la moyenne de *v* sur la classe $\mathbf{O}_c$, $1 \le c \le C$. Le rapport de corrélation se présente sous deux formes ; d'une part, celle d'un rapport de variances :

$$Dis(\pi / v) = \mathrm{var}(v_\pi) / \mathrm{var}(v) \qquad (1)$$

et d'autre part, celle d'une corrélation au carré :

$$Dis(\pi / v) = (Corr(v_\pi, v))^2 \ . \qquad (2)$$

C'est cette interprétation corrélative du coefficient de discrimination d'une classification par une variable qui constituera le fil directeur fondamental de notre développement. Le second membre de (1) peut très directement être morcelé selon les différentes classes.

L'extension en cas de données booléennes ou de contingence est assez naturelle dès lors qu'on adopte une représentation mathématique adéquate du tableau des données. Ainsi, dans le cas d'un tableau de contingence de dimension *(n,p)*, l'ensemble des lignes est assimilé à l'ensemble **O** des objets, alors que l'ensemble des colonnes représente l'ensemble **A** des attributs. De la sorte, un objet représente une catégorie d'individus. La valeur de l'attribut *j* sur l'objet *i* est définie par la proportion relative d'individus de la catégorie *i* qui sont dans la catégorie *j*. C'est bien la représentation euclidienne de l'analyse des correspondances qui est adoptée, où l'espace de représentation géométrique R^p est muni de la métrique du χ^2. Qu'il s'agisse de données booléennes ou de contingence, ce sont des expressions très spécifiques par rapport à la nature des données qui sont obtenues, en harmonie avec les expressions (1) et (2) ci-dessus [LER 04].

3 Le cas de données relationnelles

Une variable relationnelle peut formellement être définie par un graphe valué sur l'ensemble **O** des objets. La comparaison de variables de ce type a intéressé de nombreux chercheurs [DAN 44] [HUB 83] [LEC 76] [LER 77] [LER 81] [LER 92] [MANT 67] [OUA 91]. Cependant, les premières publications [DAN 44] [MANT 67] dont certes, on peut recueillir l'aspect technique pouvaient être plus ou moins particulières et ne correspondaient pas à l'optique générale que nous développons. Compte tenu du fait qu'une variable qualitative a un très faible nombre de valeurs (modalités, catégories) eu égard au nombre d'objets, le graphe mentionné ci-dessus est à considérer sur un ensemble quotient. Ce dernier est défini par un ensemble de classes, caractérisées chacune par une même valeur de l'attribut. La nature de ce graphe que nous dirons également quotient, va directement dépendre de la structure dont se trouve munie l'ensemble des valeurs. Le cas le plus pauvre où l'ensemble des arcs du graphe quotient est vide, est celui où il n'y a aucune structure ; il s'agit du cas où la variable qualitative est nominale. Un autre cas classique est celui de la variable qualitative ordinale, où l'ensemble des modalités est muni d'un ordre total et strict. Dans ce cas, le graphe quotient se réduit à une chaîne. On peut également considérer le cas plus général où l'ensemble des valeurs est muni d'un préordre partiel. Le cas d'une variable « préordonnance » est celui où l'ensemble de ses modalités est muni d'une similarité ordinale : il s'agit d'un préordre total sur l'ensemble des couples de valeurs. Nous coderons ce préordre à partir de la notion de « rang moyen », ce qui correspond à une valuation particulière d'une relation binaire. Il y a enfin le cas d'une variable relationnelle valuée quelconque.

Pour opérer l'extension d'un coefficient tel que (2), il y a lieu de retrouver à partir d'une approche combinatoire et avec une vision relationnelle, le coefficient de Bravais Pearson. Un tel coefficient peut être retrouvé selon une même procédure de normalisation et cela de deux façons différentes : soit en

regardant l'attribut numérique comme une relation unaire valuée ou bien, en le regardant d'une certaine façon, comme une relation binaire valuée. Dans chacun des deux cas le coefficient de corrélation $\rho(v,w)$ entre deux variables numériques se met sous la forme :

$$\rho(v,w) = \frac{Q(v,w)}{\sqrt{Q(v,v) \times Q(w,w)}} \qquad (3)$$

où $Q(v,w)$ s'obtient à partir de la normalisation statistique (centrage par la moyenne et réduction par l'écart type) par rapport à une hypothèse d'absence de liaison à caractère permutationnel d'un indice brut d'association qui se présente sous la forme d'une somme de produits $v(i) \times w(i)$ dans le premier cas et $(v(i)-v(i')) \times (w(i)-w(i'))$, dans le second cas, $1 \leq i \neq i' \leq n$. Cette deuxième forme constitue une articulation pour considérer –par analogie cohérente et généralisation- le cas de la comparaison de deux variables qualitatives de types quelconques, définissant des graphes valués sur l'ensemble **O** des objets. Dans le problème qui nous concerne il s'agira de confronter chacune des variables descriptives avec une classification « naturelle » issue de cette description ; laquelle définit une variable partition sur **O** [ADO 03] [LER 92] [LER 04].

4 Application à des données génotypiques[1]. Le programme v-class

Le génotype d'un individu est associé à son ADN qui comprend la suite de ses gènes. Plus exactement, la suite dont chaque composante est définie par deux copies d'un même gène provenant respectivement des deux parents. À une position donnée (dite *locus*) d'une copie donnée du gène est inclus un nucléotide pris dans l'ensemble {A, C, G, T} des quatre nucléotides. L'énorme majorité des loci est occupée par le même nucléotide pour les deux copies du même gène. Cependant, pour certains, dits SNPs (Single Nucleotid Polymorphism), une mutation peut avoir lieu et alors, provoquer dans certains cas, un changement phénotypique. Il s'agit dans notre cas de l'apparition de la maladie de l'hémochromatose (surcharge en fer du sang). Le bilan martial (teneur enfer) est évalué au moyen de 5 attributs numériques : Ferritine, Fer, Coefficient de saturation de la transferrine, Transferrine et Coefficient de fixation de la transferrine. Le problème général consiste alors à mettre en évidence une interaction possible entre des profils génotypiques et des profils relatifs à la teneur en fer. Diverses expériences ont été menées sur un échantillon d'un peu plus de 300 individus issu d'une population « normale ». L'une d'entre elles consiste en un croisement avec des indices adéquats, entre une classification obtenue à partir de la description numérique donnant les différents phénotypes et une classification obtenue à partir de la description génotypique. Pour cette dernière les variables SNPs (chacune à 3 valeurs) ont été codées en termes de préordonnances [ADO 03] [LER 04].

Nous présentons ici le résultat d'une autre expérience qui illustre directement le sujet traité dans ce papier. On commence par la construction d'une classification hiérarchique qui prend en compte une description hétérogène incluant les 5 variables numériques correspondantes au bilan martial et 20 variables préordonnances correspondantes aux SNPs [LER 03]. C'est une version particulière du logiciel CHAVL (Classification Hiérarchique par Analyse de la Vraisemblance du Lien) [LER 93] dûe à Ph. Peter (Maître de Conférences à l'École Polytechnique de l'Université de Nantes) qui a été utilisée. C'est relativement à la partition récoltée au niveau dit le plus « significatif » de l'arbre des classifications issu du programme CHAVL que les indices $(Corr(v_\pi, v))^2$ et $Corr(v_\pi, v)$ sont calculés (voir Tableau 1). C'est la distribution empirique le long de la suite des niveaux de l'arbre d'une statistique combinatoire normalisée qui permet de détecter les niveaux les plus « significatifs » [LER 81]. Le programme v-class [ADO 03] réalise les calculs des indices précédents en récoltant à partir de CHAVL le niveau pertinent qui l'intéresse. Ce qui permet de reconnaître les rôles respectifs des différentes variables de description en termes de discrimination de la classification concernée.

[1] Les données nous ont été fournies par l'UMR-CNRS «Génétique et Développement » de l'Université de Rennes 1.

VARIABLE	COEF2	COEF
1	0.515617	0.718065
2	0.286276	0.535047
3	0.142609	0.377636
4	0.098659	0.314100
5	0.374676	0.612107
6	0.000857	0.029271
7	0.007744	0.087999
8	0.003300	0.057447
9	0.002462	0.049619
10	0.022951	0.151496
11	0.001986	0.044569
12	0.015162	0.123136
13	0.024573	0.156757
14	0.015085	0.122821
15	0.015539	0.124655
16	0.014207	0.119193
17	0.000255	0.015969
18	0.021895	0.147971
19	0.002510	0.050096
20	0.016842	0.129776
21	0.001558	0.039472
22	0.007386	0.085945
23	0.007399	0.086018
24	0.000326	0.018063
25	0.000137	0.011715

Variables numériques (1 à 5)
Variables qualitatives préordonnances SNPs (6 à 25)

COEF correspond au coefficient de corrélation et COEF2 à son carré

Tableau 1

5 Conclusion et perspectives

À partir d'une méthode très générale pour la définition d'un coefficient d'association entre deux structures combinatoires définissant deux graphes valués sur un ensemble d'objets, nous avons montré comment on pouvait réduire le problème de l'extension d'un rapport de corrélation au cas de variables qualitatives de toutes sortes et cela, en préservant fidèlement toute la richesse sémantique dont se trouve munies les ensembles de valeurs des variables descriptives. Nous avons appliqué cette approche au cas de données difficiles touchant la maladie de l'hémochromatose. Certains effets statistiques ont été perçus. Cependant, le caractère « normal » de la population, la relative petitesse de l'échantillon et

la difficulté d'apparition de la maladie lorsqu'il y a une mutation et cela même, pour le seul SNP dont l'influence a été découverte (le C282Y) font qu'il importe de travailler à une bien plus grosse échelle en termes de taille de l'échantillon et comprenant en son sein un sous ensemble d'individus carencés.

Remerciements

Nous tenons à remercier le Professeur Yves Deugnier, « Service des Maladies du Foie & Centre d'Investigation Clinique INSERM 0203 (CHU de Rennes) » et le Professeur Jean Mosser, UMR-CNRS 6061 « Génétique du Développement », pour les discussions que nous avons pu avoir et la mise à disposition des données.

6 Bibliographie

[ADO 03] ADOUE V., " *Élaboration d'un logiciel d'explication de classes pour une classification de données génotypiques ",* Rapport DESS CCI, IFSIC, Université de Rennes 1, Septembre 2003.

[CEL 89] CELEUX G., DIDAY E., GOVAERT G., LECHEVALLIER Y., RALAMBONDRAINY H., " *Classification automatique des données ",* 1989*,* Dunod, Paris.

[DAN 44] DANIELS H.E., " The relation between measures of correlation in the universe of sample permutations ", *Biometrika*, vol. 33, 1944, p. 129-135.

[HUB 83] HUBERT L., " Inference procedures for the evaluation and comparison of proximity matrices ", *Numerical Taxonomy,* Ed. J. Felsenstein, NATO ASI Series, Berlin, Springer-Verlag, p. 209-228.

[LEC 76] LECALVÉ G., " Un indice de similarité pour des variables de types quelconques ", *Statistique et Analyse des Données,* 01-02, 1976, p. 39-47.

[LER 77] LERMAN I.C., " Formal analysis of a general notion of proximity between variables ", *Congrès Européen de Statisticiens,* Grenoble, Sept. 1976, North Holland, 1977.

[LER 81] LERMAN I.C., " *Classification et analyse ordinale des données* ", 1981, Dunod, Paris.

[LER 92] LERMAN I.C., " Conception et analyse de la forme limite d'une famille de coefficients statistiques d'association entre variables relationnelles. II ", *Mathématiques Informatique & Sciences humaines,* 29e année, n° 119, 1992, p. 75-100.

[LER93] LERMAN I.C., PETER., Ph., LEREDDE H., " Principes et calculs de la méthode implantée dans le programme CHAVL (Première partie) ", *La revue de modulad,* n° 13, 1993, p. 33-70.

[LER03] LERMAN I.C., PETER., Ph., " Indice probabiliste de vraisemblance du lien entre objets quelconques ; analyse comparative entre deux approches ", *Revue de Statistique Appliquée,* LI(1), 2003, p. 5-35.

[LER 04] LERMAN I.C., *" Coefficient numérique général de discrimination de classes d'objets par des variables de types quelconques. Application à des données génotypiques "*, Publication Interne Irisa n° 1652 , Décembre 2004, 32 pages.

[MAN 67] MANTEL N., "Detection of disease clustering and generalized regression approach ", *Cancer Research*, vol 27, p. 209-220.

[OUA 91] OUALI-ALLAH M., "*Analyse en préordonnances des données qualitatives. Applications aux données numériques et symboliques* ", Thèse de l'Université de Rennes 1, Décembre 1991.

[YOU 04] YOUNESS G., " *Contributions à une méthodologie de comparaison de partitions* ", Thèse de l'Université de Paris 6, Juin 2004.

Compression et classification de données de grande dimension

Sylvain Lespinats, Alain Giron, Bernard Fertil

Unité INSERM 494, CHU Pitié-Salpétrière
91 bd de l'hôpital, 75634 PARIS (France)

RÉSUMÉ. Les données de grande dimension posent des problèmes spécifiques en classification. Sur un exemple de donnés de grande dimension, nous montrons les conséquences dramatiques d'une méthode de compression originale dont le but est pourtant de préserver le voisinage local.

MOTS-CLÉS : données de grande dimension, projection locale non linéaire, visualisation, exploration, point de vue, Analyse en composantes curvilignes, placement d'objets dans un champ de forces, signature génomique, bioinformatique.

1 Introduction

Les données de grande dimension posent des problèmes inhabituels d'analyse, étant donné que les propriétés des espaces qui les contiennent ne sont pas nécessairement "évidentes". La notion de voisinage en particulier doit être réexaminée pour tenir compte du nombre de dimensions. En particulier (notamment dans le cas d'un espace euclidien), on est souvent confronté au problème de l'espace vide et de la concentration de mesure : quand le nombre de dimensions est élevé, le voisinage immédiat d'une donnée est très peu occupé tandis que la plupart des autres données se trouvent à des distances très comparables de cette dernière. D'une manière générale, les distances entre données de grande dimension sont très concentrées autour de leur moyenne.
L'exploration et l'analyse des données de grande dimension s'effectuent souvent à l'aide de méthodes de réduction dimensionnelle. L'analyse en composantes principales (ACP), les techniques de "multidimensional scaling" (MDS) [COX 94], les cartes de Kohonen (SOM) [KOH 97] sont des outils classiques dans ce contexte. D'une manière générale, une fonction de coût (loss function) permet de construire les règles de projection de l'espace original des données vers l'espace cible de dimension plus faible. Pour les problèmes de classification, la conservation du voisinage apparaît un des aspects importants à maîtriser. Dans ce travail, on présente une méthode de projection non linéaire de type MDS dont le but explicite est de préserver "au mieux" le voisinage des données. On analysera ensuite les conséquences de cette réduction de dimension pour un problème de classification de données de grande dimension concernant les signatures génomiques.

2 Principes du modèle de réduction de la dimension des données

L'approche qui est présentée ici est du type MDS. Il convient donc de définir des métriques pour l'espace d'origine des données et pour l'espace cible, une fonction de coût qui s'intéresse à caractériser l'erreur réalisée lors de la projection des données, enfin un algorithme de projection. D'une manière générale, les caractéristiques des données à analyser sont à prendre en considération pour choisir ces différents éléments.

2.1 Métrique et fonction de coût

Les données d'études concernent la signature génomique. Cette dernière caractérise la molécule d'ADN par 256 variables fréquentielles, définies sur un intervalle borné [0-1]. La signature génomique est spécifique à chaque espèce vivante. Elle peut être obtenue à partir de l'examen d'une fraction relativement faible du matériel génétique de l'espèce. En pratique, une séquence de 2000 nucléotides en donne une bonne approximation. La métrique euclidienne permet de montrer des différences statistiquement significatives entre les signatures génomiques des espèces [DES 99]. Cette métrique sera donc choisie pour illustrer la méthode.

Pour définir la fonction de coût, on s'intéresse à l'ensemble des distances entre toutes les données (ou une partie d'entre elles, voir étape d'initialisation, en 2.4), dans l'espace des données et dans l'espace cible. La fonction de coût caractérise l'erreur de projection par les écarts entre distances entre objets mesurés dans ces deux espaces. Cependant, pour préserver préférentiellement les distances concernant le voisinage proche, une pondération est appliquée progressivement pendant la phase d'optimisation de la projection pour réduire l'impact des erreurs liées aux grandes distances sur la construction locale de l'espace cible. Cette approche s'inspire des travaux de P. Demartines et J. Herault [DEM 97] ainsi que ceux de T. Kohonen [KON 97].

2.2 Algorithme d'optimisation

En général la position optimale des données à projeter dans l'espace cible ne peut être obtenue de manière analytique. Il faut mettre en oeuvre un algorithme de minimisation de fonction possédant des caractéristiques de robustesse et de convergence reconnues. Classiquement, dans le contexte des MDS, on utilise la méthode de Newton-Raphson généralisée, la méthode TABU Search [GLO 95], les algorithmes génétiques [GOL 89], le recuit simulé [DOW 95]. Nous proposons de mettre en place une méthode dynamique fondée sur le concept de « placement d'objets dans un champ de force » ("Force Directed Placement" ou FDP) [FRU 91]. La méthode FPD, qui a été décrite au début des années 80, est très utilisée par exemple pour déterminer de manière optimale la position des différents éléments d'un circuit imprimé (VLSI). Elle est par contre peu connue dans le domaine de l'analyse de données. La métaphore peut être explicitée de la manière suivante : Les données exercent des forces les unes sur les autres dont l'intensité dépend de l'écart (pondéré) entre les distances entre elles dans l'espace d'origine et dans l'espace cible. Dans le cadre de notre implémentation, les forces sont générées par l'action de ressorts dont la longueur au repos correspond à la distance entre les données qu'ils lient dans l'espace d'origine. A partir d'un état initial où les données sont placées le plus judicieusement possible dans l'espace cible, le système converge vers un état d'énergie minimum pour lequel les contraintes d'interaction entre les données sont satisfaites au mieux. Cette approche est très intéressante dans le cas des MDS, étant donné sa vitesse de convergence et ses possibilités à échapper aux minimums locaux.

Pour les problèmes de quelques milliers de données, il est possible de mettre directement en place la procédure FDP pour le placement des données dans l'espace cible. Pour les problèmes de plus grande dimension, il est souvent intéressant de sélectionner un certain nombre de données pour définir grossièrement la topologie de l'espace d'arrivée, dans un premier temps. Les autres données sont ensuite positionnées par rapport aux précédentes, en satisfaisant préférentiellement les contraintes locales. Nous avons observé que cette approche hiérarchique de la projection est très efficace, surtout lorsque les données initiales sont choisies après clustering.

2.3 Exemple de projection par la méthode FDP-MDS

Les données à projeter ont trois dimensions. Elles sont organisées en 2 boîtes cubiques avec un couvercle ouvert ne pointant pas dans la même direction. La projection dans un espace à 2 dimensions par FDP-MDS développe correctement les 2 boîtes et effectue une torsion de l'espace à grande échelle (Fig. 1). Les relations de voisinage sont conservées de manière satisfaisante.

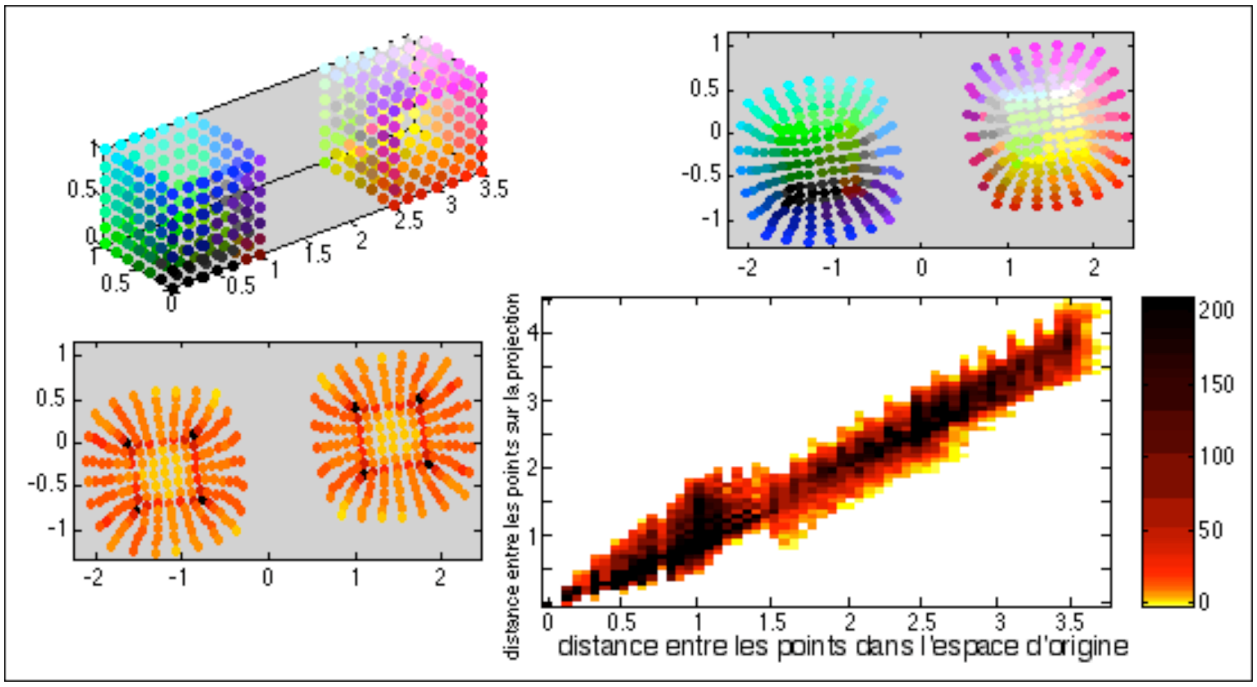


Figure 1 : Projection de 2 boites ouvertes (3D) dans un espace 2D. Haut gauche : données d'origine, haut droit : données projetées, bas gauche, satisfaction des contraintes de distance (l'indice de satisfaction croit du noir au blanc), bas droite, conservation des distances (l'intensité code la densité de points). Une version couleur des figures de cet article est disponible sur notre site web:
< http://e6.imed.jussieu.fr/afficherpub.php/sfc05.pdf>

2.4 Classification des signatures génomiques

Les données concernées par cette étude sont de deux types. Les signatures globales de 5000 espèces constituent un échantillon de la diversité des molécules d'ADN du vivant. La signature d'une espèce, *B. subtilis* a été étudié en détail. 8420 signatures correspondant à l'analyse de l'ADN dans une fenêtre glissante de taille prédéfinie ont été calculées. La signature de chacune de ces fenêtres (appelée signature locale par la suite) porte en général les caractéristiques de *B. subtilis*. Il s'agit de retrouver l'espèce d'origine des signatures locales par recherche de l'espèce la plus proche (classification au plus proche voisin), dans l'espace d'origine (256 dimensions) et dans un espace cible de dimension 3. Dans le cadre de cette courte présentation, deux situations sont étudiées : 1- l'espace cible est appris à l'aide de signatures d'espèces, 2- l'espace cible est appris à l'aide de signatures d'espèces, mais aussi de signatures locales de *B. subtilis*.

La référence de classification est calculée dans l'espace d'origine : 64 % des signatures sont correctement attribuées à *B. subtilis*. Certaines espèces dont la signature est proche de celle de *B. subtilis* réduisent de manière importante l'efficacité de la classification : En fait, *B. subtilis* est l'un des 5 plus proches voisins dans 87% des cas. Il faut noter qu'un ensemble important de signatures locales est mal classé pour des raisons biologiques connues. Lorsque l'espace cible est appris à l'aide de signatures d'espèces (cas 1), le taux de bonne classification devient négligeable : 0,7%. On retrouve 24% de signatures bien classées quand l'espace cible est appris à l'aide de l'ensemble des signatures (espèces et locales) (cas 2). L'observation des données dans l'espace cible montre que la région correspondant aux signatures locales a été développée pour satisfaire les contraintes de distances entre signatures locales quand ces dernières sont introduites dans l'échantillon d'apprentissage (Fig. 2). Malgré tout, la qualité de classification est faible.

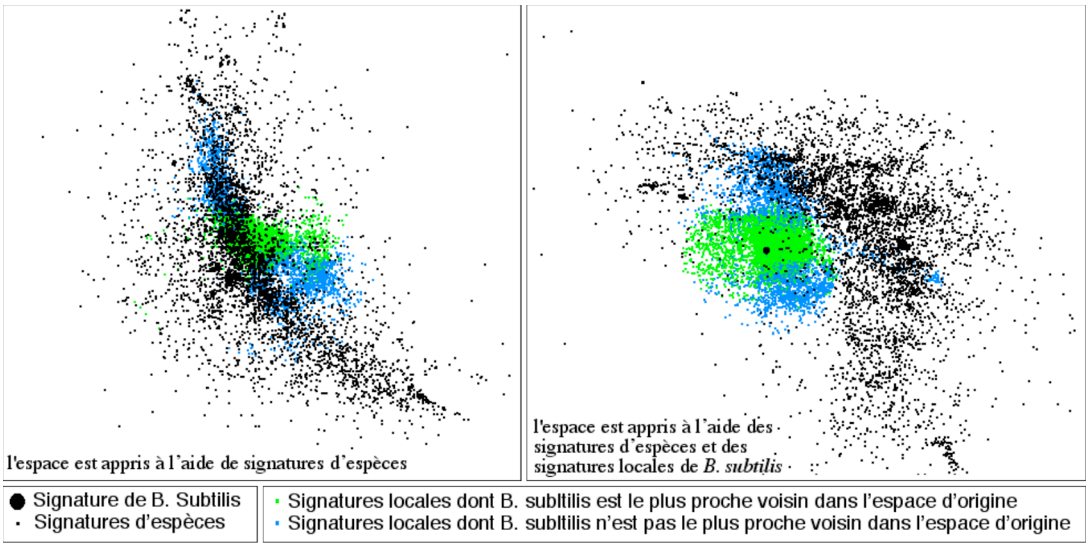


Figure 2: Signatures génomiques dans un espace de faible dimension. Une version couleur des figures de cet article est disponible sur notre site web < http://e6.imed.jussieu.fr/affichepub.php/sfc05.pdf>

3 Discussion et conclusion

L'approche de projection non linéaire décrite dans cet article a été conçue pour préserver au maximum le voisinage des données. Pour les problèmes de petite dimension, il apparaît que son efficacité est très bonne. Ce n'est malheureusement pas le cas pour les données de grande dimension où l'efficacité de classification au plus proche voisin des signatures locales est fortement réduite lors de la projection. Il est clair que la méthode de classification utilisée est particulièrement sensible aux "erreurs" de placement puisqu'il suffit dans ce cas d'une seule espèce "mal placée" pour générer une erreur. Cette situation est sûrement très fréquente lors de réduction de dimensions aussi importante (256 vers 3). Il paraît utile de rappeler que l'analyse des données résultant de taux de compression conséquents doit être effectuée avec d'infinies précautions.

Bibliographie

[COX 94] COX T., COX M., Multidimensional Scaling. Chapman & Hall, London, 1994.

[DEM 97] DEMARTINES P., HERAULT J., *Curvilinear Component Analysis : A Self-Organizing Neural Network for Nonlinear Mapping of Data Sets*, IEEE Trans. Neural Networks, 1997, 8: 148-154.

[DES 99] DESCHAVANNE P.D., GIRON A. et al, *Genomic signature: characterization and classification of species assessed by chaos game representation of sequences.* " Mol. Biol. Evol. 1999, 16: 1391-1399.

[DOW 95] DOWSLAND K.A., In C.R. Reeves ed, Modern Heuristic techniques for combinatorial problems, Chap. 2, McGraw-Hill Book Company, Bershire, 1995.

[FRU 91] FRUCHTERMAN T., REINGOLD E., *Reingold E., " Graph Drawing by Force-directed placement "* Software-Practice and Experience, 1999, 21, 1129-1164.

[GLO 95] GLOVER F., LAGUNA M., *Tabu search*, In C.R. Reeves ed. Modern Euristic techniques for combinatorial problems, Chap. 3, McGraw-Hill Book Company, Bershire, 1995.

[GOL 89] GOLDBERG D.E., *Genetic algorithms in search, Optimization, and Machine Learning*, Addison-Wesley, Reading, Massachusetts, 1989.

[KOH 97] KOHONEN T., *Self-Organizing Maps,* Springer-Verlag, 1997.

[VER 01] VERLEYSEN M., *Learning high-dimensional data,* NATO Advance Research Workshop on limitation and future trends in neural computing, Siena (Italy), 2001.

4 Remerciements

INSERM (dotation annuelle), Action inter-EPST Bio-informatique 2001 contrat N° 120910

Régression linéaire pour la prédiction de variables de type intervalle

Eufrasio de A. Lima Neto et Francisco de A.T. de Carvalho

Centro de Informatica - CIn,
Universidade Federal de Pernambuco,
Av. Prof. Luiz Freire, s/n – Cidade Universitára
CEP : 50740-540, Recife-PE, Brésil
{ealn,fatc}@cin.ufpe.br

RÉSUMÉ. Nous présentons deux approches pour ajuster une régression linéaire à des données de type intervalle. Dans la première approche on prédit le centre et l'étendue de l'intervalle de la variable dépendante à partir de l'ajustement de deux régressions linéaires sur respectivement, les centres et les étendues des intervalles des variables indépendantes. Dans la deuxième approche la prévision du centre et de l'étendue de l'intervalle de la variable dépendante est obtenue par deux régressions linéaires qui tiennent compte simultanément des centres et des étendues des intervalles des variables indépendantes. L'évaluation de ces deux approches est basée sur l'estimation de la moyenne du coefficient de détermination et de la racine carrée de la moyenne de la somme des carrés des résidus pour des données synthétiques de type intervalle dans le cadre d' une expérience Monte Carlo.

MOTS-CLÉS : Analyse des Données Symboliques, Régression Linéaire, Données de Type Intervalle.

1 Introduction

Dans un tableau de données chaque cellule contient soit une valeur numérique (correspondant à une variable quantitative) soit une catégorie (ordonné ou non) correspondant à une variable qualitative. L'analyse des données symboliques [BOC 03] a introduit des nouvelles variables dites « symboliques » qui permettent de tenir compte de la variabilité et/ou de l'incertitude présente dans les données. Par conséquent, dans un tableau de données symboliques une cellule peut contenir un intervalle, un ensemble de catégories ou encore une distribution de poids (fréquences).

Les variables de type intervalle sont souvent rencontrées dans la pratique : un intervalle peut décrire la plus petite et la plus grande valeur d'une mesure concernant un individu pendant une journée ou encore l'étendue des salaires dans une entreprise. Il peut aussi indiquer que la valeur exacte d'une mesure ne peut pas être obtenue, mais que cette valeur est dans cet intervalle.

Nous nos intéressons à l'ajustement des méthodes de régression linéaire aux données de type intervalle. Une première approche a été présentée par [BIL 02] et il consiste à ajuster un modèle usuel de régression linéaire sur le centre des intervalles et appliquer ce modèle aux limites inférieure et supérieure des variables indépendantes pour prédire, respectivement, la limite inférieure et la limite supérieure de l'intervalle de la variable dépendante.

Dans ce travail nous considérons deux autres approches dont la performance est supérieure à celle présenté par [BIL 02]. Dans la première approche, on ajuste deux modèles usuels de régression linéaire. Dans le premier modèle, l'estimation du centre d'un intervalle assumé par la variable dépendante est basée

sur les centres des intervalles assumés par les variables indépendantes. Dans le deuxième modèle, l'estimation de l'étendue d'un intervalle assumé par la variable dépendante est basée sur les étendues des intervalles assumés par les variables indépendantes. On obtient la prévision de la limite inférieure et de la limite supérieure d'un intervalle assumé par la variable dépendante à partir de l'estimation du centre et de l'étendue de ce même intervalle selon leur correspondant modèle de régression.
La deuxième approche diffère de la première essentiellement en ce qui concerne l'estimation du centre et de l'étendue d'un intervalle assumé par la variable dépendante : dans ces deux cas cette estimation maintenant est basée à la fois sur le centre et l'étendue des intervalles assumés par les variables dépendantes. Enfin, l'évaluation de ces approches est basée sur l'estimation de la moyenne du coefficient de détermination et de la moyenne de la racine carrée de la moyenne de la somme des carrés des résidus (root mean squared error) dans le cadre d'un expérience Monte Carlo.

2 Description des données et des modèles

Soit E = $\{e_1,\ldots,e_n\}$ les exemples qui sont décrit par *p+1* variables de type intervalle : $X_1,\ldots,X_p$, Y. Chaque exemple $e_i \in E$ est représenté par un vecteur d'intervalles $\mathbf{z}_i = (x_{i1},\ldots,x_{ip},y_i)$ où $X_j(e_i) = x_{ij} = [a_{ij},b_{ij}] \in \Im = \{[a,b] : a, b \in \Re, a \leq b\}$ et $Y(e_i) = y_i = [y_{iL}, y_{iU}] \in \Im$.

2.1 La méthode du centre (MC)

Ce méthode a été introduit par [BIL 02]. Les exemples sont décrit par *p+1* variables quantitatives $X_{1C},\ldots,X_{pC}$, Y_C qui assument comme valeur, respectivement, le centre des intervalles assumés par les variables $X_1,\ldots,X_p$, Y. Chaque exemple $e_i \in E$ est représenté par un vecteur de valeurs réelles $\mathbf{z}_{iC}=(x_{i1C},\ldots,x_{ipC},y_{iC})$ où $X_{jC}(e_i) = x_{ijC} = (a_{ij} + b_{ij})/2$ et $Y_C(e_i) = y_{iC} = (y_{iL} + y_{iU})/2$. Le modèle est donc :

$$\mathbf{y}_C = \mathbf{X}_C\boldsymbol{\beta}_C + \boldsymbol{\varepsilon}_C$$

où $\mathbf{y}_C = (y_{1C},\ldots,y_{nC})^T$, $\mathbf{X}_C = ((\mathbf{x}_{1C})^T,\ldots,(\mathbf{x}_{nC})^T)^T$, avec $\mathbf{x}_{iC} = (1, x_{i1C},\ldots,x_{ipC})^T$, $\boldsymbol{\beta}_C = (\beta_{0C},\ldots,\beta_{pC})^T$ et $\boldsymbol{\varepsilon}_C=(\varepsilon_{1C},\ldots,\varepsilon_{nC})^T$. L'estimation de $\boldsymbol{\beta}_C$ par la méthode des moindres carrées est donnée par :

$$\hat{\boldsymbol{\beta}}_C = \left(\left(\mathbf{X}_C\right)^T\left(\mathbf{X}_C\right)\right)^{-1}\left(\mathbf{X}_C\right)^T\mathbf{y}_C$$

Étant donnée nouvel exemple *e* décrit par $\mathbf{z} = (x_1,\ldots,x_p,y)$ et par $\mathbf{z}_c = (x_{1C},\ldots,x_{pC},y_C)$, où $x_j = [a_j,b_j]$, $y=[y_L,y_U]$, $x_{jC}=(a_j + b_j)/2$ et $y_C = (y_L + y_U)/2$, la valeur $y = [y_L,y_U]$ de Y sera prédite par

$$\hat{y}_L = \left(\mathbf{x}_L\right)^T\hat{\boldsymbol{\beta}}_C \text{ et } \hat{y}_U = \left(\mathbf{x}_U\right)^T\hat{\boldsymbol{\beta}}_C$$

où $\mathbf{x}_L = (1, a_1,\ldots,a_p)^T$, $\mathbf{x}_U = (1, b_1,\ldots,b_p)^T$ et $\hat{\boldsymbol{\beta}}_C = (\beta_{0C},\ldots,\beta_{pC})^T$.

2.2 Première méthode du centre et de l'étendue(MCE1)

Les exemples sont en plus décrit par décrit *p+1* variables quantitatives $X_{1R},\ldots,X_{pR}$, Y_R qui assument comme valeur, respectivement, l'étendue des intervalles assumés par les variables $X_1,\ldots,X_p$, Y. Chaque exemple $e_i \in E$ est aussi représenté par un vecteur de valeurs réelles $\mathbf{z}_{iR}=(x_{i1R},\ldots,x_{ipR},y_{iR})$ où $X_{jR}(e_i) = x_{ijR} = b_{ij} - a_{ij}$ et $Y_R(e_i) = y_{iR} = y_{iU} - y_{iL}$. Le modèle correspondant est donc :

$$\mathbf{y}_R = \mathbf{X}_R\boldsymbol{\beta}_R + \boldsymbol{\varepsilon}_R$$

où $\mathbf{y}_R = (y_{1R},\ldots,y_{nR})^T$, $\mathbf{X}_R = ((\mathbf{x}_{1R})^T,\ldots,(\mathbf{x}_{nR})^T)^T$, avec $\mathbf{x}_{iR} = (1, x_{i1R},\ldots,x_{ipR})^T$, $\boldsymbol{\beta}_R = (\beta_{0R},\ldots,\beta_{pR})^T$ et $\boldsymbol{\varepsilon}_R=(\varepsilon_{1R},\ldots,\varepsilon_{nR})^T$. L'estimation de $\boldsymbol{\beta}_R$ par la méthode des moindres carrées est donnée par :

$$\hat{\boldsymbol{\beta}}_R = \left(\left(\mathbf{X}_R\right)^T\left(\mathbf{X}_R\right)\right)^{-1}\left(\mathbf{X}_R\right)^T\mathbf{y}_R$$

Étant donnée un nouvel exemple *e* décrit par $\mathbf{z} = (x_1,\ldots,x_p,y)$, par $\mathbf{z}_C = (x_{1C},\ldots,x_{pC},y_C)$ et par $\mathbf{z}_R=(x_{1R},\ldots,x_{pR},y_R)$, où $x_j = [a_j,b_j]$, $y =[y_L,y_U]$, $x_{jC}=(a_j + b_j)/2$, $x_{jR}=b_j - a_j$, $y_C = (y_L + y_U)/2$ et $y_r = y_U - y_L$, la valeur $y = [y_L,y_U]$ de Y sera prédite par

$$\hat{y}_L = \hat{y}_C - (1/2)\hat{y}_R,\ \hat{y}_U = \hat{y}_C + (1/2)\hat{y}_R \text{ avec } \hat{y}_C = (\mathbf{x}_C)^T\hat{\boldsymbol{\beta}}_C \text{ et } \hat{y}_R = (\mathbf{x}_R)^T\hat{\boldsymbol{\beta}}_R$$

où $\mathbf{x}_C = (1, x_{1C},\ldots,x_{pC})^T$, $\mathbf{x}_R = (1, x_{1R},\ldots,x_{pR})^T$, $\hat{\boldsymbol{\beta}}_C = (\beta_{0C},\ldots,\beta_{pC})^T$ et $\hat{\boldsymbol{\beta}}_R = (\beta_{0R},\ldots,\beta_{pR})^T$.

2.3 Seconde méthode du centre et de l'étendue (MCE2)

Les exemples sont décrit par $2(p+1)$ variables quantitatives $X_{1C},\ldots,X_{pC}$, Y_C, $X_{1R},\ldots,X_{pR}$, Y_R qui assument comme valeur, respectivement, le centre et l'étendue des intervalles assumés par les variables $X_1,\ldots,X_p$, Y. Chaque exemple $e_i \in E$ est représenté par deux vecteurs de valeurs réelles $\mathbf{z}_{iC}=(x_{i1C},\ldots,x_{ipC}, x_{i1R},\ldots,x_{ipR}, y_{iC})$ et $\mathbf{z}_{iR}=(x_{i1C},\ldots,x_{ipC}, x_{i1R},\ldots,x_{ipR}, y_{iR})$ où $X_{jC}(e_i) = x_{ijC} = (a_{ij} + b_{ij})/2$, $Y_C(e_i) = y_{iC} = (y_{iL} + y_{iU})/2$, $X_{jR}(e_i) = x_{ijR} = b_{ij} - a_{ij}$ et $Y_R(e_i) = y_{iR} = y_{iU} - y_{iL}$. Les modèles correspondant sont donc :

$$\mathbf{y}_C = \mathbf{X}\boldsymbol{\beta}_C + \boldsymbol{\varepsilon}_C \text{ et } \mathbf{y}_R = \mathbf{X}\boldsymbol{\beta}_R + \boldsymbol{\varepsilon}_R$$

où $(\mathbf{y}_C) = (y_{1C},\ldots,y_{nC})^T$, $(\mathbf{y}_R) = (y_{1R},\ldots,y_{nR})^T$, $\mathbf{X} = ((\mathbf{x}_1)^T,\ldots,(\mathbf{x}_n)^T)^T$, avec $\mathbf{x}_i = (1,x_{i1C},\ldots,x_{ipC}, x_{i1R},\ldots,x_{ipR})^T$, $\boldsymbol{\beta}_C = (\beta_{0C},\ldots,\beta_{2pC})^T$, $\boldsymbol{\beta}_R = (\beta_{0R},\ldots,\beta_{2pR})^T$, $\boldsymbol{\varepsilon}_C = (\varepsilon_{1C},\ldots,\varepsilon_{nC})^T$ et $\boldsymbol{\varepsilon}_R = (\varepsilon_{1R},\ldots,\varepsilon_{nR})^T$. L'estimation de $\boldsymbol{\beta}_C$ et $\boldsymbol{\beta}_R$ par la méthode des moindres carrées est donné par

$$\hat{\boldsymbol{\beta}}_C = \left((\mathbf{X})^T(\mathbf{X})\right)^{-1}(\mathbf{X})^T\mathbf{y}_C \text{ et } \hat{\boldsymbol{\beta}}_R = \left((\mathbf{X})^T(\mathbf{X})\right)^{-1}(\mathbf{X})^T\mathbf{y}_R$$

Étant donnée un nouvel exemple *e* décrit par $\mathbf{z} = (x_1,\ldots,x_p,y)$, par $\mathbf{z}_C = (x_{1C},\ldots,x_{pC}, x_{1R},\ldots,x_{pR}, y_C)$ et par $\mathbf{z}_R = (x_{1C},\ldots,x_{pC}, x_{1R},\ldots,x_{pR}, y_R)$, où $x_j = [a_j,b_j]$, $y = [y_L,y_U]$, $x_{jC}=(a_j + b_j)/2$, $x_{jR}=b_j - a_j$, $y_C = (y_L + y_U)/2$ et $y_r = y_U - y_L$, la valeur $y = [y_L,y_U]$ de Y sera prédite par

$$\hat{y}_L = \hat{y}_C - (1/2)\hat{y}_R,\ \hat{y}_U = \hat{y}_C + (1/2)\hat{y}_R \text{ avec } \hat{y}_C = (\mathbf{x})^T\hat{\boldsymbol{\beta}}_C \text{ et } \hat{y}_R = (\mathbf{x})^T\hat{\boldsymbol{\beta}}_R$$

où $\mathbf{x} = (1, x_{1C},\ldots,x_{pC}, x_{1R},\ldots,x_{pR})^T$, $\hat{\boldsymbol{\beta}}_C = (\beta_{0C},\ldots,\beta_{2pC})^T$ et $\hat{\boldsymbol{\beta}}_R = (\beta_{0R},\ldots,\beta_{2pR})^T$.

3 Évaluation des méthodes

Pour l'évaluation de ces méthodes nous considérons ici plusieurs jeux de données synthétiques de type intervalle présentant des différents dégrées de difficultés en ce qui concerne l'ajustement d'un modèle de régression linéaire. Ces données synthétiques sont obtenus de la façon suivante :

a) Les centres des intervalles assumés par les variables indépendantes sont obtenus par tirage aléatoire selon une loi uniforme ;
b) Le centre d'un intervalle assumé par la variable dépendante est supposé être en relation linéaire avec les centres des intervalles des variables indépendantes. Les coefficients et le terme d'erreur du modèle sont obtenu par tirage aléatoire selon une loi uniforme ;
c) Une fois obtenues les centres des intervalles, les étendues correspondantes sont obtenus aussi par tirage aléatoire selon une loi uniforme ;
d) Dans chaque réplication de l'expérience Monte Carlo, le jeu de données est divisé en un ensemble d'apprentissage (250 observations) et un ensemble test (125 observations).

On obtient neuf différentes configurations selon deux facteurs (l'étendue et le terme d'erreur) avec des différents degrés de variabilités : faible, moyenne et forte. Par exemple, une des configurations représente des données de type intervalle avec une forte variabilité de l'étendue et une pauvre relation linéaire entre la variable dépendante et les variables indépendantes du a la forte variabilité du terme d'erreur associé aux centres des intervalles.

La performance de ces modèles (MC, MCE1 et MCE2) est basée sur l'estimation de la moyenne des mesures suivantes : la racine carrée de la moyenne de la somme des carrés des résidus de la limite inférieure ($RMSE_L$), la racine carrée de la moyenne de la somme des carrés des résidus de la limite supérieure ($RMSE_U$), le coefficient de détermination de la limite inférieure (R_L^2) et le coefficient de détermination de la limite supérieure (R_U^2). Ces mesures sont obtenues à partir des valeurs observées $y_i = [y_{iL}, y_{iU}]$ et des valeurs prédites $\hat{y}_i = [\hat{y}_{iL}, \hat{y}_{iU}]$:

$$RMSE_L = \sqrt{\frac{\sum_{i=1}^{n}(y_{iL} - \hat{y}_{iL})^2}{n}} \text{ et } RMSE_U = \sqrt{\frac{\sum_{i=1}^{n}(y_{iU} - \hat{y}_{iU})^2}{n}}$$

$$R_L^2 = \left(\frac{Cov(\mathbf{y}_L, \hat{\mathbf{y}}_L}{S_{\mathbf{Y}_L} S_{\hat{\mathbf{Y}}_L}} \right) \text{ et } R_U^2 = \left(\frac{Cov(\mathbf{y}_U, \hat{\mathbf{y}}_U}{S_{\mathbf{Y}_U} S_{\hat{\mathbf{Y}}_U}} \right)$$

Ces mesures sont estimées pour chaque modèle (MC, MCE1 et MCE2) dans le cadre d'une simulation Monte Carlo avec 100 réplications pour chacune des 9 différentes configurations fixées avec des différents nombres de variables indépendantes. Dans chaque réplication, on ajuste un modèle de régression linéaire sur les données d'apprentissage selon chaque modèle (MC, MCE1 et MCE2) qui ensuite est utilisé pour prédire la valeur de la variable dépendante sur l'ensemble test et on calcule les mesures $RMSE_L$, $RMSE_U$, R_L^2 et R_U^2. Pour chacune de ces mesures, on calcule leur moyenne et l'écart type sur les 100 réplications de la simulation Monte Carlo et la performance moyenne des différentes modèles est comparée selon un test statistique standard de différence de moyennes. En outre, toute cette procédure est répétée considérant 100 différentes valeurs pour le vecteur de paramètres β qui lie le centre des intervalles assumés par la variable dépendante aux centres des intervalles assumés par les variables indépendantes.

D'une façon générale, les tests statistiques indiquent, au risque de 1%, que la performance moyenne (selon les mesures $RMSE_L$, $RMSE_U$, R_L^2 et R_U^2) des méthodes MCE1 et MCE2 est supérieure à celles de la méthode MC. Aussi, la supériorité des méthodes MCE1 et MCE2 par rapport à la méthode MC est encore plus évidente quand le nombre de variables indépendantes présentes dans le modèle augmente.

Le dégrée de variabilité de l'étendue des intervalles étant fixée, plus le dégrée de variabilité du terme d'erreur du modèle linéaire qui lie le centre des intervalles de la variable dépendante aux centres des intervalles des variables indépendantes diminue plus la performance moyenne des méthodes MCE1 et MCE2 est supérieure à celle de la méthode MC. C'est-à-dire, plus la relation linéaire entre les variables est riche plus la performance moyenne des méthodes MCE1 et MCE2 est supérieure à celle de la méthode MC.

De l'autre coté, le dégrée de variabilité du terme d'erreur du modèle linéaire qui lie le centre des intervalles de la variable dépendante aux centres des intervalles des variables indépendantes étant fixée, plus le dégrée de variabilité de l'étendue des intervalles diminue, plus est semblable la performance moyenne entre les méthodes MC, MCE1 et MCE2. C'est un résultat attendu car plus l'étendue des intervalles s'approche de zéro, plus la méthode MC dévient un cas particulier des méthodes MCE1 et MCE2.

Enfin, indépendamment du nombre de variables indépendantes présentes, les résultats obtenues montrent qu'il n'y a pas de différence statistiquement significative entre la performance moyenne des méthodes MCE1 et MCE2. Par conséquent, on va preferer la méthode MCE1 car le nombre de paramètres à être estimé dans le cadre du modèle MCE2 est presque le double de celui à être estimé dans le cadre du modèle MCE1.

4 Bibliographie

[BIL 02] BILLARD, L., DIDAY, E., "Regression Analysis for Interval-Valued Data", *Data Analysis, Classification and Related Methods: Proceedings of the Seventh Conference of the International Federation of Classification Societies, IFCS-2000*, Namur (Belgium), Kiers, H.A.L. et al. Eds, 2000, p. 369—374, Springer, Berlin Heidelberg.

[BIL 03] BILLARD, L., DIDAY, E., "From the Statistics of Data to the Statistics of Knowledge: Symbolic Data Analysis", *Journal of the American Statistical Association*, vol. 98, 2003, p. 470-487.

[BOC 00] BOCK H-H., DIDAY, E., *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data*, Springer, 2000.

[CAR 04] DE CARVALHO, F. A. T., LIMA NETO, E. A., TENÓRIO, C. P. "A New Method to Fit a Linear Regression Model for Interval-Valued Data", *Advances in Artificial Intelligence: Proceedings of the 27th German Conference on Artificial Intelligence - KI'2004. Lectures Notes on Artificial Intelligence, LNAI 3238*. Ulm (Germany), 2004, p.295 – 306, Springer, Berlin Heidelberg.

Une description symbolique minimisant l'inertie et l'impureté

Mohamed Mehdi Limam[1], Edwin Diday[1] et Suzanne Winsberg[2]
[1] LISE-CEREMADE, [2] IRCAM

Résumé

Notre objectif est de trouver une partition d'une classe prise dans une population donnée, chaque classe de cette partition est décrite par une conjonction de propriétés caractéristiques des variables descriptives des individus. Cette méthode utilise en entrée un tableau de données dont les cases contiennent des histogrammes ou des intervalles.

1 Introduction

Ici, notre but est de produire une description d'une classe qui induit une partition de cette classe. Cette partition satisfait simultanément un critère d'homogénéité intra-classe et un critère de discrimination par rapport à une partition à priori. Notre méthode possède donc un aspect supervisé et un aspect non supervisé. En plus, on doit pouvoir générer une description qui minimise le débordement c'est à dire qui décrit le moins possible d'individus n'appartenant pas à la classe à décrire qu'on appelera C. Dans le cas classique par exemple, à partir d'un tableau de donnée contenant des variables descriptives comme le niveau d'étude ou l'activité sportive, on extrait une classe à décrire, C, qui peut être les jeunes dont l'âge est compris entre 15 et 25 ans et une partition a priori qui peut être les fumeurs et les non fumeurs. On cherche à trouver une description de C qui induit une partition de C constituée de classes homogènes de jeunes, chacune d'entre elles bien discriminée par rapport au fumeurs et aux non fumeurs. Chaque classe de la partition recherchée est décrite par une conjonction de propriétés caractéristiques des variables descriptives des individus.

Notre méthode est basée sur une approche divisive monothétique comme CART ou ID3 (voir [1], [2]). Cette méthode utilise en entrée dans sa première version réalisée par [8] un tableau de données dont chacune des cases contiennent un histogramme. On appelle histogramme une distribution de fréquences ou de probabilités sur un ensemble de modalités qui peuvent être ordonnées ou pas. Ici, on présentera une version plus complète permettant de traiter un tableau dont chacune des cases contient des histogrammes ainsi que des intervalles. De plus, on proposera une méthode permettant d'adapter automatiquement le mélange des critères d'homogénéité et de discrimination. Naturellement, les

données classiques représentent un cas particulier des données de type histogramme et de type intervalle considérées ici. Ainsi, cette procédure fonctionne avec des données classiques qualitatives et quantitatives. D'autres algorithmes divisifs pour données symboliques utilisent soit un critère d'homogénéité, soit un critère de discrimination basé sur une partition a priori mais non les deux simultanément. [3] a proposé une méthode pour apprentissage non supervisé, par contre [6], et [7] ont proposé des méthodes pour l'apprentissage supervisé.

2 La méthode

Cinq entrées sont demandées pour cette méthode : 1) les données, constituées de n unités statistiques, chacune décrite par K variables symboliques ; 2) la partition a priori en plusieurs classes ; 3) la classe, C à décrire ; 4) le nombre de noeuds terminaux qui correspond à la règle d'arrêt et 5) un coefficient qui donne plus ou moins d'importance au pouvoir discriminant de la partition à priori ou à l'homogénéité de la description de la classe C donnée.

La méthode utilise une approche descendante hiérarchique divisant un ensemble en deux noeuds fils. A chaque étape k (k noeuds correspondants à une partition en k+1 classes), un des noeuds terminaux de l'arbre est divisé en deux noeuds fils (ou feuilles) dans le but d'optimiser un critère d'évaluation Q de la partition en k+1 classes. La division d'un noeud N en deux noeuds N_1(noeud gauche) et N_2 (noeud droit) est réalisée par une "coupure", avec y la variable de coupure et c la valeur de coupure.

L'algorithme génère deux types de sorties. Le premier est une représentation graphique, dans laquelle la classe à décrire, C, est représentée par un arbre binaire. Les noeuds terminaux représentent les classes obtenues après partition de la classe C, et chaque branche représente une coupure (y, c). Le deuxième est une description : chaque noeud terminal est décrit par la conjonction des coupures qui le génèrent. La classe, C, est alors décrite par la disjonction de ces conjonctions.

Le critère d'homogénéité utilisé est le critère d'inertie et le critère de discrimination est l'indice de Gini (voir [8]). Soient *H(N)* et *h(N_1,N_2)* respectivement le critère d'homogénéité du noeud C et le critère d'homogénéité du couple de noeud *(N_1, N_2)* avec $h(N_1, N_2) = H(N_1) + H(N_2)$. On défini alors $\Delta H(N) = H(N) - h(N_1, N_2)$. Soient *D(N)* et *d($N_1$,$N_2$)* respectivement le critère de discrimination du noeud N et le critère de discrimination du couple de noeuds *(N_1, N_2)* avec $d(N_1, N_2) = p_1 D(N_1) + p_2 D(N_2)$ et $p_i = card(N_i)/card(N)$. Comme précédemment, on défini $\Delta D(N) = D(N) - d(N_1, N_2)$. La qualité Q (respectivement q) d'un noeud N (respectivement du couple de noeuds (N_1, N_2)) est la somme pondérée des deux précédents critères : $Q(N) = \alpha H(N) + \beta D(N)$ (respectivement $q(N_1, N_2) = \alpha h(N_1, N_2) + \beta d(N_1, N_2)$) avec $\alpha + \beta = 1$. Donc, la variation de qualité induite par cette partition de N en (N_1, N_2) est $\Delta Q(N) = Q(N) - q(N_1, N_2)$. Ici, on maximise $\Delta Q(N)$. On note que puisqu'on est en train d'optimiser les deux critères, les critère doit être normalisé. L'utilisateur peut changer les valeurs des coefficients α et β selon l'importance qu'il

veut donner à chaque critère. Pour déterminer la coupure (y ;c) et le noeud à couper, il faut : d'abord, pour chaque noeud N de cardinalité n, sélectionner la variable et la valeur de coupure qui minimise $q(N_1, N_2)$; Ensuite, sélectionner et couper le noeud N qui maximise la différence entre la qualité avant coupure et la qualité après coupure, $max\Delta Q(N) = max[\alpha\Delta H(N) + \beta\Delta D(N)]$.

Nous rappelons que nous somme en train de traiter des variables à valeurs histogrammes. Pour cela, nous avons à définir les coupures pour ce type de données et les valeurs de coupure. L'idée est que la valeurs de coupure d'une variable à valeurs histogrammes est définie sur la valeur de la fréquence d'une seule modalité, sur la valeur de la somme des fréquences de deux modalités ou sur la valeur de la somme de plusieurs modalités. On commence par ordonner les individus dans l'ordre croissant en utilisant la valeur de la fréquence d'une modalité, On pourra alors trouver $n-1$ valeurs de coupure c entre deux valeurs consécutives et donc $n-1$ partitions en deux classes. On refait cette opération pour chaque modalité. De même, on ordonne les individus dans l'ordre croissant de la somme des valeurs des fréquences de deux modalités. On peut aussi remarquer qu'on a $2^{M-1}-1$ manières de trier les individus dans l'ordre croissant qui représente le nombre de sous-classes non vides d'un ensemble de M modalités. Pour le cas ordinal, on aura $M-1$ manières de trier les individus dans l'ordre croissant qui représente le nombre de division en deux ensembles non vides d'un ensemble de M modalités. D'où, pour une variable à M modalités, on a au maximum $(2^{M-1}-1)(n-1)$ partitions pour le cas nominal et $(M-1)(n-1)$ partitions pour le cas ordinal.

Pour une variable intervalle, notons $\{b_i\}_{i=1,...,2.n}$ l'ensemble ordonné des bornes distinctes (inférieurs et supérieurs) pour toutes les valeurs des individus sur la variable intervalle. On peut alors définir les $n-1$ valeurs de coupure distinctes c_i $(i = 1, ..., n-1)$ en posant $c_i = \frac{b_i+b_{i+1}}{2}$. On obtient alors au maximum $2.(n-1)$ valeurs de coupures ou questions binaires dans ce cas.

Nous allons nous intéresser maintenant au choix du coefficient α. Pour cela, on doit avant tout fixer le nombre de noeud terminaux. L'influence du coefficient α peut être déterminant dans la construction de l'arbre et dans la qualité de prédiction. La variation de α (ou de β puisque $\alpha + \beta = 1$) de 0 à 1 augmente l'importance de l'homogénéité et diminue l'importance de la discrimination. Ces variations influent sur les coupures et par conséquent sur les noeuds terminaux. Nous avons besoins de critères qui mesurent la qualité de l'arbre : on calcule l'inertie totale (I) pour mesurer l'homogénéité et le taux de mauvais classement (TMC) pour la discrimination (voir [4]). Alors, on peut déterminer la valeur de α qui optimise simultanément I et le TMC c'est-à-dire l'homogénéité et la discrimination simultanément. Si au contraire l'utilisateur fixe la valeur de $\alpha = 0$, considérant seulement le critère de discrimination, et si en plus les données sont classiques, l'algorithme fonctionne comme CART. Donc CART est un cas particulier de cet algorithme.

L'idée est de construire pour chaque valeur de α plusieurs arbres à partir de plusieurs échantillons et de calculer I et TMC pour chaque arbre. Partant de l'ensemble de départ de n individus, on extrait B échantillon par bootstrap de taille n (par tirage avec remise). Pour chaque échantillon et pour chaque valeur

de α entre 0 et 1, on construit un arbre et on calcule les deux paramètres (I et TMC). En variant α entre 0 et 1 (avec un pas de 0.1 par exemple) donne 11 couples de valeurs de I et TMC correspondant au moyennes de ces paramètres pour les B échantillons. Afin de visualiser la variation des deux paramètres, on sort une courbe représentant l'I et une deuxième représentant le TMC en fonction de α. La valeur optimale de α est celle qui minimise la somme normalisée des deux paramètres. Cette valeur est donnée automatiquement par le programme.

3 Conclusion

L'idée principale de la méthode est de mixer un critère d'homogénéité et un critère de discrimination par rapport à une partition a priori pour décrire un ensemble . L'ensemble à décrire peut être une classe de la partition a priori, l'ensemble de la population ou une classe de la population. Ayant choisi cette classe, l'intérêt de la méthode est que l'utilisateur puisse choisir les poids α et $\beta = 1 - \alpha$ des critères d'inertie et de discrimination respectivement, ce choix dépend de l'importance de ces deux critères pour atteindre le but fixé. Sinon, l'utilisateur peut optimiser les deux critères simultanément, en choisissant la valeur de α trouvée automatiquement. Nous avons présenté dans cet article le traitement des données histogrammes et intervalles. Le traitement d'autre types de variables comme les variables taxonomiques est présenté dans [4]. D'autres améliorations comme les affectations de type flou des individus aux noeuds de l'arbre ont été éfféctuées (voir [5]).

Références

[1] Breiman, L., Friedman, J.H., Olhsen, R.A., and Stone, C.J. (1984) : Classification and Regression Trees. Wadsworth, Belmont, California.

[2] Quinlan, J.R (1986) : Induction of Decision Trees. Machine Learning, 1, 81-106.

[3] Chavent, M. (1997) : Analyse de données symboliques, Une Méthode Divisive de Classification. Thèse de Doctorat, Université Paris IX Dauphine.

[4] Limam, M., Diday, E. et Winsberg, S. (2003) : Symbolic Class Description with Interval Data. Journal of Symbolic Data Analysis, Vol. 1.

[5] Limam, M., Diday, E. et Winsberg, S. (2004) : Probabalistic allocation for Symbolic Class Description. Proceedings IFCS'2004. Chicago.

[6] Périnel, E. (1999) : Construire un arbre de discrimination binaire à partir de données imprécises. Revue de statistique Appliquée, 47, 5-30.

[7] Gettler-Summa, M. (1999) : Marking and Generalisation by Symbolic Objects in the Symbolic Official Data Analysis Software. CEREMADE, France.

[8] Vrac, M. et Diday, E. (2001) : Description Symbolique de Classes, Cahiers du CEREMADE, Paris, France.

Classification de courbes par apprentissage

J-M.Loubes, O.Roudenko, M.Sebag et O.Wintenberger*

** : SAMOS, Université Paris 1 Panthéon-Sorbonne,*
72 rue Régnault
75013 PARIS, FRANCE

RÉSUMÉ

Nous souhaitons dans ce papier résumer l'information apportée par un grand nombre de données, ici les courbes d'évolution journalière de la vitesse moyenne des voitures observées sur une portion de route pendant deux ans. Nous utilisons une nouvelle méthode de classification non supervisée afin de regrouper les observations en classes cohérentes. Une méthode classique consiste à choisir les classes présentes à un même niveau dans une classification ascendante hiérarchique (CAH), puis de s'entraîner sur un échantillon distinct pour atteindre le niveau optimal. A partir de la CAH, nous effectuons une optimisation multicritère pour sélectionner de manière plus systématique les classes recouvrant le plus de courbes ayant le plus de similarités. L'algorithme atteint de meilleurs résultats que la méthode classique car il prend en considération les asymétries possibles dans la CAH.

MOTS-CLÉS : Classification, optimisation multicritères.

1 Introduction

Nous souhaitons résumer l'information d'un grand nombre d'observations grâce à une nouvelle méthode de classification non supervisée. Nous procédons classiquement en trois étapes :

- Construire une CAH utilisant une distance spécifique aux données puis sélectionner les classes recouvrant le plus de courbes ayant le plus de similarités. Cette étape est appelée l'étape d'initialisation.
- Utiliser un échantillon de courbes distinctes pour sélectionner parmi ces ensembles de classes le meilleur recouvrement. C'est l'étape d'apprentissage.
- Enfin comparer les résultats obtenus avec la méthode classique sur un troisième échantillon test. C'est l'étape de validation.

2 Classification

2.1 Présentations des données

Nous observons la vitesse moyenne $X_j(t_k)$ des véhicules sur une portion de route toutes les 6 minutes aux instants t_k, $k = 1,...,d$ et chaque jour $j=1,...,J$ pendant deux ans. Nous scindons les courbes d'évolution journalière de la vitesse $X_j \in \Re^d$ en trois groupes :

- Les n premières observations X_j, $j=1,...,n$, seront utilisées dans la CAH.
- Les N courbes suivantes, notées Y_j, $j=1,...,N$, constituent l'échantillon d'apprentissage et seront utilisées pour sélectionner le meilleur recouvrement.

- Les courbes restantes, notées Z_j, j=1,...,T seront utilisées pour tester les performances de notre algorithme.

2.2 Choix de la distance de classification

Nous avons choisi une distance spécifique à la nature de nos données afin de construire la CAH. Pour $x, y \in \Re^d$, nous utilisons la distance Δ définie par $\Delta(x,y)=\sqrt{{}^t(x-y)W(x-y)}$ où W est la matrice d par d définie par $W_{i,j} = \dfrac{d-|i-j|}{d}$ pour tout i et j$\in$ {1,…,d}. Cette distance prend en compte la dimension temporelle de l'évolution de la vitesse des véhicules. Ainsi, les phénomènes similaires seront ordonnés selon les délais entre leurs moments d'apparition :

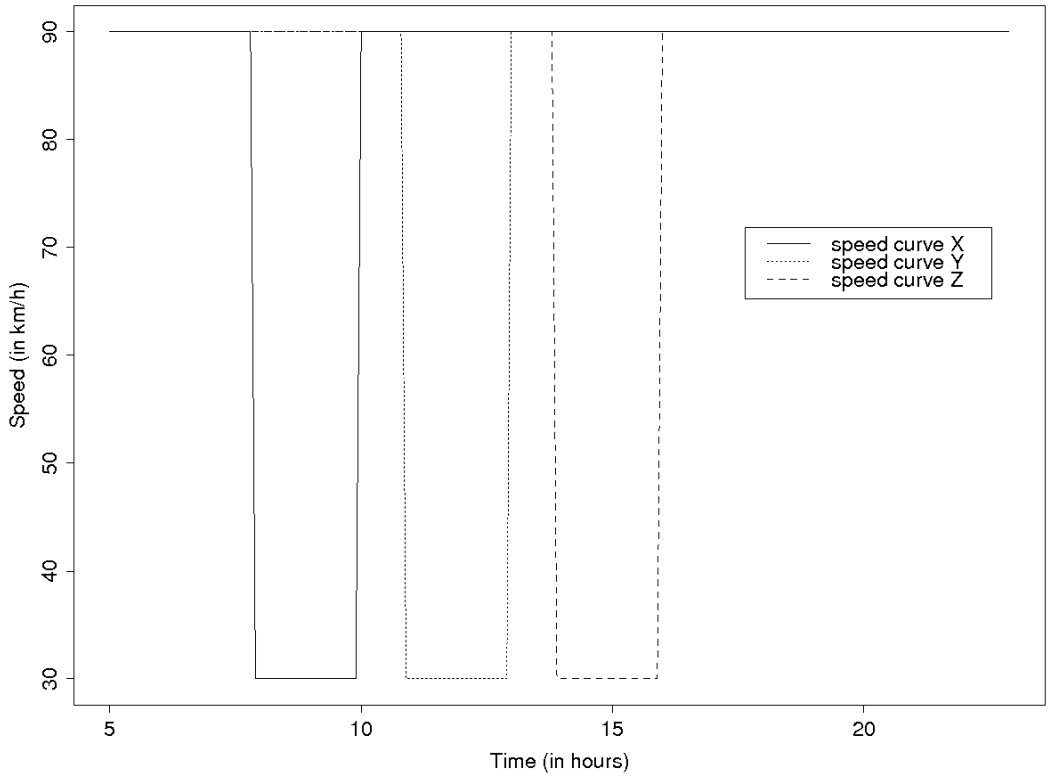


Figure 1

Nous avons $\Delta(X,Y)=\Delta(Y,Z)=637$ et $\Delta(X,Z)=967$ alors que la distance euclidienne ne fait aucune distinction entre ces 3 courbes. Les pics de vitesses représentent dans notre cadre d'étude les bouchons. Notre distance permet de faire une distinction suivant l'heure à laquelle ces phénomènes ont lieu.

2.3 Classification Ascendante Hiérarchique

Nous construisons une CAH en utilisant notre distance Δ pour notre sous-échantillon X_j, j=1,...,n. Pour cela, nous suivons l'algorithme de Johnson, la distance entre deux classes étant fixée comme la plus grande distance des éléments de ces classes. Nous observons une forte asymétrie dans la CAH obtenue :

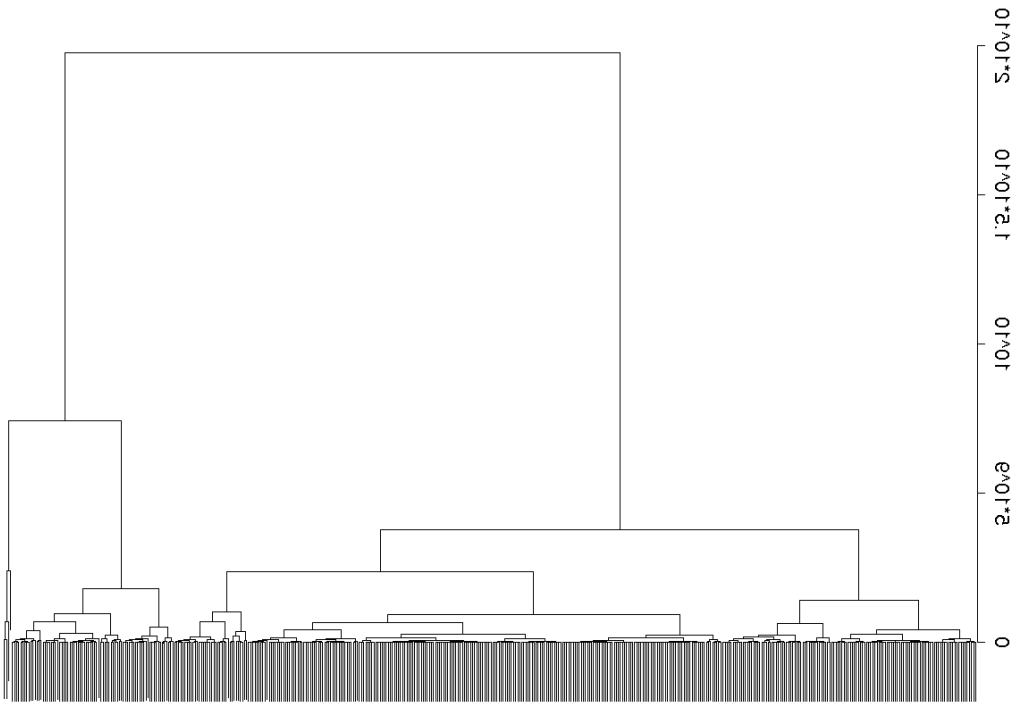


Figure 2

Nous formons classiquement les classes C_j, $j \in J$ de courbes en regroupant les feuilles descendantes d'un même noeud. Nous résumons l'information de chaque classe C_j grâce à un représentant (appelé archétype) : $\overline{X}_j = \arg\min_{X \in C_j} \sum_{X' \in C_j} \Delta(X, X')$, qui est la courbe la plus proche du barycentre du groupe. D'autres choix sont possibles (médiane…).

3 Optimisation et apprentissage

3.1 Optimisation multicritères

Nous supposons que l'ensemble des archétypes résume convenablement l'information apportée par les observations. Mais cet ensemble est grand, le nombre de classes J de l'arbre étant élevé. Nous allons rechercher à conserver le maximum d'information dans le plus petit sous-ensemble d'archétypes possible. Pour cela, nous nous entraînons sur l'échantillon d'apprentissage en nous ramenant à un programme d'optimisation multicritères. Nous devons trouver l'ensemble d'indices $\Lambda \subset J$ qui minimise

- La précision : $-\frac{1}{N}\sum_{i=1}^{N} \min_{j \in \Lambda} \Delta(Y_i, \overline{X}_j)$
- La généralité : $|\Lambda|$

3.2 Présentation succincte de l'algorithme

Nous ne pouvons pas nous permettre de tester la totalité des sous-ensembles Λ possibles. Un algorithme nous permet d'en faire une sélection. Il consiste à itérer deux étapes jusqu'à obtenir un recouvrement convenable des courbes ($\bigcup_{j\in\Lambda} C_j \approx \{ X_j , j=1,...,n \}$):

- Choix d'une feuille : nous choisissons une courbe notée F parmi X_j , $j=1,...,n$ qui maximise les critères d'une pondération.
- Choix d'une classe : nous choisissons dans l'arborescence de la feuille F le nœud (et donc la classe C) qui minimise le critère : $(1-\mu)\Delta(\overline{X},F)-\mu|C|$, où $\overline{X}$ est l'archétype de C. $\overline{X}$ est alors rajouté à notre sous-ensemble Λ. Puis nous mettons à jour la pondération de l'ensemble des feuilles.

La pondération est primordiale, elle permet en particulier de sélectionner à la première étape des feuilles qui sont éloignées de tous les archétypes déjà sélectionnés, tout en évitant de ne sélectionner que les outliers.

3.3 Apprentissage

L'algorithme dépend essentiellement du paramètre μ : à μ fixé, nous sélectionnons toujours les mêmes archétypes. Nous faisons alors varier notre paramètre jusqu'à obtenir un Λ_{opt} tel qu'il n'existe pas d'autres candidats qui décrivent plus précisément l'échantillon d'apprentissage : $\Lambda_{opt} = \arg\min_{\mu} \frac{1}{N}\sum_{i=1}^{N} \min_{j\in\Lambda} \Delta(Y_i, \overline{X}_j)$.

3.4 Comparaison avec la méthode classique

Habituellement, le choix du meilleur ensemble de classes dans la CAH se fait en coupant l'arbre au niveau optimal et en gardant les classes correspondantes. Ainsi, seuls les sous ensembles constitués de nœuds (et donc de classes) d'un même niveau dans la CAH sont considérés. Le problème de cette méthode est que toutes les branches de l'arbre sont traitées de la même manière.

Nous comparons cette méthode classique avec la notre grâce à l'échantillon test. Le tableau ci-dessous présente la moyenne, l'écart type, le minimum et le maximum des quantités $\min_{j\in\Delta}\Delta(Z_i,\overline{X}_j)$:

Méthodes	Moyenne	Ecart type	Min.	Max.
Niveau de coupe fixe	0.43	5.0	18	28
Multiobjectif	0.34	4.2	18	19

Tableau 1

Nous obtenons de meilleurs résultats que la méthode classique. Nous appliquons notre méthode au trafic routier. A une date T et une heure H, nous disposons de l'historique du trafic aux dates 1,…,T-1. Nous obtenons un résumé optimal de cet ensemble de courbes grâce à notre algorithme. Puis nous estimons la vitesse en H+1 en choisissant l'archétype le plus proche aux heures 0,…,H de la courbe en T et en prenant la vitesse de cet archétype en H+1.

4 Bibliographie

[GLM 04] GAMBOA F., LOUBES J.M., MAZA E., *Structural estimation for high dimensional data*, submitted to Annals of Statistics, 2004.

[LLM 04] LAVIELLE M., LOUBES J.M., and MAZA E , *Classification and forecasting in travel time*, submitted to Canadian Journal of Statistics, 2004.

Comparaison des critères de Kolmogorov-Smirnov, de Gini et de l'entropie sur des données de type intervalle

Chérif Mballo [1,2] **& Edwin Diday** [2]

1- ESIEA Recherche, Pole E.C.D
38, Rue des Docteurs Calmette et Guérin
53000 Laval. France
mballo@esiea-ouest.fr

2 - LISE-CEREMADE,
Université Paris Dauphine, Place du Maréchal de Lattre de Tassigny 75775 Paris cedex 16 France
diday@ceremade.dauphine.fr

RÉSUMÉ. Le critère de découpage binaire de Kolmogorov-Smirnov (Friedman 1977) nécessite un ordre total des valeurs prises par les variables explicatives. Nous pouvons ordonner des intervalles fermés bornés de $\Re$ (ensemble des nombres réels) de différentes façons. Les relations d'ordre sur des données de type intervalle nous ont permis d'étendre le critère de Kolmogorov-Smirnov à ce type de données. Nous comparons ce critère avec les critères de Gini et de l'entropie. Pour ce faire, nous axons notre attention sur le taux d'erreur mesuré sur l'échantillon de test.

MOTS-CLÉS : Critère de Kolmogorov-Smirnov, arbre de décision, ordre sur des intervalles..

1 Introduction

Le critère de découpage binaire de Kolmogorov-Smirnov ([FRI 77] ; [UTG 96]) a été introduit pour une partition binaire à expliquer avec des variables explicatives quantitatives classiques. Il a été étendu aux variables qualitatives classiques par ([ASS 98]). Il nécessite un ordre total des valeurs prises par les variables explicatives. Nous pouvons ordonner des intervalles ([BOC 00]) fermés bornés de $\Re$ (ensemble des nombres réels) de différentes façons ([DID 03]) : la borne supérieure, la borne inférieure, la moyenne (centre) et la longueur (étendue). Nous proposons une approche exploratoire consistant à construire un arbre pour chaque ordre. Nous comparons ce critère avec les critères de Gini et de l'entropie.

2 Présentation du tableau de données en entrée

Soit Ω un ensemble constitué d'objets (individus) destinés à être classés à l'aide d'un arbre de décision. Chaque objet est décrit par (p + 1) variables : p variables de type intervalle $X_1,...,X_p$ (variables explicatives) et une variable classe Y (variable à expliquer). Au départ du processus de construction d'un arbre de décision par le critère de Kolmogorov-Smirnov, notre tableau de données en entrée se présente alors comme le tableau (Tableau 1).

Variables / *Individus* (*objets*)	X_1	……	X_p	Y
Objet_1	$[a_1,b_1]$	……..	$[c_1,d_1]$	i
…….	…….	…….		…..
Objet_n	$[a_n,b_n]$	…….	$[c_n,d_n]$	j

Tableau 1: Format du tableau en entrée

3 Le critère de découpage binaire de Kolmogorov-Smirnov pour la segmentation

Pour utiliser le critère de Kolmogorov-Smirnov (noté KS dans la suite), Friedman ([FRI 77]) suppose que les coûts de mauvais classement et les probabilités a priori des classes sont inversement proportionnels. Ce critère permet de séparer une population en deux sous populations plus homogènes en se basant sur les deux fonctions de répartition des classes a priori (cas de deux classes) pour chaque variable explicative. Dans le cas où le nombre de classes a priori est strictement supérieur à 2, les fonctions de répartition sont induites par le regroupement de ces k classes en deux groupes appelés super classes par la méthode « *twoing splitting process* » ([BRE 84]). Il y a $(2^{k-1}-1)$ possibilités de regrouper k classes en deux groupes, mais cette complexité exponentielle a été réduite à une complexité polynomiale par ([ASS 98]). Cette méthode « *twoing splitting process* » est utilisée pour générer deux super classes G_1 et G_2 auxquelles sont associées deux fonctions de répartitions F_1 et F_2 d'une variable aléatoire (variable explicative). Soit D_{X_j} le domaine (ensemble fini d'intervalles fermés bornés de $\Re$) d'une variable explicative X_j. La fonction de répartition a pour rôle de compter toutes les valeurs inférieures à un certain seuil. Les deux fonctions de répartition théoriques F_1 et F_2 ne sont pas connues en pratique. S'il n'y a pas d'ordre sur les observations, on ne peut pas estimer la fonction de répartition théorique F_i par la fonction de répartition empirique $\hat{F}_i$ (i = 1,2) comme dans le cas continu. Dans notre cas, nous pouvons estimer cette fonction de répartition théorique car nous avons un ordre des intervalles fermés bornés de $\Re$ et l'ensemble $\{y \in D_{X_j} / y \le x\} \cap \{y \in D_{X_j} / y \in G_i\}$ est toujours fini en pratique ($j = 1,2,...,p$; $i = 1,2$ et « ≤ » un ordre d'intervalles). Selon l'ordre choisi pour ordonner les observations d'une variable explicative X_j, la fonction de répartition empirique $\hat{F}_i^j$ qui estime F_i^j en $x \in D_{X_j}$ est donnée par : $\hat{F}_i^j(x) = \frac{Cardinal\left(\{y \in D_{X_j} / y \le x\} \cap \{y \in D_{X_j} / y \in G_i\}\right)}{Cardinal\left(\{y \in D_{X_j} / y \in G_i\}\right)}$. Ce sont les proportions réelles des observations pour chaque variable explicative X_j relative à une classe a priori (ou super classe) G_i. Ainsi, le critère KS est défini par : $KS = \sup_{x \in D_{X_j}} \left| \hat{F}_1^j(x) - \hat{F}_2^j(x) \right| \ \forall$ j = 1,..,*p*.

C'est une extension naturelle du critère KS ([MBA 04]), seulement l'argument sélectionné pour le seuil de coupure est ici un intervalle et non un réel comme dans le cas classique. On peut donc utiliser toutes les autres étapes (communes à tout type de variable) pour construire l'arbre de décision. Les auteurs ayant construit des arbres de décision sur des variables de type intervalle ([PER 96]) avec les critères classiques (Gini, likelihood et gain ratio) prennent le centre de l'intervalle comme seuil de coupure.

4 Comparaison des critères de Kolmogorov-Smirnov, de Gini et de l'entropie

Comme il y a plusieurs façons d'ordonner des intervalles, l'approche que nous proposons consiste à construire un arbre pour chaque ordre. Le but de cette partie est de comparer les critères de KS, de Gini et de l'entropie pour chaque ordre d'intervalles selon le taux d'erreur réel $\hat{R}_r$ mesuré sur l'échantillon de test. Les fichiers[1] que nous utilisons (Tableau 2) sont obtenus à partir de bases de données classiques à l'aide du module DB2SO (*Data Base two Symbolic Objects*) du logiciel libre Sodas[2]. En fait nos individus (ou objets) ne sont pas des individus au sens classique du terme (individus de premier ordre), mais des

[1] Fichiers Sodas disponibles à l'URL http://www.ceremade.dauphine.fr/~touati/exemples.htm
[2] disponible à l'URL http://www.ceremade.dauphine.fr/%7Etouati/sodas-pagegarde.htm

concepts (individus de second ordre). Au tableau (Tableau 2), la colonne « *Nb_cl* » indique le nombre de classes a priori de la variable à expliquer (variable nominale) et la colonne « *Répartition par classe a priori* » donne la répartition des objets par classe a priori. Par exemple la notation (20 ;10) indique que vingt objets sont de la classe « 1 » et dix de la classe « 2 » pour une variable nominale ayant deux modalités. La dernière colonne « *Nb_var* » indique le nombre de variables explicatives. Comme les échantillons de test et d'apprentissage sont indépendants et pris aléatoirement dans le fichier de données, la précision de l'estimation ne dépend que du nombre d'objets de l'ensemble de test et de la valeur du risque réel ([COR 02]). Le nombre minimum d'individus pour une feuille est fixé à cinq pour les fichiers F_1 à F_6, dix pour les fichiers F_7 à F_8, quinze pour F_9 et cent pour F_10. Les figures (Figure 1, Figure 2, Figure 3 et Figure 4) présentent les résultats obtenus.

Numéro	Nom du fichier	Taille du fichier	Nb_cl	Répartition par classe a priori	Nb_var
F_1	Wine	23	9	(2 ;2 ;2 ;2 ;6 ;5 ;2 ;1 ;1)	21
F_2	Joueur	29	2	(16 ;13)	7
F_3	Wave	30	3	(10 ;10 ;10)	21
F_4	Auto	33	4	(10 ;8 ;8 ;7)	8
F_5	Accident	48	3	(36 ;9 ;3)	5
F_6	Shuttle	102	7	(78 ;1 ;1 ;15 ;5 ;1 ;1)	9
F_7	Cholesterol	193	2	(99 ;94)	2
F_8	Age_color	231	4	(50 ;51 ;82 ;48)	4
F_9	Profession/size	720	5	(36 ;441 ;32 ;35 ;176)	1
F_10	Regions_NU	10000	4	(1900 ;2300 ;3620 ;2180)	4

Tableau 2. Inventaire des fichiers utilisés

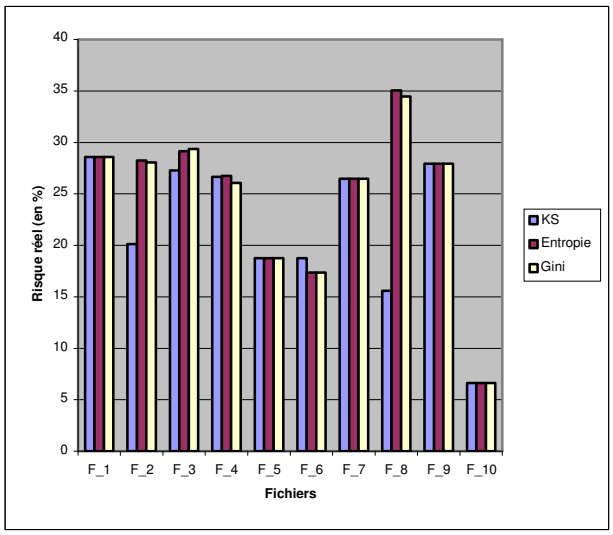


Figure 1. Résultats obtenus en ordonnant les intervalles par la borne supérieure

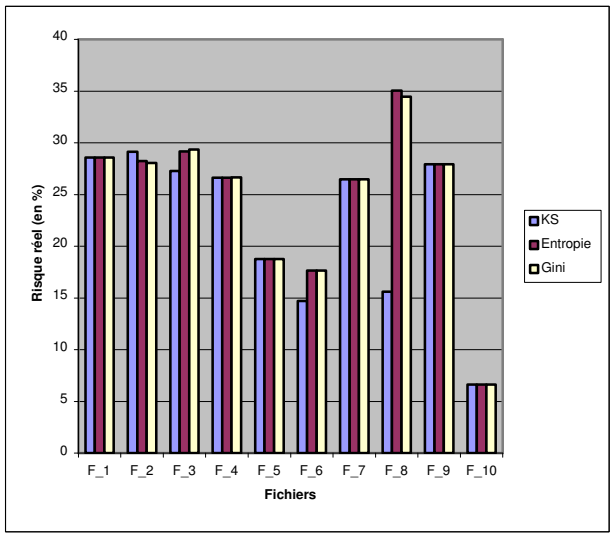


Figure 2. Résultats obtenus en ordonnant les intervalles par la borne inférieure

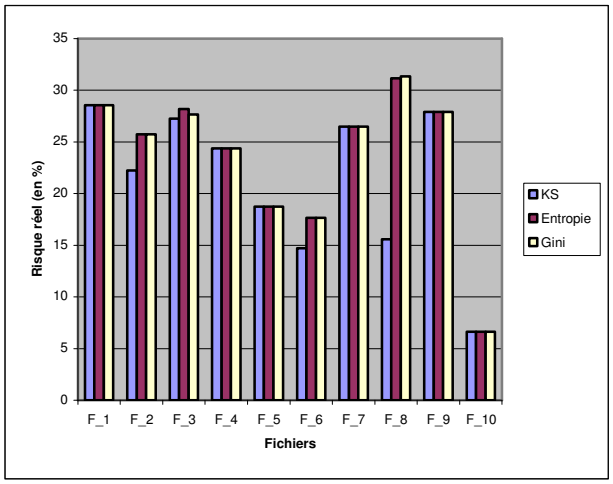


Figure 3. Résultats obtenus en ordonnant les intervalles par la moyenne

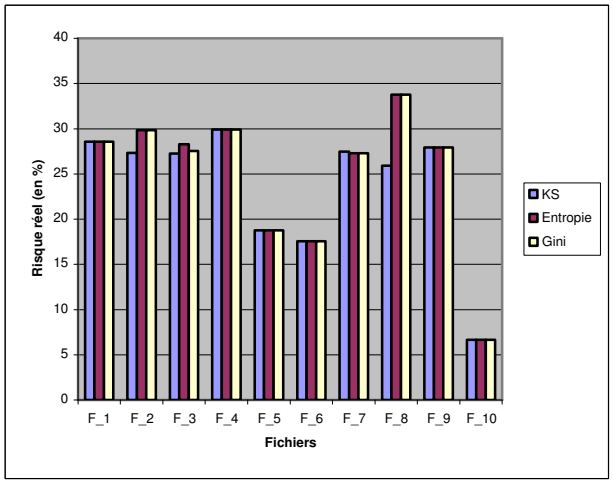


Figure 4. Résultats obtenus en ordonnant les intervalles par la longueur

Ces résultats nous montrent des risques réels très similaires d'un critère à un autre dans le cas général. Cependant on note quelques disparités : pour toutes les figures, le risque réel associé au critère KS pour le fichier F_8 est faible tandis qu'il reste très élevé pour les critères de Gini et de l'entropie sur ce même fichier. Cette même remarque est observée pour les fichiers F_2 (figure 1, figure 3 et figure 4) et F_6 (figure 2 et figure 3).

5 Bibliographie

[ASS 98] ASSERAF M., Extension et optimisation pour la segmentation de la distance de Kolmogorov-Smirnon , *Thèse de Doctorat, Mathématiques Appliquées, Université Paris Dauphine* , 1998.

[BOC 00] BOCK, H. H. & DIDAY, E., *Analysis of symbolic data : Exploratory methods for extracting statistical information from complex data*, Springer-Verlag, Berlin-Heidelberg, 2000.

[BRE 84] BREIMAN L., FRIEDMAN J., OLSHEN R., STONE C., *Classification and regression trees*, Chapman & Hall, 1984.

[COR 02] CORNUEJOLS A. & MICLET L., *Apprentissage Artificiel : concepts et algorithmes*. Eyrolles, 2002.

[DID 03] DIDAY E., GIOIA F., MBALLO C., Codage qualitatif d'une variable intervalle, *XXXV*^{ième} *Journées de Statistique*, pp 415-418, Lyon, France, Juin 2003.

[FRI 77] Friedman J. H., A recursive partitioning decision rule for non parametric classification; *IEEE Transactions on Computers*, C-26, pp 404-408, 1977.

[MBA 04] MBALLO C., ASSERAF M. et DIDAY E., Binary decision trees for interval and taxonomic variables ; A *Statistical Journal for Graduates Students* (*incorporating Data & Statistics*), Volume 5, Number 1, April 2004, pp 13-28.

[PER 96] PERINEL, E., Segmentation et Analyse des données symboliques : Application à des données probabilistes imprécises, *Thèse de Doctorat, Mathématiques Appliquées, Université Paris Dauphine,* 1996.

[UTG 96] UTGOFF, P.E., CLOUSE, J.A., A Kolmogorov-Smirnov metric for decision tree induction, *University of Massachusetts, Amherst*, Number 96-3, 1996.

EClaViSeR : Classification visuelle et interactive pour la recherche d'informations sur le Web

F. Mokaddem*, F. Picarougne[*], H. Azzag[*], C. Guinot[**], G. Venturini[*]

[*]*Université François-Rabelais de Tours, Laboratoire d'Informatique (EA 2101),*
64, Avenue Jean Portalis, 37200 Tours, France
fewzi.mokaddem@etu.univ-tours.fr, {hanene.azzag, fabien.picarougne,venturini}@univ-tours.fr

[**]*CE.R.I.E.S., 20, rue Victor Noir, 92521 Neuilly-sur-Seine Cédex, France*
christiane.guinot@ceries-lab.com

RÉSUMÉ. Nous exposons dans cet article notre algorithme EClaViSeR : Extracting, Clustering and Visualizing Search Results. Cet outil en ligne permet d'aider les utilisateurs à effectuer leurs recherches d'informations sur le Web en classant les résultats et en permettant de visualiser et de manipuler graphiquement les résultats de la classification. Il extrait les résultats du moteur de recherche Google, et ensuite les classes grâce à un algorithme biomimétique à base de nuages d'agents. Il stabilise ce nuage à l'aide d'un algorithme d'affichage de graphes à base de forces et de ressorts. Enfin il présente les résultats à l'aide de techniques de visualisation de données et permet à l'utilisateur d'interagir avec la classification à l'aide d'un outil de pointage jusqu'à ce qu'il trouve l'information qu'il recherche.
MOTS-CLÉS : Classification, Visualisation, Interaction, Web, Moteur de recherche, Algorithme biomimétique.

1 Introduction

Les outils de classification présentent un grand intérêt pour la recherche d'information sur Internet en offrant la possibilité de structurer en classes non prédéfinies à l'avance les résultats fournis par les moteurs de recherche. En effet, les nombreuses réponses fournies par les moteurs de recherche sont rarement structurées et placent donc l'utilisateur devant le difficile problème d'analyse des résultats proposés. Les outils de classification ont donc été appliqués pour trouver des classes au sein de ces résultats, comme par exemple le système Grouper [ZAM 1999], ou encore Scatter/Gather [HEA 1996]. Parallèlement, il existe des techniques de visualisation des résultats d'un moteur de recherche qui vont également grandement aider l'utilisateur à mieux comprendre les résultats proposés. Ces techniques vont par exemple visualiser les liens qui existent entre les résultats (moteurs Kartoo ou MapStan), ou encore positionner les résultats par rapport à des mots-clés comme cela est fait dans SQWID [MCC 1997]. Nous proposons dans cet article un système qui va combiner à la fois l'établissement d'une classification avec la visualisation de cette classification et la possibilité d'interagir avec celle-ci.

Les principes de notre approche sont les suivants : l'extraction et la classification des résultats a lieu initialement en utilisant un algorithme à base de déplacements d'un nuage d'agents [MON 2002]. Cet algorithme propose en sortie un regroupement en classes ainsi qu'un positionnement 2D des résultats en fonction de leur similarité textuelle. Cependant, afin de fixer efficacement la position finale des agents (les déplacements doivent s'arrêter pour rendre la visualisation manipulable par l'utilisateur), nous stabilisons le nuage par un algorithme d'affichage de graphes à base de forces et de ressorts. Ensuite, à partir de la vue globale de cet ensemble de résultats, nous utilisons la visualisation d'attributs des résultats, le filtrage des résultats, et l'affichage d'autres détails sur les résultats à la demande de l'utilisateur.

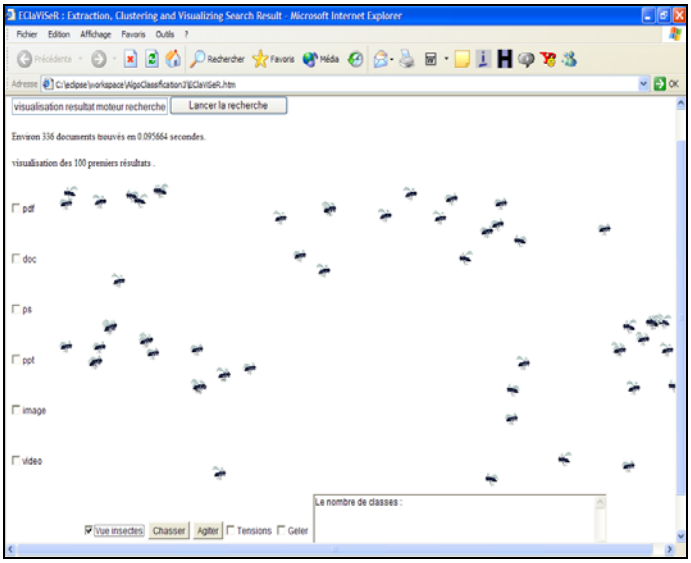


Figure 1 – Déplacement du nuage d'agents dans EClaViSeR.

2 Principes de notre approche

2.1 Interrogation des moteurs et génération des attributs d'un résultat

EClaViSeR est implémenté en JAVA et peut être exécuté via une connexion Internet comme applet Java à travers un navigateur Web. L'utilisateur saisit sa requête à travers une page HTML. Les mots clés sont ensuite transmis au moteur de recherche (actuellement Google) dont nous obtenons directement les résultats exprimés en XML. Après extraction, notre outil génère des informations sur chaque résultat comme les mesures de présence des mots-clés, le résumé court, les liens HTML. Ces informations vont permettre de définir une mesure de similarité entre les résultats. Cette mesure est simplement définie par le nombre de mots en commun.

2.2 Classification par nuage d'agents

Le module de classification que nous avons employé est une implémentation de l'algorithme de classification de [MON 2002]. Il s'inspire du comportement observé chez certains animaux volants ou nageants ayant un comportement social pour le déplacement. Les motivations pour l'utilisation de cet algorithme biomimétique sont les suivantes : il ne nécessite pas d'informations a priori (nombre de classes, partition initiale), il est rapide et peut être interrompu à tout moment (résultat intermédiaire), il visualise les résultats de la classification sous la forme de groupes de données similaires évoluant de manière coordonnée. Les principes des nuages d'agents utilisés au sein de notre outil peuvent être décrits de la façon suivante (fig.1) : chaque résultat du moteur de recherche sera considéré comme un agent. Les agents vont être placés dans un environnement 2D et sont caractérisés par trois éléments : leurs coordonnées dans l'environnement, leurs vecteurs vitesse et des règles comportementales pour gérer les déplacements. Ces règles sont communes à tous les agents et utilisent le voisinage local d'un agent pour décider de changer le vecteur vitesse. Elles vont prendre en compte la similarité des résultats portés par les agents afin de former des nuages de résultats homogènes. A la fin de l'exécution de cet algorithme seront générées des classes en fonction de la proximité des agents (cf. calcul des classes [MON 2002]).

2.3 Stabilisation à base de ressorts et de points d'intérêts

Après la classification des résultats et le calcul de coordonnées 2D initiales pour chacun d'eux, nous devons maintenir la visualisation dans un état d'équilibre (sans déplacement), moyennant un mécanisme à base de ressorts et de points d'intérêts [EAD 1984] [WIS 1999]. Le principe de cet algorithme d'affichage

Figure 2 – Interactions et visualisation des attributs des résultats.

de graphes est de positionner des nœuds en 2D et de les déplacer de manière à atteindre la valeur de longueur voulue pour chaque arc. Ainsi, chaque arc agit comme un ressort tentant de rapprocher ou d'éloigner les noeuds qu'il relie. Un des inconvénients de ce type d'algorithme est le positionnement initial des nœuds (généralement aléatoire) qui peut engendrer un état final sous optimal. Ici, chaque nœud représente un résultat du moteur. Nous relions deux nœuds avec un arc lorsqu'ils appartiennent à la même classe et qu'ils sont fortement similaires. Tous les autres nœuds sont aussi reliés entre eux avec des ressorts qui sont transparents sur l'écran. Ces ressorts servent à maintenir les nœuds les plus similaires cote à cote et repousser les autres. Nous initialisons la position des nœuds en utilisant les coordonnées des agents. L'affichage se stabilise rapidement (et plus efficacement qu'en utilisant le nuage d'agents seul). Cette approche hybride combine donc utilement les avantages des deux méthodes.

2.4 Visualisation des attributs des résultats

Nous nous sommes basés sur l'état de l'art [MOK 2004] pour définir les formes et icônes à utiliser pour représenter visuellement les attributs des résultats du moteur de recherche : 1) la pertinence de chaque résultat par rapport à la requête est exprimée selon la couleur et la dimension de l'icône avec lequel il est représenté ; 2) des icônes de même couleur appartiennent à une même classe ; 3) des histogrammes représentent la présence des mots-clés [VEE 1996] ; 4) la distance entre deux icônes est représentative de la similarité entre les résultats ; 5) quelques icônes de résultats sont reliés entre eux par des traits (la présence d'un trait entre deux icônes symbolise une forte similarité).

2.5 Interactions visuelles et graphiques avec la classification

L'utilisateur peut interagir avec la classification de la manière suivante : le clic sur le bouton gauche de la souris (fig.2 (a)) permet d'afficher des barres histogrammes indiquant la présence des mots-clés, le titre du résultat, son résumé, son lien http. Ensuite, l'utilisation du clic droit de la souris permet d'agir sur un résultat comme la possibilité d'ouvrir le lien dans une nouvelle fenêtre, de fixer ou de libérer la position de ce résultat, de le sélectionner/désélectionner (son icône sera entouré d'un cercle noir), d'afficher son rang dans Google, etc (fig.2 (b)). De plus, notre outil affiche au milieu de l'icône d'un résultat un nombre représentant le nombre de fois ou l'utilisateur a consulté la page Web de celui-ci durant cette session. L'utilisateur a la possibilité de filtrer les résultats par type de fichiers : pdf, doc, ppt, ps, image (des drapeaux représentant le type de documents s'afficheront à coté des icônes). Enfin, l'utilisateur peut déplacer les résultats pour construire sa propre carte de résultats. A chaque déplacement de l'utilisateur, notre algorithme à base de ressorts réajuste dynamiquement la visualisation. L'utilisateur a donc la

possibilité de distinguer rapidement les résultats les plus pertinents, de pouvoir consulter les attributs de chaque résultat, et de pouvoir les comparer. Ce temps est beaucoup moins important que s'il avait consulté les cent premiers résultats de Google un à un.

3 Conclusion et Perspectives

Nous avons présenté dans cet article notre outil EClaViSeR, un algorithme opérationnel[1] d'aide à la recherche d'information sur le Web à travers le moteur de recherche Google. Il extrait, classe et présente les résultats sous forme visuelle. Il combine plusieurs techniques (classification à l'aide d'algorithmes biomimétiques, stabilisation à l'aide de points d'intérêts, de forces et ressorts, représentation visuelle d'attributs des résultats, interaction avec la visualisation). Parmi les perspectives liées à ce travail, nous souhaitons améliorer EClaViSeR afin de permettre à l'utilisateur de reformuler interactivement la requête à partir d'une représentation visuelle de celle-ci. De même, d'autres attributs pourraient être visualisés, comme des distances entre phrases appartenant à différents documents, en faisant intervenir des techniques de fouille de texte. Par ailleurs, nous souhaitons quantifier plus précisément l'apport de l'algorithme à base d'agents dans l'initialisation de l'algorithme d'affichage de graphes, en comparant les temps de convergence et la qualité du résultat final (avec initialisation à base d'agents ou avec une initialisation aléatoire). Nous pensons également que cet outil pourrait être appliqué à la visualisation/exploration de toute classification de données à partir d'une matrice de similarités. Enfin, nous souhaitons évaluer et comparer notre outil de recherche avec des utilisateurs sur des problèmes réels, comme cela est fait par exemple dans [CUG 2000].

4 Références

[CUG 2000] Cugini, J. V., Presenting Search Results: Design, Visualization and Evaluation. In: Information Doors-Where Information Search and Hypertext Link: San Antonio, TX, May 30 2000.

[EAD 1984] Eades, P. A heuristic for graph drawing. Congressus Numerantium, 42, pp 149-160. 1984.

[HEA 1996] Hearst M. A. et J. O. Pedersen, Reexamining the cluster hypothesis: Scatter/Gather on retrieval results, in:Proceedings of the 19th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'96), 1996, pp 76-84.

[MCC 1997] McCrickard, D. Scott; Kehoe, Colleen M.: Visualizing Search Results using SQWID. In: WWW 6: Sixth International World Wide Web Conference. Conference: Santa Clara, CA, 1997.

[MOK 2004] Mokaddem F., « Etat de l'art sur les techniques visuelles utilisées dans la recherche d'information sur Internet » Mémoire DIRS, 3ème cycle, Laboratoire d'Informatique, Equipe RTIC, Ecole polytechnique de l'Université de Tours, septembre 2004.

[MON 02] Monmarché N., C.Guinot, G.Venturini. Fouille visuelle et classification de données par nuage d'insectes volants. Extraction des Connaissances et Apprentissage : Méthodes d'optimisation pour l'extraction de connaissances et l'apprentissage. volume 6, pp 729-752, 2002.

[VEE 1996] Veerasamy, A., Belkin, N. J.: Evaluation of a Tool for Visualization of Information Retrieval Results. In: Frei, Hans-Peter; Harman, Donna K.; Schäuble, Peter et al. (Eds.): SIGIR 1996: 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Switzerland, August 18 -22 1996. New York 1996. pp. 85-92.

[WIS 1999] Wise, James A.: The Ecological Approach to Text Visualization. In: Journal of the American Society for Information Science (JASIS), 50, 1999, 13, pp. 1224-1233.

[ZAM 1999] Zamir, O., Etzioni, O., Grouper: A dynamic clustering interface to web search results. Proceedings of WWW8, Toronto, Canada, 1999.

[1] http://www.antsearch.univ-tours.fr/mokaddem/index.php

HGT-Simulator : logiciel pour simuler des transferts horizontaux de gènes

Dung Nguyen, Alix Boc et Vladimir Makarenkov

Département d'informatique,
Université du Québec à Montréal,
Case postale 8888, succursale Centre-ville
Montréal (Québec) Canada, H3C 3P8
Courriels : nguyen.van_dung2@uqam.ca, alix.boc@ uqam.ca et makarenkov.vladimir@uqam.ca.

RÉSUMÉ. Le problème de la détection et de la classification de transferts horizontaux de gènes (i.e. transferts latéraux de gènes) est parmi les plus ardus en biologie moléculaire. Dans cet article nous présentons un logiciel, appelé HGT-Simulator, permettant de simuler un modèle d'évolution comprenant les transferts horizontaux de gènes. Des transferts aléatoires sont générés entre les branches d'un arbres phylogénétique donné tout en respectant le modèle stochastique d'évolution choisi. Ce logiciel a été employé pour tester notre nouvelle méthode de détection des transferts horizontaux.

MOTS-CLÉS : arbre phylogénétique, transfert horizontal de gène, modèle d'évolution stochastique, évolution réticulée.

1 Introduction

L'évolution des êtres vivants a longtemps été modélisé uniquement à l'aide des arbres phylogénétiques (i.e. arbres additifs). Dans un arbre phylogénétique deux espèces sont toujours reliées par un chemin passant par leur ancêtre commun. Un tel modèle ne peut inclure des scénarios d'évolution réticulée comprenant les transferts horizontaux de gènes et l'hybridation. La recombinaison homologue, l'hybridation, le transfert latéral de gènes, la duplication d'un gène suivie de sa perte et l'évolution convergente sont les principaux mécanismes d'évolution réticulée [LEG 02]. Les deux premiers phénomènes peuvent être représentés seulement à l'aide des modèles en réseaux, tandis que les trois derniers nécessitent plus d'un arbre phylogénétique pour leur représentation.

Le transfert horizontal consiste en un échange direct de matériel génétique d'une lignée à une autre [DOO 99]. Il est très fréquent chez les procaryotes. Bactéries et Archéobactéries ont développé des mécanismes sophistiqués pour acquérir rapidement de nouveaux gènes à l'aide du transfert latéral. Ces mécanismes ont été favorisés par la sélection naturelle par rapport à l'évolution génétique par mutations. Les trois principaux mécanismes de transfert de gènes sont les suivantes : la transformation par acquisition d'ADN directement de l'environnement, la conjugaison qui est enclenchée par des plasmides conjugaux ou par des transposons conjugaux et la transduction par transfert d'ADN par phage. Ces mécanismes peuvent introduire des séquences d'ADN de l'espèces donneur ayant très peu de similarité avec le reste de l'ADN de l'espèce hôte.

Plusieurs méthodes pour modéliser et détecter les transferts horizontaux sont disponibles : Page et Charleston [PAG 98] ont décrit un ensemble de règles d'évolution qui doivent être prises en compte lors de la modélisation des transferts, Mirkin, Muchnik et Smith [MIR 95] ont décrit une méthode de réconciliation d'arbres permettant de combiner plusieurs phylogénies de gènes en arbre d'espèces unique, Hallet et Lagergren [HAL 01] ont proposé un modèle de détection de transferts permettant d'inscrire les phylogénies de gènes en phylogénie d'espèces. Par ailleurs, Boc et Makarenkov [BOC 03]

et Makarenkov, Boc et Diallo [MAK 04] ont introduit deux méthodes de détection impliquant des scénarios unique et multiples des transferts horizontaux.

Dans cet article nous décrivons un outil de simulation des transferts latéraux de gènes permettant aux chercheurs de générer les transferts à l'intérieur d'un arbre phylogénétique donné. Ce programme incluant de nombreux modèles d'évolution connus peut être utilisé pour comparer les méthodes d'inférence de transferts horizontaux. Les règles biologiques pertinentes spécifiées dans [PAG 98] et [MAK 05] ont été incorporées dans le modèle implanté. Ce logiciel a premièrement été utilisé dans les simulations statistiques [MAK 05] effectuées pour tester une nouvelle méthode de détection de transferts latéraux.

2 Description du logiciel HGT-Simulator

Notre logiciel de génération des transferts horizontaux utilise les résultats du logiciel Seq-Gen [RAM 97]. Seq-Gen est un programme permettant de simuler l'évolution de séquences d'ADN le long d'une phylogénie donnée. Seq-Gen inclut plusieurs modèles stochastiques d'évolution ayant faits leurs preuves en analyse phylogénétique.

La nouvelle application HGT-Simulator étend la possibilité initiale de Seq-Gen de modéliser l'évolution arborescente en se basant sur les principes de réseaux réticulés [LEG 02]. À son entrée HGT-Simulator récupère les séquences associées aux nœuds de l'arbre initial qui ont été simulées par Seq-Gen. En fonction du nombre de transferts et du modèle d'évolution des séquences d'ADN choisis par l'utilisateur, le programme génère les transferts en affichant à sa sortie la liste des transferts engendrés, l'arbre modifié suite à ces transferts, de même que les nouvelles séquences d'ADN associées aux nœuds de cet arbre modifié.

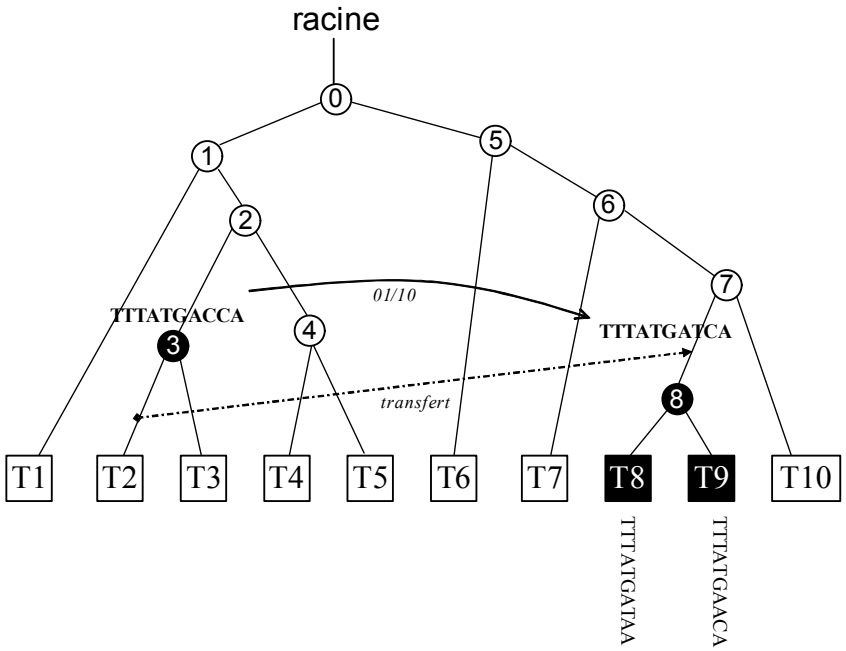


Figure 1. Transfert horizontal entre les branches (3, T2) et (7, 8) a eu lieu. Ce transfert explique la ressemblance entre les séquences associées aux nœuds 3 et T2 d'un côté et aux nœuds 8, T8 et T9 de l'autre.

Notre procédure algorithmique se divise en trois étapes principales :
***Étape* 1**. Identifier un transfert (i.e. deux branches de l'arbre) en respectant les règles d'évolution.
***Étape* 2**. Générer les séquences associées aux nœuds dans le sous-arbre affecté par le transfert. La figure 1 montre un transfert horizontal entre les branches (3, T2) et (7, 8). Ce transfert affecte tout d'abord le nœud 8 (la séquence associée au nœud 8 a maintenant seulement une différence par rapport à la

séquence associée au nœud 3) ainsi que les feuilles T8 et T9. L'algorithme choisit arbitrairement l'emplacement du départ du transfert sur la branche d'origine (3, T2) de même que l'emplacement de son arrivée sur la branche cible (7, 8). Une nouvelle distance entre les nœuds 3 et 8 est calculée en fonction de ces emplacements et du modèle d'évolution retenu. Ici une seule différence existe entre les séquences d'ADN TTTATGACCA et TTTATGATCA associées respectivement aux nœuds 3 et 8. Dans ce modèle, nous supposons que le gène de l'espèce donneur remplace complètement le gène homologue de l'hôte en transformant la phylogénie de départ en un arbre phylogénétique différent (figure 2).
***Étape* 3**. Reprendre Étape 1 tant qu'il reste des transferts à engendrer.

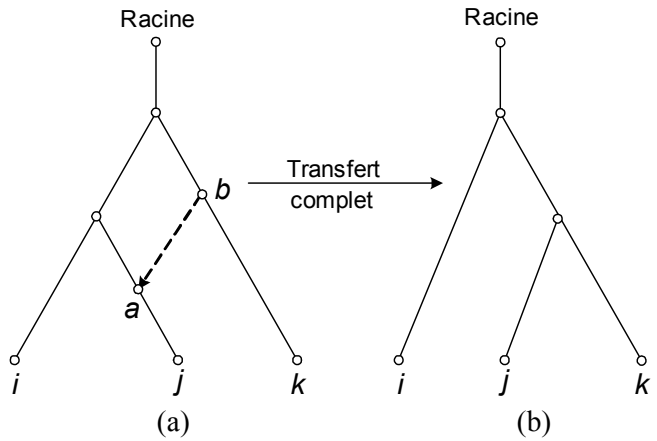


Figure 2. Modèle du transfert complet. Le gène de l'espèce donneur remplace le gène homologue de l'hôte ce qui transforme la phylogénie initiale (a) en arbre phylogénétique différent (b).

De plus, l'introduction de quelques règles d'évolution de base est nécessaire afin de renforcer la plausibilité biologique du modèle (voir [PAG 98] pour plus de détails sur ces règles). Par exemple, les transferts impliquant des espèces appartenant à la même lignée doivent être interdits (figure 3).

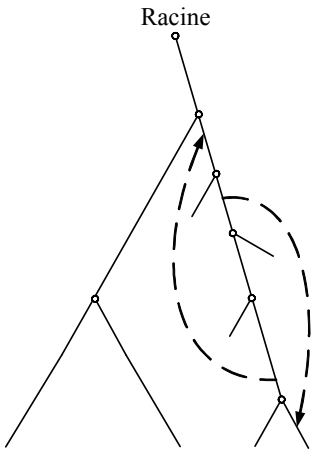


Figure 3. Transferts horizontaux sur la même lignée sont interdits.

Finalement, HGT-Simulator préserve naturellement les caractéristiques de fluctuations stochastiques de Seq-Gen, mais apporte en plus une nouvelle dimension permettant de simuler les transferts horizontaux.

Une seconde version du programme, indépendante de Seq-Gen, a aussi été développée. Cette version permet de simuler les transferts horizontaux pour un ensemble de phylogénies aléatoires qui peuvent être engendrées par le programme. Les séquences peuvent être générées selon 3 modèles d'évolution : Jukes-Cantor [JUK 69], Kimura 2 paramètres [KIM 80] et Jin-Nei [JIN 90]. À l'entrée, ce programme prend le nombre d'espèces, la taille des séquences, le nombre de transferts, le nombre d'arbres à

considérer et le modèle d'évolution. À la sortie, il fournit les matrices de distance entre les feuilles des arbres modifiés par les transferts ainsi que la liste de transferts obtenue pour chaque arbre.

3 Utilisation du logiciel HGT-Simulator dans une étude Monte-Carlo

Une étude Monte-Carlo a été effectuée pour tester les performances d'une nouvelle méthode [MAK 05] de détection de transferts latéraux. Nous avons examiné comment la procédure d'inférence des transferts se comporte dépendamment du modèle d'évolution des séquences d'ADN et du nombre d'espèces. Les résultats présentés sur la figure 4 ont été obtenus pour des arbres phylogénétiques binaires ayant 8, 16, 24, 32, 48 et 64 feuilles (i.e. espèces). Dans chaque cas, une vraie topologie d'arbre *T*, a été obtenue aléatoirement en utilisant la procédure de génération d'arbres proposée par [KUH 94]. Les longueurs des branches de *T* ont été calculées à l'aide d'une loi exponentielle. Suivant l'approche décrite dans [GUI 02], nous avons ajouté du bruit sur les branches des vraies phylogénies pour créer une déviation de l'hypothèse de l'horloge moléculaire. Toutes les longueurs des branches de *T* ont été multipliées par le coefficient 1+**a***x,* où la variable *x* a été obtenue d'une distribution exponentielle standard ($P(x>k) = \exp(-k)$) et la valeur de la constante **a** a été fixée à 0.8. Les arbres générés par une telle procédure ont la profondeur $O(\log(n))$, où *n* est le nombre d'espèces. Chaque arbre phylogénétique enraciné a par la suite été soumis à HGT-Simulator qui, à son tour, a simulé l'évolution des séquences d'ADN le long de ses branches. Les modèles d'évolution de Jukes et Cantor [JUK 69], de Kimura 2 paramètres [KIM 80] et de Jin-Nei Gamma [JIN 90] ont été considérés. Par la suite, la procédure de génération de transferts a engendré des transferts horizontaux de gène tout en respectant les règles d'évolution spécifiées dans la section précédente. Un seul transfert par arbre a été engendré dans cette étude. HGT-Simulator a régénéré des séquences d'ADN pour chaque nœud de l'arbre situé sous la branche affectée par un transfert (i.e. dans le sous-arbre qui a changé sa place dans la phylogénie à cause d'un transfert latéral). Pour chaque taille de données, 500 phylogénies aléatoires différentes ont été examinées. La méthode NJ [SAT 87] a été utilisée pour reconstruire les arbres de gène à partir des distances obtenues des séquences terminales (i.e. séquences associées aux feuilles) ; les vraies phylogénies *T*, utilisées comme arbres d'espèces, ont été supposées connues.

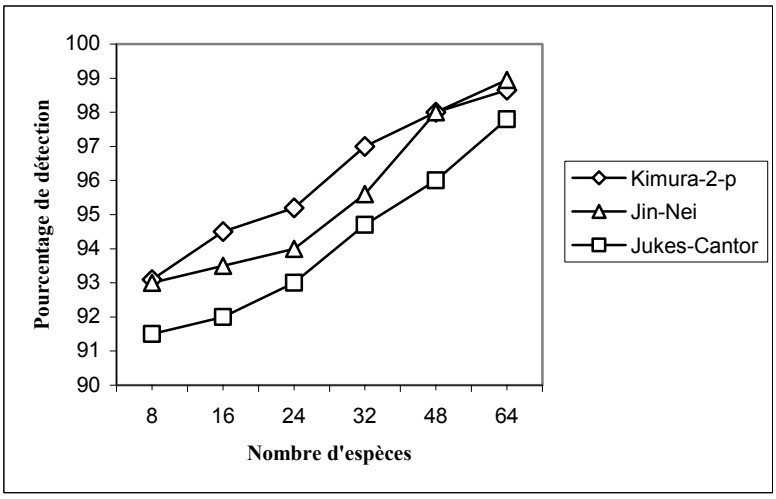


Figure 4. Pourcentage de détection des transferts horizontaux obtenus en utilisant la distance topologique de Robinson et Foulds (RF) comme critère d'optimisation pour réconcilier les topologies de gène et d'espèces [MAK 05]. Tests ont été effectués avec des arbres ayant de 8 à 64 feuilles. Les modèles d'évolution de Jukes et Cantor [JUK 69], Kimura 2 paramètres [KIM 80] et Jin-Nei Gamma [JIN 90] ont été comparés.

La figure 4 présente les résultats des simulations obtenus pour le modèle de transfert horizontal complet [MAK 05]. Pour tous les trois modèles d'évolution considérés, la méthode de détection de

transferts a pu retrouver le transfert en question avec au moins 91.4% de succès. Les meilleurs résultats ont été obtenus avec le modèle d'évolution Kimura 2 paramètres, suivi par ceux de Jin-Nei et Jukes-Cantor. Pour les phylogénies avec 64 feuilles, le pourcentage de détection a atteint 98.5-99.1% avec les modèles Kimura 2 paramètres et Jin-Nei. Le pourcentage de détection augmente quand le nombre d'espèces augmente; cette tendance est certainement due au problème bien connu de reconstruction de petites phylogénies. Les résultats de la méthode de détection [MAK 05] sont surtout très prometteurs pour des larges phylogénies ou le pourcentage de détection tend vers 98 –99%.

4 Bibliographie

[DOO 99] DOOLITTLE W. F., "Phylogenetic classification and the universal tree", *Science,* 284:2124-2129.

[GUI 02] GUINDON S., GASCUEL O., "Efficient biased estimation of evolutionary distances when substitution rates vary across sites", *Mol. Biol. Evol*., 19:534-543.

[HAL 01] HALLET M., LAGERGREN J., "Efficient algorithms for lateral gene transfer problems", pp. 149-156, *proceedings de RECOMB 2001*, ACM Press, New-York.

[JIN 90] JIN L., NEI M., "Limitations of the evolutionary parsimony method of phylogenetic analysis", *Mol. Biol. Evol*., 7:82-102.

[JUK 69] JUKES T.H., CANTOR C., "Mammalian Protein Metabolism", pp. 21-132 dans H. N. Munro, editor, *Evolution of protein molecules*, Academic Press, New York.

[KIM 80] KIMURA M., "A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences", *J. Mol. Evol*., 16:111-120.

[KUH 94] KUHNER M., FELSENSTEIN J., "A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates", *Mol. Biol. Evol*. 11:459-68.

[LEG 02] LEGENDRE P., MAKARENKOV V., "Reconstruction of biogeographic and evolutionary networks using reticulograms", *Systematic Biology*, 51:199-216.

[MAK 03] MAKARENKOV V., BOC A., "New Efficient Algorithm for Detection of Horizontal Gene Transfer Events", pp. 190-201 dans G. Benson and R. Page, eds. *Algorithms in Bioinformatics*. Springer Verlag, proceedings of WABI 2003, Budapest.

[MAK 04] MAKARENKOV V., BOC A., DIALLO B., "Representing lateral gene transfer in species classification", Unique scenario. Pp. 439:446 dans D. Banks, L. House, F. R. McMorris, P. Arabie et W. Gaul, eds. *Classification, Clustering and Data Mining Applications*, Springer Verlag, proceeding of IFCS 2004, Chicago.

[MAK 05] MAKARENKOV V., BOC A., DELWICHE C. F., PHILIPPE H., "A novel approach for detecting horizontal gene transfers: Modeling partial and complete gene transfer scenarios", soumis.

[MIK 95] MIRKIN B. G., MUCHNIK I., SMITH T. F., "A Biologically Consistent Model for Comparing Molecular Phylogenies", *J. of Comp. Biol*., 2:493-507.

[PAG 98] PAGE R. D. M., CHARLESTON M. A., "Trees within trees: phylogeny and historical associations", *Trends in Ecol. and Evol*., 13:356-359.

[RAM 97] RAMBAUT A., GRASSLY N.C., "Seq-Gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees", *Comput. Appl. Biosci.,* 13: 235-238.

[SAI 87] SAITOU N., NEI M., "The neighbour-joining method: a new method for reconstructing phylogenetic trees", *Mol. Biol. Evol*., 4:406-425.

Hiérarchies pour la classification supervisée

Christophe Osswald et Arnaud Martin

Laboratoire E^{3}I^2, ENSIETA
2 rue François Verny 29806 Brest Cedex 9
Christophe.Osswald@ensieta.fr et Arnaud.Martin@ensieta.fr

RÉSUMÉ. *Les méthodes pour construire une hiérarchie sont nombreuses. Ici, nous explorons les fonctions de Lance et Williams pour déterminer des paramètres de construction d'une classification automatique adaptés à des données constituées d'imagettes sonar. Nous considérons que la qualité d'une hiérarchie pour nos données est liée au fait qu'elle contient des classes homogènes relativement à l'étiquetage existant. Lorsque nous utilisons ces paramètres pour construire une hiérarchie mêlant données étiquetées et non étiquetées, la même mesure de qualité de règle permet de proposer des étiquettes pour les données apparaissant dans les classes les plus homogènes.*

MOTS-CLÉS : *Hiérarchies, fonctions d'agrégation, mesure de qualité, caractérisation des fonds sous-marins.*

Introduction

La caractérisation des fonds sous-marins permet de constituer des cartes à l'usage des sédimentologues, de la navigation sous-marine autonome ou de la détection de pollution. La carte est composée de grandes images sonar (voir Martin *et al.* [MSL04]), découpée en 4003 imagettes de 64x384 pixels. La distinction entre deux types de sédiments (sable, rocher, cailloutis, ride, vase, ombre) est souvent malaisée, même pour un expert. Les types de sédiment sont présents de façon variée : le sable représente près de 55% des imagettes, les rochers 21%, les cailloutis moins de 1%. De plus, 40% des imagettes contiennent plus d'un type de sédiment : l'étiquette indique alors uniquement le sédiment occupant la plus grande surface sur l'imagette.

Ici, nous mettons en oeuvre une méthode de classification supervisée qui autorise un étiquetage multiple (plusieurs propositions sur une seule imagette) ou aucune étiquette. Ce type d'étiquetage devient particulièrement pertinent dans le cadre d'une fusion multi-capteur ultérieure.

Nous rappelons tout d'abord les mécanismes d'un algorithme de classification ascendante hiérarchique, ainsi que le formalisme de Lance et Williams pour les fonctions d'agrégation utilisées dans ce cadre. Nous donnons des résultats concernant les espaces de définition de ces fonctions, et présentons une famille de ces fonctions que nous utilisons pour couvrir les méthodes usuelles. A l'aide de la mesure de Jaccard, nous optimisons notre méthode de construction de hiérarchie dans une optique de classification supervisée.

1. Classification Ascendante Hiérarchique

Les algorithmes de classification ascendante hiérarchique (CAH) constituent des méthodes usuelles pour construire un système de classes à partir d'un ensemble d'objets sur lequel on peut évaluer une dissimilarité $d(x, y)$. L'algorithme construit une distance entre classes : $d(\{x\}, \{y\}) = d(x, y)$. A chaque étape il fusionne les deux classes les plus proches en une seule classe, et réévalue les dissimilarités entre cette nouvelle classe et les classes préexistantes. Les classes $C = A \cup B$ ainsi créées, munies de l'indice $d(A, B)$ forment une hiérarchie indicée ; la dissimilarité associée est une ultramétrique.

1.1. *Fonctions de Lance et Williams*

Lance et Williams [LW67] synthétisent de nombreuses méthodes d'évaluation de distance inter-classes par la formule :

$$d^p(C,D) = \alpha_A d^p(A,D) + \alpha_B d^p(B,D) + \beta d^p(A,B) + \gamma |d^p(A,D) - d^p(B,D)|$$

Chen [Che96] restreint les formes des termes α, β et γ à des fonctions ne dépendant que de paramètre issus des cardinaux des classes, afin de garantir que les valeurs obtenues amènent bien à une ultramétrique. La plupart des algorithmes usuels de CAH peuvent être définis par ces trois fonctions α, β et γ ainsi qu'un réel p.

$$r_A = \frac{|A|}{|A \cup B|} \quad r_B = \frac{|B|}{|A \cup B|} \quad r_D = \frac{|D|}{|A \cup B|} \quad p \neq 0$$
$$d^p(C,D) = \alpha(r_A, r_D) d^p(A,D) + \alpha(r_B, r_D) d^p(B,D) + \beta(r_A, r_B, r_D) d^p(A,B) + \gamma(r_D) |d^p(A,D) - d^p(B,D)|$$

Algorithme	$\alpha(u,w)$	$\beta(u,v,w)$	$\gamma(w)$	p
lien simple	$-1/2$	0	$1/2$	1
lien complet	$1/2$	0	$1/2$	1
méthode de Ward	$(u+w)/(1+w)$	$-w/(1+w)$	0	2

1.2. *Fonctions admissibles et internes*

Pour s'assurer qu'il n'y ait pas d'inversion lors de la CAH, *i.e.* qu'on n'ait pas $A \subsetneq B$ avec $f(A) > f(B)$, il faut que l'algorithme soit monotone (Dragut [Dra01]). On parle alors d'une fonction admissible :

i. $\alpha(u,w) + \alpha(1-u,w) + \beta(u,1-u,w) \geqslant 1$
ii. $\alpha(u,w) \geqslant 0$
iii. $\gamma(w) \geqslant \max\{-\alpha(u,w), -\alpha(1-u,w)\}$

Un algorithme de Lance et Williams conserve l'espace si :

$$\min\{d(A,D), d(B,D)\} \leqslant d(A \cup B, D) \leqslant \max\{d(A,D), d(B,D)\}$$

Les ultramétriques sont alors des points fixes. C'est le cas pour le lien simple et le lien complet, mais pas pour la méthode de Ward.

De nombreux algorithmes de CAH ne peuvent s'écrire comme des algorithmes de Lance et Williams. Notamment, toute fonction d'agrégation interne amène à un algorithme de CAH conservant l'espace. Osswald [Oss03] étudie une famille de telles fonctions qui utilisent les médianes.

2. Hiérarchies

Le lien simple est connu pour favoriser les hiérarchies déséquilibrées : dès que l'algorithme engendre une classe d'une taille conséquente, la plupart des agrégations ultérieures vont se faire avec cette classe. Cela est dû au fait que si x est un élément de A et y un élément de B, on a $d(x,y) \geqslant d(A,B)$ et $d(A,B)$ est souvent petit lorsque A et B sont de cardinal important. A l'inverse, pour le lien complet comme pour la méthode de Ward, on a $d(x,y) \leqslant d(A,B)$ et l'effet obtenu est inverse : les hiérarchies équilibrées sont favorisées par l'algorithme.

Ici, la mesure de distance entre nos imagettes est la distance euclidienne entre les vecteurs constitués par les paramètres extraits d'un filtrage de Gabor [MSL04]. Les hiérarchies obtenues sur un ensemble de 24 imagettes pour lesquelles tous les types de sédiments sont représentés à quatre reprises, avec éventuellement des frontières sont représentées en figure 1.

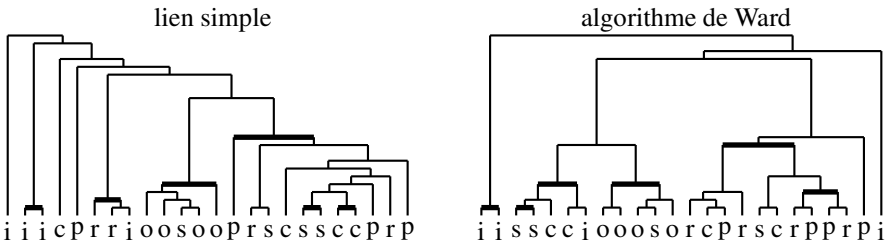


FIG.1. *Hiérarchies construites sur 24 imagettes avec frontières*

2.1. *Mesure de qualité d'une hiérarchie*

Nous considérons qu'une hiérarchie est d'autant plus fidèle à l'analyse de l'expert qu'elle contient des classes représentant au mieux les types de sédiments. L'intérêt d'une hiérarchie est donc défini par sa forme, *i.e.* l'ensemble des classes qu'elle contient. La valeur des indices des classes n'a donc pas d'intérêt pour cet objectif.

Pour chaque type de sédiment i, on considère la classe A de la hiérarchie qui « ressemble » le plus à l'ensemble M_i des imagettes de type i. Pour ce faire, nous utilisons la mesure de Jaccard $J(A \iff M_i)$, où $P(A)$ représente la proportion des éléments de A par rapport à l'ensemble tout entier, pour obtenir une mesure globale $q(H)$ de la qualité d'une hiérarchie. Elle est assez résistante au nombre d'objets considérés [TKS02] :

$$J(A \iff M_i) = \frac{P(A \cap M_i)}{P(A) + P(M_i) - P(A \cap M_i)}$$

$$q(H) = \prod_i \max_{A \in H} J(A \iff M_i)$$

2.2. *Paramètres pour les fonctions de Lance et Williams*

Nous plongeons les méthodes d'agrégation dans un espace constitué par quatre segments de l'espace des fonctions admissibles. Le tableau suivant définit ces quatre segments, le réel $x \in [0, 1]$ est le paramètre qui permet de parcourir chaque segment.

Algorithme	$\alpha(u, w)$	$\beta(u, v, w)$	$\gamma(w)$	p
simple à complet	$-1/2 + x$	0	$1/2$	1
complet à intermédiaire	$\frac{1-x}{2} + x\frac{u+w/2}{1+w}$	0	$\frac{x-1}{2}$	1
intermédiaire à Ward	$\frac{u+(1+x)w/2}{1+w}$	$\frac{-xw}{1+w}$	0	2
Ward à Ward-γ_1	$(u+w)/(1+w)$	$-w/(1+w)$	x	2

Le premier segment est composé de fonctions internes, qui conservent donc l'espace. Les deux suivants lient le lien complet à la méthode de Ward, et dilatent l'espace. Le dernier segment modifie l'algorithme de Ward en augmentant la valeur de $\gamma(w)$ et donc la propension de l'algorithme à construire une hiérarchie équilibrée. Il se conclut par la méthode Ward-γ_1, qui est une telle extension de la méthode de Ward

Une combinaison linéaire de la fonction d'agrégation du lien complet et de celle de la méthode de Ward n'est pas toujours admissible. Le lien intermédiaire permet de rester au sein des fonctions admissibles qui dilatent l'espace en liant ces deux algorithmes usuels. Le lien moyen est une méthode qui conserve l'espace, et s'il est possible de paramétrer le passage du lien simple au lien complet en passant par le lien moyen, les fonctions de Lance et Williams obtenues perdent largement en lisibilité par rapport à celles que nous utilisons. Dans la mesure où les résultats obtenus mettent en exergue l'intérêt des méthodes qui ne conservent pas l'espace, nous n'avons pas retenu le lien moyen parmi notre famille de fonctions d'agrégation.

La mesure de qualité ne tient compte que de la forme de la hiérarchie produite pour un paramètre x. Ainsi, pour un segment et un ensemble d'imagettes donnés, l'algorithme de Lance et Williams considéré produit une hiérarchie $H(x)$, et la fonction qui à x associe $q(H(x))$ est constante par morceaux. La figure 2 montre les moyennes de $q(H(x))$ pour 100 jeux de 47 imagettes.

Dans le cadre de cette étude, nous ne conservons que les imagettes ne contenant qu'un seul type de sédiment : cela permet de se limiter aux imagettes pour lesquelles les informations fournies par l'expert sont les plus fiables, et pour lesquelles le jeu de paramètres extraits est efficace.

Les différents types de sédiment sont présents en proportions très différentes. Afin d'analyser le comportement des méthodes de CAH vis-à-vis des classes de tailles variées, nous utilisons des échantillons dans lesquels le nombre d'imagette de type i est proportionnel à n_i^λ où n_i est le nombre d'imagettes de type i total, et où λ prend les valeurs 0.3, 0.5 et 0.7. La présence de classes naturelles de tailles déséquilibrées n'est donc en rien une gêne pour cette méthode, alors qu'elle amène souvent à des difficultés dans les applications de classification supervisée.

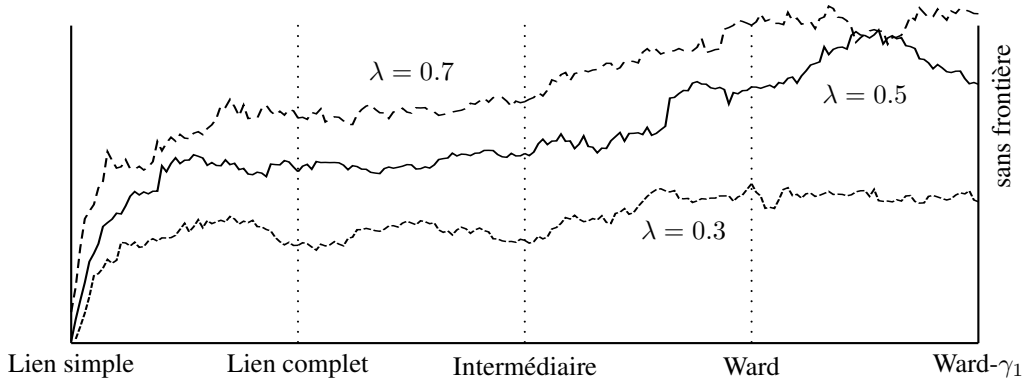


FIG.2. *Qualité des hiérarchies obtenues sur 47 imagettes*

3. Conclusion

Cette démarche nous permet d'associer classification supervisée et classification non-supervisée, en optimisant les paramètres de classification non-supervisée sur un sous-ensemble de données connu. Par la suite, la mesure de Jaccard nous permet d'identifier pour chaque type de sédiment une meilleure classe dans la hiérarchie. L'étiquetage ainsi réalisé n'est pas univoque, et les imagettes peuvent aisément recevoir zéro ou deux étiquettes. Nous obtenons une structure qui étend minimalement les partitionnements, et qui est assez simple à prendre en compte dans le cadre de la fusion de classifieurs.

Références

[Che96] Z. Chen. Space-conserving agglomerative algorithms. *Journal of Classification*, 13 :157–168, 1996.

[Dra01] A. Dragut. Characterization of a set of algorithms verifying the internal similarity. *Mathematical Reports*, 53(3-4) :225–232, 2001.

[LW67] G. N. Lance and W. T. Williams. A general theory of classificatory sorting strategies. *the Computer Journal*, 9(4) :373–380, 1967. and vol. 10, n°3, pp 271-277.

[MSL04] A. Martin, G. Sévellec, and I. Leblond. Characteristics vs decision fusion for sea-bottom characterization. In *colloque caractérisation in-situ des fonds sous-marins*, Brest, France, 2004.

[Oss03] C. Osswald. Robustesse aux variations de méthode pour la classification hiérarchique. In *XXXVèmes Journées de Statistiques*, Lyon, France, 2003.

[TKS02] P.-T. Tan, V. Kumar, and J. Srivastana. Selecting the right interestingness measure for association patterns. In *SIGKDD'02*, Edmonton, Canada, 2002.

UNIFORMISATION RELATIONNELLE DES PARAMETRES DE DESCRIPTION

Mohamed OUALI ALLAH

Centre de Recherche des Ecoles de Saint-Cyr
56381 GUER CEDEX
Phone : +33 297 707 632 Fax : +33 297 707 615
E-mail : mohamed.ouali-allah@st-cyr.terre.defense.gouv.fr

RÉSUMÉ :

La concomitance au sein d'un même corpus de données de paramètres de description de natures différentes, suppose un système de codage unique et homogène. Celui que nous proposons est basé sur la représentation de chaque paramètre de description par une matrice de pondération sur l'ensemble des couples d'objets.

Ce codage intègre les divers types de descripteurs, tout en préservant leur spécificité : pour le quantitatif on utilise l'écart de valuation entre couples d'objets, et pour le qualitatif, un préordre total sur l'ensemble des couples de modalités. Ce procédé a même autorisé l'incorporation de descripteurs de données symboliques.

L'homogénéité de cette représentation permet de quantifier aisément les accords entre différents types de descripteurs. Leurs similarités sont mesurées par un coefficient d'association, qui s'apparente asymptotiquement à une corrélation linéaire.

MOTS-CLÉS : —Numérique / symbolique —Mesure de similarités —Préordonnance —Qualitatif / quantitatif

1 Introduction

Les données combinent de plus en plus le quantitatif et le qualitatif et parfois, même le numérique et le symbolique. L'objectif de ce travail est d'introduire un système de codage autorisant l'intégration de tous ces types de descripteurs dans une même structure de représentation, afin de pouvoir les comparer.
Notre démarche s'inscrit dans une approche relationnelle, où chaque paramètre de description engendre une relation binaire sur l'ensemble des objets. Un indice basé sur le produit scalaire entre les valuations induites par ces relations permet de mesurer les similarités entre descripteurs et de procéder par la suite à une classification hiérarchique.

2 Typologie des paramètres de description

Un paramètre de description ω est une application définie sur l'ensemble des objets O à valeurs dans un domaine d'observation Ω : $\omega : O \rightarrow \Omega$
$i \rightarrow \omega(i)$

2.1 Premier distinguo : données numériques / données à modalités

Les valeurs prises par le paramètre de description ω peuvent être de type :

• Numérique (i.e. : $\Omega = \mathbb{R}$), c'est le cas des variables quantitatives classiques auxquelles on préférera le terme de descripteurs numériques ;

• Enuméré, où toutes les valeurs (on parle de modalités) sont représentées dans un ensemble d'indexation $M_\omega = \{1, 2, ..., m_\omega\}$ (donc : $\Omega = M_\omega$). Ce cas correspond aux variables qualitatives classiques, qu'on intitulera ici descripteurs à modalités.

Cette première distinction nous permet d'établir une démarcation entre le numérique et le symbolique. En effet, si on se réfère à la définition donnée par E. Diday dans [DID 91], tout descripteur à modalités est de type symbolique, puisqu'il ne peut être considéré comme un élément de $\mathbb{R}^n$ (n étant le nombre d'objets).

2.2 Second distinguo : qualitatif / symbolique

Mais évidemment, les variables qualitatives classiques qui ont le même niveau et le même espace (chaque objet ne peut posséder qu'une et une seule modalité d'un même descripteur), ne peuvent suffire pour décrire des données *hétérogènes* (un groupe d'objets peut être considéré comme un seul objet), *irrégulières* (un objet n'est pas forcément décrit par tous les descripteurs), *multivaluées* (un objet peut posséder plusieurs modalités d'un même descripteur) et *structurées* (la description d'un objet peut dépendre de celle des autres).

Le concept descriptif que nous proposons [OUA 00-a] et qui est intitulé *descripteurs à modalités non disjointes*, permet d'appréhender toutes les caractéristiques des données symboliques exposées ci-dessus. Tout objet peut posséder une ou plusieurs modalités d'un même descripteur, comme il peut n'en posséder aucune (i.e. : $\Omega = P(M_\omega)$, ensemble des parties de M_ω). Concrètement, à chaque objet –pour un descripteur donné– est associé une distribution de fréquences sur l'ensemble de ses modalités.

3 Système de codage

3.1 Bases du codage des descripteurs à modalités

A chaque descripteur à modalités est associé un préordre total sur l'ensemble des couples de ses modalités. Ce préordre appelé *préordonnance*, peut être fourni par le praticien pour concrétiser les particularités des différents descripteurs. Sinon, il sera élaboré à partir d'un graphe valué qui caractérise la structure associée à l'ensemble M des modalités de chaque descripteur.

3.1.1 Préordonnance fournie

Il s'agit des variables préordonnances au sens classique du terme, c'est à dire que le praticien fournit un préordre total sur l'ensemble des paires de modalités de chaque descripteur.

Cette préordonnance est ensuite quantifiée, en affectant un rang à chacun de ses éléments (constitué des deux modalités k et l). En présence d'ex æquo, on attribue aux éléments de la classe d'ex æquo, la moyenne arithmétique des rangs qu'ils auraient eus s'ils étaient totalement ordonnés. Ces rangs —dits *moyens*— présentent l'avantage d'avoir une somme constante, quel que soit le préordre choisi.

On procède enfin, à un "centrage-réduction" par le rang des paires de mêmes modalités pour aboutir à des rangs —notés r_{kl}^{ω}— dans l'intervalle [–1, +1].

3.1.2 Préordonnance calculée

Chaque descripteur ω est représenté par le graphe valué : $G_\omega = \langle M_\omega , \Gamma_\omega , f_\omega \rangle$, où : Γ_ω est un sous-ensemble de $M_\omega \times M_\omega$ et $f_\omega : \Gamma_\omega \to \mathbb{R}$ est une fonction de valuation caractérisant la structure associée à M_ω, et qui engendre un préordre total sur Γ_ω.

- $\mathbf{M_\omega}$ est sans structure particulière

Les modalités n'étant pas ordonnées, il en résulte que Γ_ω est réduit à $P_2(M_\omega)$, ensemble des *paires* de modalités et que deux objets décrits par un tel descripteur ne peuvent être que réunis s'ils possèdent la même modalité ou séparés dans le cas contraire. On montre alors que : $r_{kl}^\omega = \begin{cases} -1 & si \quad k \neq l \\ 0 & si \quad k = l \end{cases}$

- $\mathbf{M_\omega}$ est muni d'une structure booléenne

M_ω est composé des deux modalités vrai/faux (ou présence/absence) qui ne sont pas symétriques. La présence d'un tel paramètre chez un couple d'objets peut être plus significative que son absence (ou inversement), on en déduit : $r_{kl}^\omega = \begin{cases} -1 & si \quad k \neq l \\ 0 \ ou \ 1 & si \quad k = l \end{cases}$

- $\mathbf{M_\omega}$ est muni d'une structure ordinale

Etant donné que les modalités sont structurées ordinalement, il faut considérer ici l'ensemble des couples de modalités $\Gamma_\omega = M_\omega \times M_\omega$. Mais contrairement au premier cas, deux objets décrits par un tel descripteur ne sont pas seulement réunis ou séparés mais en cas de séparation, il faut tenir compte de son amplitude et de son sens. On démontre dans [OUA 91] que : $r_{kl}^\omega = \frac{(k-l)\left(2m_\omega - |k-l|\right)}{m_\omega^2 + 1}$

3.2 Matrices de codage

Chaque descripteur qu'il soit numérique ou à modalités, est représenté par une relation binaire sur l'ensemble d'objets O. Ainsi un descripteur ω définit une matrice de codage (ou de pondération) sur l'ensemble O × O, soit en considérant $I = \{1, ..., i, ..., n\}$ l'ensemble d'indexation de O : $C_\omega = \left(c_{ij}^\omega\right)_{(i,j)\in I\times I}$

3.2.1 Descripteurs à modalités disjointes

Pour un descripteur à modalités disjointes ω, un objet i (resp.j) possède une seule modalité k (resp.l). Le codage associé au couple d'objets (i, j) est naturellement le rang du couple de modalités (k, l) : $c_{ij}^\omega = r_{kl}^\omega$

3.2.2 Descripteurs à modalités non disjointes

Dans le cas d'un descripteur à modalités non disjointes, un objet peut être décrit par tout un ensemble de modalités. Afin que le codage de ces descripteurs ne soit pas biaisé par la disparité entre leurs nombres de modalités, on construit — pour chaque objet i — une distribution de fréquences sur l'ensemble des modalités de chaque descripteur ω : $\left\{ f_{k,\omega}^i = \frac{\Im_{k,\omega}^i}{m_\omega^i} \quad 1 \le k \le m_\omega \right\}$, où : $\Im_{k,\omega}^i = 1 / 0$ est l'indicateur de présence/absence de la modalité k du descripteur ω chez l'objet i et $m_\omega^i = \sum_k \Im_{k,\omega}^i$ représente le nombre total de modalités de ω que possède i.

On considère alors la moyenne arithmétique des rangs des différents couples de modalités possédées par le couple d'objets (i, j). On obtient donc le codage : $c_{ij}^\omega = \sum_{k,l} f_{k,\omega}^i f_{l,\omega}^j r_{kl}^\omega$

3.2.3 Descripteurs numériques

Dans le cas numérique, le codage doit tenir compte de l'écart entre les valeurs obtenues par les couples d'objets. Et afin de ne pas biaiser cette pondération par la différence entre les ordres de grandeur des descripteurs, on la réduit par l'écart maximum. On obtient ainsi pour chaque descripteur numérique ω, une matrice de codage à valeur dans [−1, +1] : $c_{ij}^{\omega} = \frac{\omega(i)-\omega(j)}{\max_k(\omega(k)) - \min_k(\omega(k))}$

4 Mesure de la similarité

4.1 Coefficient d'association

Pour comparer deux descripteurs ω et ϖ, on considère un coefficient d'association classique, produit scalaire des valuations engendrées par ω et ϖ sur l'ensemble des couples d'objets : $s(\omega,\varpi) = \sum_{i\neq j} c_{ij}^{\omega} c_{ij}^{\varpi}$

L'étude et la normalisation de ce critère ont fait l'objet de nombreux travaux, et ce, dans différents contextes. Le nôtre s'inscrit dans la méthode de la vraisemblance du lien [LER 92], où une hypothèse d'indépendance, consiste à associer à l'indice *s* une variable aléatoire S asymptotiquement normale.

4.2 Coefficient Centré-Réduit (CCR)

La normalisation statistique de l'indice (centrage-réduction par l'espérance μ et l'écart-type σ de la variable aléatoire *S*), permet la définition du Coefficient Centré-Réduit : $Q(\omega,\varpi) = \frac{s(\omega,\varpi)-\mu}{\sigma}$

Les expressions du CCR développées dans [OUA 91] sont fort complexes et nécessitent l'utilisation de moments factoriels ($n^{[2]} = n(n-1)$) centrés des matrices de codage :

- Covariance entre C_ω et C_ϖ : $\Phi_{\omega\varpi} = \frac{1}{n^{[2]}}\sum_{i\neq j} c_{ij}^{\omega} c_{ij}^{\varpi} - \left(\frac{1}{n^{[2]}}\sum_{i\neq j} c_{ij}^{\omega}\right)\left(\frac{1}{n^{[2]}}\sum_{i\neq j} c_{ij}^{\varpi}\right)$
- Variance de C_ω : $\Phi_\omega = \Phi_{\omega\omega} = \frac{1}{n^{[2]}}\sum_{i\neq j} c_{ij}^{\omega 2} - \left(\frac{1}{n^{[2]}}\sum_{i\neq j} c_{ij}^{\omega}\right)^2$
- Variance des marges de C_ω : $\Psi_\omega = \frac{1}{n^{[2]}(n-1)}\sum_i \left(\sum_{j,j\neq i} c_{ij}^{\omega}\right)^2 - \left(\frac{1}{n^{[2]}}\sum_{j,j\neq i} c_{ij}^{\omega}\right)^2$

L'une des formes simplifiées du CCR s'écrit : $Q(\omega,\varpi) = \frac{\sqrt{n}}{2}\frac{\Phi_{\omega\varpi}}{\sqrt{\Psi_\omega\Psi_\varpi + \frac{1}{2n}(\Phi_\omega - 2\Psi_\omega)(\Phi_\varpi - 2\Psi_\varpi)}}$

4.3 Expressions asymptotiques du CCR

Nous montrons par ailleurs que les moments Φ et Ψ sont asymptotiquement indépendants de *n*. D'où, lorsque *n* est suffisamment grand, le deuxième terme de la variance devient négligeable devant le premier.

Nous aboutissons ainsi à la *forme limite* du CCR : $Q(\omega,\varpi) = \frac{\sqrt{n}}{2}\frac{\Phi_{\omega\varpi}}{\sqrt{\Psi_\omega\Psi_\varpi}}$

Le recours enfin, à une « réduction géométrique » du CCR (en le divisant par la moyenne géométrique des deux coefficients diagonaux correspondants) permet d'obtenir sa *forme corrélative*, qui n'est autre que le coefficient de corrélation linéaire entre les matrices de codage C_ω et C_ϖ: $Q(\omega,\varpi) = \frac{\Phi_{\omega\varpi}}{\sqrt{\Phi_\omega\Phi_\varpi}}$

4.4 Expressions contingentielles du CCR

Pour les données à modalités, l'utilisation des tables de contingences permet de déplacer le support du codage de l'ensemble des couples d'objets à l'ensemble des couples de modalités. On obtient de la sorte des expressions dites *contingentielles* qui se prêtent mieux au calcul.
Pour appliquer ces expressions contingentielles aux descripteurs à modalités non disjointes, on introduit la notion de « *tables de contingences généralisées* » [OUA 00-a], où les termes ne sont plus des entiers mais des réels représentant les poids des couples de modalités : $\eta_{kl}=\sum_i f^i_{k,\omega} f^i_{l,\varpi}$

Un programme calculant les similarités dans le cas des données à modalités disjointes, intitulé AVARE (Association entre Variables Relationnelles) est disponible dans la bibliothèque MODULAD [OUA 00-b]. Une autre version du programme contenant le cas disjoint est opérationnelle et a été appliquée à des données réelles et a fourni des résultats concluants [OUA 91]. La version finale du programme, englobant le cas numérique, est en cours de développement.

5 Conclusion

Le traitement d'un ensemble hétérogène de paramètres de description nécessite dans un premier temps, la conception d'un système de représentation général et cohérent. La structure descriptive que nous avons présentée est assez souple pour transcrire des descripteurs aussi diverses, sans perdre de leurs caractéristiques.

D'autre part, tout système de codage aussi séduisant soit-il, ne peut être pertinent que s'il permet de comparer aisément ces descripteurs de natures différentes. Dans notre approche, les similarités entre les paramètres de description sont mesurées par un coefficient d'association, issu d'un critère classique et général, qui se révèle —dans sa forme asymptotique— être le coefficient de corrélation linéaire entre matrices de codage.

6 Bibliographie

[DID 91] DIDAY E. *Des objets de l'analyse de données à ceux de l'analyse des connaissances* Induction symbolique et numérique à partir de données, vol. 1, p. 9-75, Cépaduès-Editions, Toulouse, 1991.

[LER 92] LERMAN I.C. *Conception et analyse de la forme limite d'une famille de coefficients statistiques d'association entre variables relationnelles* Mat. Inf. Sci. hum. n° 118 p. 33-52. & n° 119 p. 75-100, 1992.

[OUA 91] OUALI ALLAH M. *Analyse en préordonnances des données qualitatives. Applications aux données numériques et symboliques* Thèse de l'université de Rennes I, 1991.

[OUA 00-a] OUALI ALLAH M. *Analyse relationnelle des données symboliques* Induction symbolique et numérique à partir de données, vol. 2, p. 247-262 Cépaduès-Editions, Toulouse, 2000.

[OUA 00-b] OUALI ALLAH M. *Programme pour le calcul des coefficients d'association entre variables relationnelles* La revue MODULAD n° 25 juin 2000.

L'arbre de régression multivariable: classification d'assemblages d'oiseaux fondée sur les caractéristiques de leur habitat

Marie-Hélène Ouellette[1], Jean-Luc DesGranges[2], Pierre Legendre[1], Daniel Borcard[1]

[1]Département de Sciences biologiques
Université de Montréal,
Case postale 6128, succursale Centre-ville,
Montréal (Québec) Canada, H3C 3J7

[2]Service canadien de la faune,
Case Postale 10100
Sainte-Foy (Québec), Canada, G1V 4H5

RÉSUMÉ. Les problèmes écologiques liés aux fluctuations des niveaux d'eau (p.ex. dues aux barrages) sont nombreux et souvent mal connus. L'analyse présentée ici s'insère dans un programme de recherche de la Commission mixte internationale de gestion des eaux des Grands Lacs et du Saint-Laurent (CMI). Elle porte sur les assemblages d'oiseaux le long du fleuve Saint-Laurent en corrélation avec leur habitat. Le but est d'utiliser ces assemblages comme bioindicateurs de l'état écologique des milieux riverains. La technique appliquée ici consiste en un arbre de régression multivariable portant sur une sélection de 128 sites. Cet arbre permet de distinguer 6 groupes de sites caractérisés par des assemblages d'oiseaux et des propriétés environnementales spécifiques.

MOTS-CLÉS : arbre de régression multivariable, catégories, répartition spatiale, espèces indicatrices, oiseaux, caractéristiques environnementales.

1 Introduction

La gestion du débit des cours d'eau au moyen de barrages perturbe l'environnement, notamment en homogénéisant la végétation au détriment de la diversité de ses habitants [DES en prépar.2]. Dans le cadre d'un programme de recherche de la Commission mixte internationale de gestion des eaux des Grands Lacs et du Saint-Laurent (CMI), Environnement Canada a étudié les assemblages d'oiseaux le long du fleuve Saint-Laurent ainsi que leur habitat afin d'utiliser ces assemblages comme bioindicateurs de l'état écologique des milieux riverains. Nous appliquerons ici la méthode de l'arbre de régression multivariable de [DEA 02] afin d'identifier les caractéristiques environnementales auxquelles les assemblages d'oiseaux répondent le plus fortement.

2 Matériel

Les analyses présentées ici portent sur 128 sites où ont été répertoriées 73 espèces d'oiseaux. Les observations ont eu lieu en l'an 2003 du lac Ontario jusqu'au lac Saint-Pierre. 95 variables environnementales ont été mesurées ou observées à chaque site. 59 d'entre elles sont issues d'analyses d'images satellitaires; les autres sont essentiellement des descripteurs de contextes paysagers dérivés de l'analyse de la végétation.

3 Méthodes

3.1 Principe de l'analyse de l'arbre de régression multivariable

Cette analyse permet de regrouper les objets multivariables de la matrice réponse en se basant sur des variables explicatives externes. C'est une forme de groupement sous contrainte qui réalise une succession de divisions (partitions) binaires des objets. Chaque partition des objets en deux groupes est faite de façon à minimiser l'impureté (ou erreur, ou statistique TESS [LEG 98]) de la variable réponse (autrement dit, maximiser l'homogénéité intragroupe). Chaque partition est définie par une seule variable environnementale (dite primaire [BRE 84]). Le processus se poursuit jusqu'à l'atteinte d'une partition en petits groupes d'objets. On émonde ensuite l'arbre obtenu en remontant vers la racine, jusqu'à atteindre la taille désirée, en testant chaque partition par validation croisée. Un arbre se décrit par sa taille (nombre de groupes) et son erreur relative [quotient de la somme, pour tous les groupes, de l'impureté totale des objets au sein de chaque groupe (somme des carrés d'écarts à la moyenne multivariable du groupe) sur l'impureté du nœud racine (somme pour tous les objets des carrés des écarts à la moyenne multivariable)]. Parce que cette mesure fournit une estimation trop optimiste des capacité prédictives de l'arbre, on recourt généralement à une autre mesure, l'erreur relative de la validation croisée. Elle varie de 0 (prédictions impeccables) à 1 (prédictions complètement erronées).

Après ces calculs, l'utilisateur peut encore décider de remplacer certaines des variables explicatives définissant les nœuds (variables primaires) par d'autres pour faciliter l'interprétation de l'arbre. Pour guider ce choix, on calcule des indices de similarité entre les noeuds et les autres variables. Ces indices tiennent compte de la répartition des objets d'un nœud par rapport à celle d'une autre variable explicative; on calcule le nombre d'objets qui changent de groupe/le nombre total d'objets, ou encore le pourcentage ajusté quantifié par le nombre d'objets qui changent de groupe/nombre d'objets dans le plus grand groupe du noeud [BRE 84]. L'utilisation de ces variables permet de modifier la topologie de l'arbre et parfois d'augmenter le pourcentage d'explication de la matrice réponse. Pour choisir l'arbre final, on a recours à un réseau d'arbres construits à l'aide de la validation croisée. L'arbre finalement retenu a la capacité de prédire des assemblages en fonction des variables explicatives, ou, à l'inverse, de prédire des caractéristiques environnementales en fonction de la structure de l'assemblage.

La méthode définit aussi des espèces délimitantes pour chaque nœud. Une espèce délimitante a une importante contribution à la variance expliquée de l'arbre à un nœud donné. Ces espèces sont les mieux expliquées (plus petite somme du carré des erreurs) par ce nœud, qui est lui-même caractérisé par une certaine variable environnementale. Cela permet d'identifier les espèces qui répondent le mieux aux variables primaires de l'arbre.

Dans la présente étude, nous avons regroupé les sites en fonction de leurs abondances d'espèces d'oiseaux et caractérisé les habitats des groupes par les variables environnementales décrites à la section *Matériel*.

3.2 Espèces indicatrices

Pour identifier les espèces indicatrices de chaque groupe, nous avons utilisé une méthode statistique de recherche des espèces indicatrices [DUF 97]. Dans cette méthode, les espèces indicatrices sont identifiées à l'aide d'un test par permutation. La statistique du test (*IndVal*) combine la fidélité des espèces (proportion de sites d'un groupe où l'espèce est présente) et leur spécificité (à quel point une espèce ne se trouve que dans le groupe considéré).

3.3 Associations d'espèces

Nous avons utilisé une analyse de concordance de Kendall pour identifier les associations significatives d'espèces, soit les espèces qui ont des distributions géographiques semblables [LEG 05]. Les espèces ont d'abord été divisées en 4 grands groupes par la méthode des *K* centroïdes (*K-means*). La partition a été faite à partir des vecteurs propres, normés à la racine carrée de leur valeur propre, d'une ACP du tableau des abondances d'espèces centrées réduites. Nous avons ensuite identifié, au sein de chaque groupe, les espèces qui ont une concordance (*W* de Kendall) significative avec les autres membres du groupe.

4 Résultats

À partir du réseau d'arbres obtenu (Fig. 1), nous avons retenu l'arbre ayant la plus faible erreur relative. Cet arbre explique 34 %, soit environ 3 % de variation de plus que l'arbre initialement choisi par l'algorithme. Les calculs ont été réalisés à l'aide de la librairie MVPART du langage R [R 04]. Le modèle final retenu (Fig. 2) indique que les variables environnementales délimitantes des 6 groupes (codés de A à F de gauche à droite) sont, en ordre décroissant de contribution au coefficient de détermination multiple : la densité de lisière de marécages arborés (ED_Marbo) qui sépare les groupes A, B, et C des groupes D, E et F ; l'axe géographique (axe.total[,1]) qui sépare les groupes D et E du groupe F ; le pourcentage de prairies humides (CP_P_hum) qui sépare le groupe A des groupes B et C ; la densité de lisière de prairies humides (ED_P_hum) qui sépare le groupe D du E ; et la densité de lisière de friches dans la place-échantillon (ED_Frich) qui sépare le groupe B de C. De A à F, le nombre de sites par groupe est 42, 46, 3, 17, 18 et 2. Dans le groupe A, le Bruant des marais, le Carouge à épaulettes et la Paruline masquée sont les espèces les plus abondantes. Aucune espèce n'est indicatrice au sens du test IndVal, n'étant ni assez spécifique ni assez fidèle. Selon la partition de la variance pour chaque nœud par espèce, le Troglodyte des marais apparaît comme étant l'espèce délimitante pour le nœud qui sépare le groupe A des groupes B et C. Dans le groupe B, les espèces les plus abondantes sont le Bruant des marais, le Carouge à épaulettes et le Troglodyte des marais. Ce groupe n'a aucune espèce indicatrice au sens de IndVal. Les espèces les plus délimitantes sont le Carouge à épaulettes et le Troglodyte des marais. Le groupe D, pour sa part, est représenté par une grande abondance de Merle d'Amérique et de Parulines jaunes. Aucune espèce n'est significative au sens de IndVal. La Paruline jaune est l'espèce la plus délimitante de son nœud. Le groupe E est caractérisé par un grand nombre d'espèces abondantes : la Paruline jaune, le Bruant chanteur, le Carouge à épaulettes, la Paruline masquée, le Merle d'Amérique, le Bruant des marais, l'Hirondelle bicolore et le Quiscale bronzé. Aucune espèce n'est indicatrice au sens de IndVal ; l'espèce la plus délimitante est la Paruline jaune. Les groupes C et F, bien qu'écologiquement intéressants, ne sont pas assez représentés dans cet échantillon (4 et 2 sites) pour justifier une discussion détaillée.

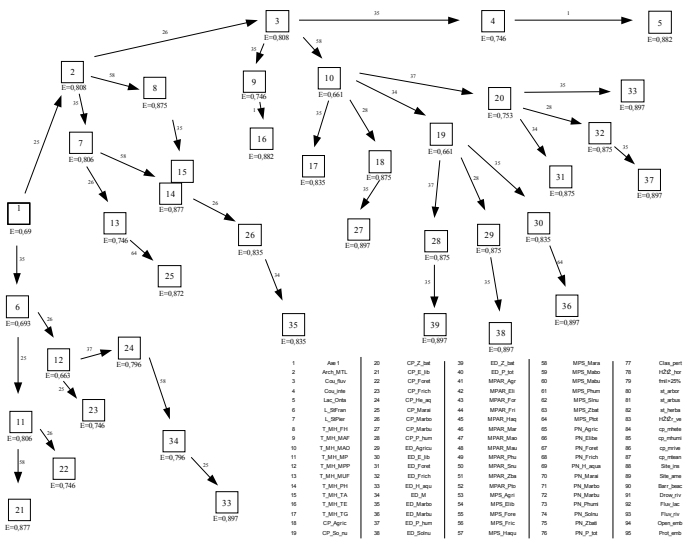


Figure 1. Réseau des arbres construits. Dans les cadres: numéros des arbres. Sur les flèches : variables primaires.

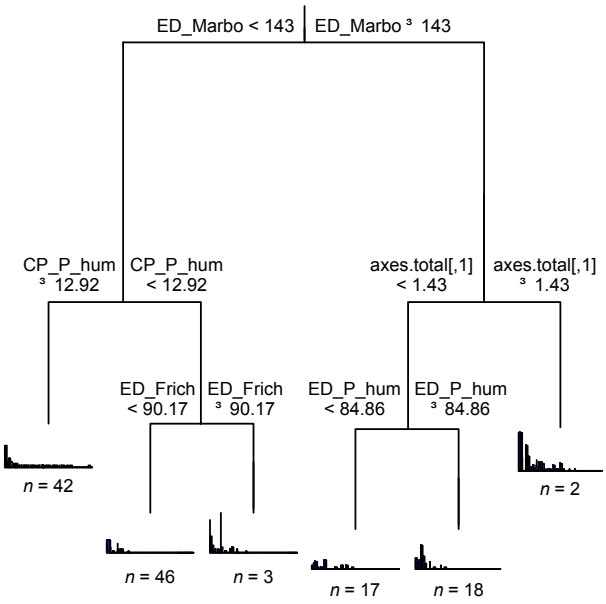


Figure 2. Arbre de régression multivariable final choisi à l'aide du réseau de la figure 1 (case 19).

5 Discussion

D'autres analyses ont été réalisées par DesGranges [DES en prépar.1], avec intégration explicite de descripteurs de l'hydrologie. Les résultats concordent avec ceux qui sont présentées ici dans la mesure où l'effet de l'hydrologie est en partie intégré par les descripteurs issus de l'analyse de la végétation. Ainsi, nos groupes A, B et C correspondent à des marais dépourvus d'arbres. Le groupe A se distingue des deux autres par l'absence de fluctuations de niveaux d'eau de longue durée au printemps. Les groupes B et C sont les plus inondés. L'identification du Troglodyte des marais comme espèce délimitante confirme les résultats obtenus d'autres analyses [DES en prépar.1]. Au pôle arboré du spectre, les groupes D, E et F représentent une toposéquence allant des sites les plus ouverts et humides (F) aux plus fermés (D).

6 Conclusion

L'arbre de régression multivariable est très intéressant lorsque l'objectif est de définir une typologie écologique qui soit non seulement explicative, mais aussi prédictive. Comparée à d'autres approches, cette méthode a le mérite de fournir un modèle d'apparence et d'interprétation simple, grâce à sa structure monothétique. Une telle caractéristique en fait un outil intéressant pour les praticiens de la conservation de l'environnement, pour lesquels l'efficacité repose sur des moyens de diagnostic et de décision simples et rapides. La comparaison des résultats présenté ici à ceux d'autres analyses réalisées sur les mêmes données montre que cette simplicité n'est pas obtenue au détriment de la qualité des résultats scientifiques.

7 Bibliographie

[BRE 84] BREIMAN L., FREIDMAN J., OLSHEN R., STONE C., *Classification and regression trees*, Chapman & Hall, 1984.

[CLA 92] CLARK L., PREGIBON D., "Tree-based models", Chambers J.M., Hastie T.J., editors, *Statistical models in S*, Wadsworth & Brooks, Pacific Grove, California, 1992, p. 377-420.

[DEA 92] DE'ATH G., FABRICIUS K., "Classification and regression trees; a powerful yet simple technique for the analysis of complex ecological data", *Ecology*, vol. 81, 2000, p. 3178-3192.

[DEA 02] DE'ATH G., "Multivariate regression trees: a new technique for modeling species-environment relationships", *Ecology*, vol.83, 2002, p. 1105-1117.

[DES en prépar.1] DESGRANGES J.-L., INGRAM J., DROLET B., SAVAGE C., BORCARD D., "Development of wetland bird assemblage predictive models and performance indicators for use in the environmental assessment of Lake Ontario and St. Lawrence River alternative water regulation plans", *Technical Report No 425*, Canadian Wildlife Service, Québec Region, en préparation.

[DES en prépar.2] DESGRANGES J.-L., LEHOUX D., DROLET B., DAUPHIN D., GIGUÈRE S. SAVAGE C., "Les oiseaux palustres : un groupe sensible aux conditions hydrologiques des zones humides du Saint-Laurent", Environnement Canada, *Série de documents d'évaluation de la science de la DGSAC*, en préparation.

[DIG 81] DIGBY P., GOWER J., "Ordination between and within groups applied to soil classification", *in: Down to earth statistics; solutions looking for geological problems*, Merriam D.F., editor, Syracuse University Geology Contributions, Syracuse, New York, 1981, p. 53-75.

[DUF 97] DUFRÊNE M., LEGENDRE P., "Species assemblages and indicator species: the need for a flexible asymmetrical approach", *Ecological Monographs*, vol.67, 1997, p.345-366.

[LEG 05] LEGENDRE P., "Species associations: the Kendall coefficient of concordance revisited", *Journal of Agricultural, Biological and Environmental Statistics*, 2005 (*sous presse*).

[LEG 98] LEGENDRE P., LEGENDRE L., *Numerical ecology. Second English edition*. Elsevier, Amsterdam, 1998.

[R 04] R DEVELOPMENT CORE TEAM, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, 2004.

Classification de cooccurrences de termes à l'aide d'un algorithme non supervisé de réseaux de neurones

Yann Prudent[1], Stéphane Trébucq[2]

(1) Laboratoire PSI,
Université et INSA de Rouen,
Mont-Saint-Aignan (Haute-Normandie), France, 76 821
(2) Centre de Recherche en Contrôle et Comptabilité Internationale,
Université Montesquieu Bordeaux IV IAE de Bordeaux,
Avenue Léon Duguit, Bordeaux(Gironde) France, 33 000

RÉSUMÉ. L'outil graphique proposé est fondé sur l'utilisation d'un algorithme ayant recours au principe de l'apprentissage Hebbien Compétitif. Une telle méthodologie est en mesure de proposer une vue synthétique des relations structurantes présentes dans des textes tout en écartant les relations à fréquence élevée, susceptibles d'être considérées comme triviales par un expert. Les vues synthétiques obtenues permettent d'isoler certaines grappes ou classes de termes associés. La sensibilité du nombre obtenu de classes et de la visualisation des relations sélectionnées aux différents paramètres de l'algorithme est ensuite développée.

MOTS-CLÉS : données textuelles, corpus, cooccurrences , réseaux de neurones, classification, visualisation

1 Introduction

L'analyse des liens de cooccurrence représente une démarche classique d'étude des corpus textuels. Cependant, les optiques retenues afin d'identifier ces types de relations sont extrêmement diverses. Cela dépend, en effet, non seulement de l'unité d'analyse choisie (phrases, paragraphes, textes), mais aussi des éléments analysés (termes lemmatisés ou non, syntagmes, classes sémantiques). Quelle que soit la méthode utilisée, l'analyste se retrouve inévitablement confronté à une masse de relations difficiles à traiter. Selon certaines approches, un filtrage statistique des relations les plus atypiques peut être obtenu [LAF 84]. En ce qui concerne la présente recherche, une liste de cooccurrences a été prise comme point de départ. Dans sa version actuelle, l'outil n'utilise que les fréquences de cooccurrence et permet de sélectionner uniquement certaines relations. La représentation graphique prend la forme d'une arborescence non hiérarchisée. Les termes associés sont alors regroupés en grappes, ces dernières correspondant à des classes, tout en conservant visuellement les liens de cooccurrence.

Cet article a pour objet d'expliciter le fonctionnement de l'algorithme utilisé, fondé sur la logique de l'apprentissage Hebbien compétitif [MAR 93]. Les propriétés d'un tel algorithme sont ensuite évaluées empiriquement, à partir d'un corpus test, composé de textes académiques dédiés à la gestion des entreprises, et plus précisément à leur gouvernance.

2 Présentation de l'outil de classification

Nous présenterons dans un premier temps la logique de l'algorithme (2.1.), puis sa mise en œuvre pratique (2.2.).

2.1 Un algorithme à base de réseau de neurones

L'outil proposé mobilise un algorithme fondé sur la logique des réseaux de neurones [TRE 05], en application des principes de l'apprentissage Hebbien compétitif [MAR 93]. Chaque terme est alors considéré comme un « neurone », et les relations de cooccurrence sont représentées graphiquement par un trait connectant chaque « neurone » à un autre. L'algorithme mis en œuvre procède à un tirage aléatoire au sein de l'ensemble S des relations décrivant l'intégralité du corpus étudié. La probabilité de tirage est proportionnelle à la fréquence de cooccurrence des termes. Dans le cas considéré, la mise en relation entre deux termes est directionnelle, sachant que le premier terme précéde le second dans le texte. L'algorithme présenté ci-après ne conserve que les connexions dont l'âge est inférieur au seuil a_{max}. Ce paramètre essentiel doit être fixé avant de lancer l'exécution du programme.

Données : un ensemble de relations S, a_{max}
Résultat : Une carte
Début
tant que le critère d'arrêt n'est pas satisfait **faire**
tirer au hasard une relation (x,y) ∈ S
// et ce proportionnellement au nombre de cooccurrences
si le neurone x n'existe pas dans la carte **alors**
créer le neurone x ;
si le neurone y n'existe pas dans la carte **alors**
créer le neurone y ;
incrémenter l'âge de toutes les connexions de x ;
incrémenter l'âge de toutes les connexions de y ;
si la connexion (x,y) existe **alors**
ramener son âge à 0 ;
sinon
la créer ;
supprimer toutes les connexions avec un âge supérieur à a_{max} ;
si ceci a pour conséquence l'isolement d'un neurone, **alors**
le supprimer ;
Fin

Selon cet algorithme, l'âge de chaque connexion débute avec une valeur de zéro. Lorsqu'une connexion, tirée précédemment et maintenue dans le réseau, est à nouveau tirée, son âge est réinitialisé. En revanche, lorsqu'un nouveau neurone apparaît, l'âge des connexions liées au neurone nouvellement connecté est incrémenté. L'application du critère a_{max} permet d'éliminer en conséquence les connexions les plus fréquentes, estimées comme triviales, et porteuses d'une information d'un intérêt limité pour l'analyste. Dans cette version, dès l'instant où toutes ces relations ont été « apprises », c'est-à-dire passées en revue un certain nombre de fois (voir paramètre « nbpasses », au point 2.2.), l'exécution du programme est alors stoppée.

2.2 Mise en œuvre de l'algorithme neuronal

La mise en œuvre de l'algorithme s'opère à partir d'une interface programmée en langage java. Le fichier proposé en entrée représente une liste de coocurrences, avec respectivement pour chacune d'entre elles la

fréquence d'occurrence, le premier terme, et le second terme, consécutif au premier. L'utilisateur est alors invité à déterminer cinq paramètres numériques, à savoir : « nbpasses » « âge » « min » « max », et « min grappe », sachant que :

- « nbpasses » est le nombre de fois où toutes les relations ont été passées en revue, critère d'arrêt évoqué précédemment ;
- « âge » correspond au paramètre a_{max} de l'algorithme ;
- « min » est un seuil correspondant à la fréquence de cooccurrence en dessous de laquelle les relations ne sont pas retenues et étudiées ;
- « max » est un seuil correspondant à la fréquence des cooccurrences au-dessus de laquelle les relations ne sont pas retenues et étudiées ;
- « mingrappe » détermine le nombre minimal de termes requis au sein d'une grappe pour que la représentation de celle-ci soit maintenue dans le graphe final. Ce dernier critère a été introduit afin de pouvoir éventuellement exclure des grappes constituées d'un nombre limité de termes et jugées d'une importance marginale.

3 Propriétés de l'algorithme neuronal

Les tests empiriques ont été menés à partir d'un corpus test de 32 textes scientifiques – soit environ 300 000 mots – rédigés par Gérard Charreaux, auteur réputé en finance d'entreprise. Le thème principal des textes traités est celui de la gouvernance des entreprises, à savoir l'ensemble des dispositifs aptes à limiter ou encadrer la marge de liberté des dirigeants afin qu'ils agissent bien dans le sens des intérêts des actionnaires ou des différentes parties prenantes de l'entreprise. A partir de ce corpus spécialisé, nous étudierons successivement les propriétés de classification de l'algorithme (3.1.), puis la qualité des représentations obtenues (3.2.), et enfin la stabilité des résultats obtenus (3.3.).

3.1 Visualisation et classification des cooccurrences

Grâce à l'outil Tropes, 8831 relations de cooccurrences[1] sont listées dans le corpus étudié. Ces dernières servent alors de base à la représentation graphique obtenue (voir Fig. 1).

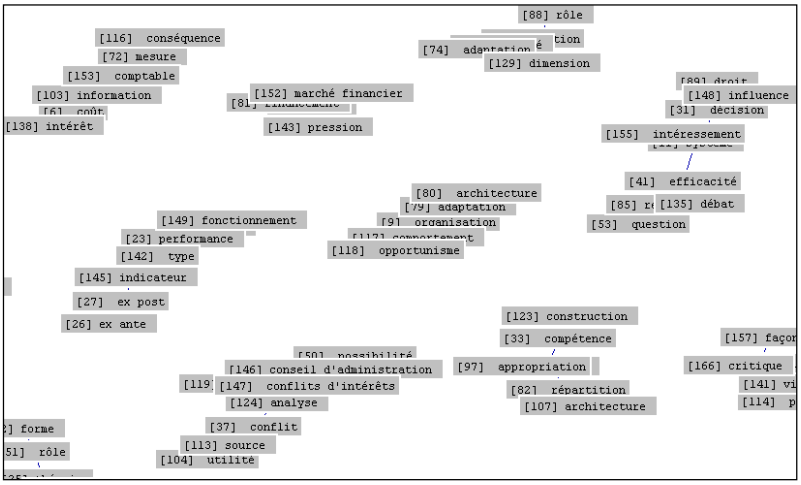


FIG. 1 : Sous-partie de l'écran de visualisation des grappes ou classes de termes

La figure 1 est obtenue avec les paramètres fixés comme suit : « nbpasses »=15, « âge »=3, « min »=1, « max »=245, et « mingrappe »=3. La représentation graphique est au total composée de 30 grappes de termes séparées et de 136 relations, soit 1,5% (136/8831) du total des relations de cooccurrences

[1] Leur distribution par fréquence d'occurrence se compose comme suit : N=11 pour l'intervalle de fréquence [100 à 245] ; N= 31 pour l'intervalle de fréquence [50 à 99] ; N=46 pour l'intervalle de fréquence [50 à 99] ; N=86 pour l'intervalle de fréquence [20 à 29] ; N=394 pour l'intervalle de fréquence [10 à 19] ; N=1120 pour l'intervalle de fréquence [5 à 9] et N=7043 pour l'intervalle de fréquence [2 à 4]. Comme on peut le constater, plus les intervalles de fréquences sont bas, plus les relations d'occurrence sont nombreuses.

initialement identifiées. Après une telle réduction, l'ensemble des relations s'avère beaucoup plus facile à gérer et à analyser.

3.2 Qualité des représentations obtenues

Il est intéressant d'observer, toujours à partir de la figure 1, que les relations retenues présentent une fréquence d'occurrence s'élevant au maximum à 48. Un tel paramétrage conduit *de facto* à écarter de la représentation les cooccurrences comptant parmi les plus fréquentes dans le corpus étudié (cf. note de bas de page n°1). Le graphe conserve 33 relations avec une occurrence de 4, 55 relations avec une occurrence de 5 à 9, 30 relations avec une occurrence variant entre 10 et 19, 12 relations avec une occurrence s'échelonnant entre 20 et 29, et 6 relations avec une occurrence allant de 30 à 48. Toutefois, un tel paramétrage n'assure pas la meilleure représentation du corpus. En effet, les grappes constituées sont d'une taille limitée. Il est donc nécessaire de jouer sur les paramètres afin d'accroître au mieux le nombre de relations sélectionnées.

3.3 Stabilité des résultats obtenus

Ainsi conçu, l'algorithme n'assure pas, pour un paramétrage donné, une stabilité parfaite des résultats. En effet, la sélection des relations repose sur un tirage aléatoire. Il n'est donc pas assuré que les graphes obtenus soient similaires, pour des paramètres fixés d'une manière rigoureusement identique. A titre expérimental, le programme a été ainsi lancé à cinq reprises avec des paramètres identiques. Pour les paramètres suivants {« nbpasses »=10, « âge »=3, « min »=1, « max »=245, et « mingrappe »=5} le nombre de grappes varie entre 8 et 15, et le nombre de relations sélectionnées entre 69 et 106. Une augmentation du paramètre « nbpasses » à 30 n'induit pas des résultats significativement différents[2] au niveau du nombre de grappes ou du nombre de relations choisies. Par ailleurs, si l'on compare les relations sélectionnées lors des cinq essais effectués, celles qui apparaissent au moins dans deux graphes représentent moins de 10% du total des relations apparaissant dans les graphes. En revanche, ce taux passe à 30% lorsque l'on joue sur le paramètre « âge » ou a_{max}.

Cependant, l'algorithme dans sa version la plus récente a été modifié afin de procéder à un tirage aléatoire sans remise. La stabilité des résultats a été ainsi améliorée significativement sans qu'il soit possible en l'état actuel, dans le cadre de cette présentation succincte, d'en présenter l'intégralité des résultats.

Par ailleurs, bien que la méthode ici proposée soit parfaitement reproductible sur tout type de corpus, il nous reste à mieux en estimer la qualité, au niveau des représentations obtenues, et ce, à partir d'un ensemble plus diversifié de corpus.

4 Conclusion

Les algorithmes à base de réseaux de neurones fournissent une solution utile pour les analystes souhaitant obtenir une représentation synthétique d'un ensemble de cooccurrences de termes. En outre, en fonction des valeurs affectées aux paramètres s'ajoutent des fonctions de classification des termes associés.

5 Bibliographie

[LAF 84] LAFON P., *Dépouillements et Statistiques en Lexicométrie*, Slatkine Champion, 1984.

[MAR 93] MARTINETZ T., "Competitive Hebbian learning rule forms perfectly topology preserving maps". In S. Gielen, B. Kappen (ed.), *Proceedings ICANN'93, International Conference on Artificial Neural Networks*, London : Springer, 1993, p. 427-434.

[TRE 05] TREBUCQ S., PRUDENT Y., ENNAJI A., " *Cooccurrences et cartes adaptatives : proposition d'un outil de visualisation et application à un corpus spécialisé* ", Actes du 3ème Atelier Visualisation et extraction de connaissances, Journées EGC (Extraction et Gestion de Connaissances), Paris, 18 janvier 2005.

[2] Cette absence de différence a été testée à partir d'une statistique non-paramétrique, selon le test de U Mann Withney.

Classification de parole en Question et NonQuestion par arbre de décision

Vũ Minh Quang[1,3] Eric Castelli[1], Alain Boucher[1,2], Laurent Besacier[3]

[1]*Centre de recherche international MICA*
Institut Polytechnique de Hanoi
1,Dai Co Viet
Hanoi – Viet Nam
{eric.castelli,minh-quang.Vu}@mica.edu.vn

[2]*Institut de la Francophonie pour l'Informatique*
Bât. D, ruelle 42, rue Ta Quang Buu
Hanoi – Viet Nam
alain.boucher@auf.org

[3]*Laboratoire CLIPS/IMAG,*
Université Joseph Fourier
385, rue de la Bibliothèque
BP53, 38041 Grenoble cedex 9, France
laurent.besacier@imag.fr

RÉSUMÉ. Le signal de parole contient, en plus des informations linguistiques, des informations extra-linguistiques qui présentent un aspect sémantique important, pouvant s'avérer fort utile surtout dans les applications telles que la recherche d'information, la réalisation automatique de résumés de documents parlés… L'étude de l'extraction de ces meta-informations à partir de signaux de parole se développe largement actuellement. Une expérimentation de classification de parole en deux catégories, Question(Q) et NonQuestion(NQ), en utilisant un arbre de décision est présentée. Sur un corpus en langue française nous avons expérimenté des conversations de « réunions de projet » et d'« entretiens d'embauche ». Notre approche originale utilisation un arbre de décision comme moteur de classification présentent des premiers résultats satisfaisants : 84,5% de taux de reconnaissance représentant l'exactitude de bonne classification des segments de parole Q, et NQ.

MOTS-CLÉS : Classification de parole, Arbre de décision, Recherche d'information

1 Introduction

Face au volume croissant des données audio disponibles, les systèmes d'indexation et de classification deviennent indispensables, afin de pouvoir localiser le plus rapidement possible les enregistrements désirés. Ces dernières années, de nombreuses études ont vu le jour dans ce domaine. Lie Lu [LU 01] a démontré avec succès que son système est capable de classifier un flux audio en parole, musique, bruit de fond et silence par un processus à deux phases : la première phase de classification consiste à séparer parole/non parole, alors que la deuxième phase discrimine ensuite le signal audio en musique, bruit de fond et silence avec un classifieur par règles. Un autre système [ZHA 98] utilise de nombreux paramètres complexes pour classifier et segmenter du signal audio en parole, musique, quelques types de bruits environnementaux et silence. Ces systèmes utilisent usuellement des méthodes de classification comme GMM (Gaussian Mixture Model), BP-ANN (Back Propagation Artificial Neural Network) et KNN (K-Nearest Neighbor). Dans cette étude, nous proposons un nouveau système de classification basé sur l'utilisation d'un arbre de décision. Le but est de classifier le signal de parole en deux catégories : *question* et *non question*. A la différence des travaux qui manipulent le signal audio d'une manière globale, nous allons traiter dans notre expérimentation un seul type de signal : la parole. Nos résultats obtenus pourraient

être appliqués à des domaines tels que la gestion de documents sonores, la réalisation automatique de résumés de discours ou de réunions, la recherche d'informations, parce que les segments de parole autours d'une question contiennent généralement des informations pouvant s'avérer très utiles dans ces applications.

Cet article est composé comme suit : le corpus est présenté dans la section 2, les détails du système de classification sont présentés section 3, et enfin, les sections 4 et 5 présentent respectivement les résultats obtenus et la conclusion du travail.

2 Le corpus DELOC

Nous utilisons le corpus du projet DELOC mené dans notre laboratoire, dont le but consistait à étudier différents types de réunions, ainsi que les différentes façons de s'exprimer (comportements langagiers selon les types de réunions). Le but du projet est de proposer des outils « collaboratifs » associés à la visioconférence, ou à n'importe quel contexte de réunions délocalisées, c'est-à-dire des outils d'aide à la rédaction du compte rendu en fin de réunion, ou d'aide à la transcription. Ce corpus se compose de différents types de réunions délocalisées réalisées par téléphone : 1) « brainstorming » ou remue-méninges, 2) (pré-)entretien d'embauche, 3) réunion de projet. Ces enregistrements ont été segmentés manuellement en phrases qui correspondent chacune à une *question* ou une *non question* : au total 852 phrases dont 295 phrases *questions* et 557 phrases *non questions*. Les phrases de courte durée correspondent à : « *Allo?* », « *D'accord* »…alors que celles à durée longue correspondent par exemple à : « *parce que chez Multicom, j'imagine qu'il y a quand même...il y a quand même des gens qui pourraient peut être compléter ?* ». Notre système peut traiter non seulement de phrases à rythme normal, mais encore celles dont le rythme est hésitant.

3 Le système de classification en *question* et *non question*

3.1 Structure globale du système

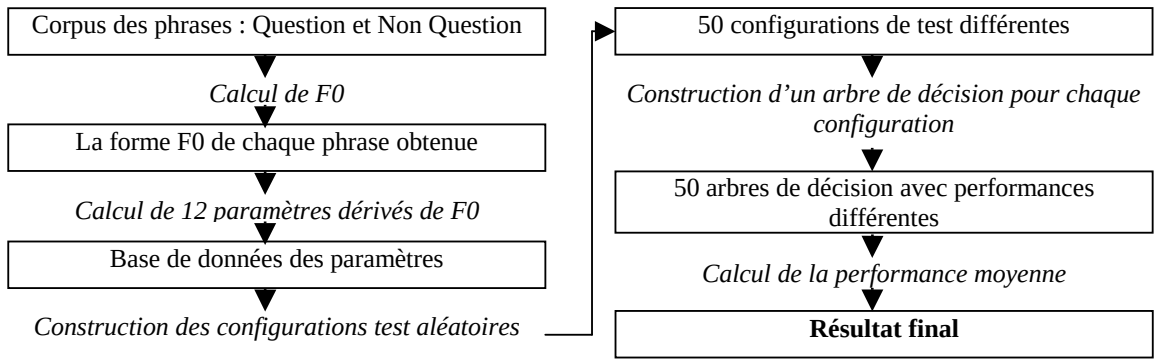


Figure 1. Structure globale du système de classification.

La structure globale du système de classification est illustrée à la figure 1. La classification commence par le calcul, pour toutes les phrases du corpus, de la fréquence fondamentale (F0 ou intonation) du signal de parole parce que la variation de l'intonation est l'une des caractéristiques principales qui différencient les types de phrases parlées. Puis, dans le but de caractériser cette intonation par un ensemble de paramètres, 12 paramètres dérivés des valeurs instantanées de F0 sont calculés. Nous construisons alors, afin de faciliter la gestion, une base de données qui comprend toutes les phrases et, pour chacune d'entre elles, ses paramètres associés.

3.2 Les paramètres de caractérisation utilisés

A la différence des travaux récents dans le domaine [LU 01, FER 03, WAN 03] qui utilisent des vecteurs sophistiqués de paramètres acoustiques tels que *la fréquence fondamentale (F0), le taux de passage par*

zéro (zero-crossing rate ZCR), le rapport d'énergie à court terme (low short-time energy ratio LSTER), le flux spectral (spectrum flux SF), nous utilisons dans notre expérimentation uniquement le paramètre F0 qui est calculé directement à partir du signal, en découpant celui-ci en fenêtres de 20ms. Cependant, à partir de F0, d'autres paramètres peuvent en être dérivés et nous proposons un ensemble de 12 paramètres listés dans le tableau 1.

No	Paramètre	Description
1	Min	Valeur minimale de F0
2	Max	Valeur maximale de F0
3	Range	Gamme de F0 pour la phrase entière (Max-Min)
4	Mean	Moyenne des valeurs de F0 d'une phrase
5	Median	Médiane des valeurs F0 d'une phrase
6	HighGreaterThanLow	Est-ce que la somme des valeurs F0 dans la première moitié de la phrase est supérieure à celle des valeurs F0 dans la dernière moitié ?
7	RaisingSum	Somme des $F0_{i+1} - F0_i$ si $F0_{i+1} > F0_i$
8	RaisingCount	Combien de $F0_{i+1} > F0_i$
9	FallingSum	Somme des $F0_{i+1} - F0_i$ si $F0_{i+1} < F0_i$
10	FallingCount	Combien de $F0_{i+1} < F0_i$
11	IsRaising	Est-ce que la forme F0 est montante ? (oui/non) Teste si RaisingSum > FallingSum
12	NonZeroFrameCount	Combien de valeurs de F0 sont non nulles ?

Tableau 1. Les 12 paramètres dérivés de F0.

Nous pouvons remarquer que ces paramètres peuvent se diviser en deux catégories distinctes : les 5 premiers paramètres sont des statistiques sur la valeur de F0, alors que les 7 derniers caractérisent le contour (la forme) de l'évolution de F0 (contour montant ou descendant). L'utilisation de ce deuxième groupe de paramètres est nouvelle et originale et, à notre connaissance, n'a jamais fait l'objet de publication. C'est l'utilisation de ces 7 derniers paramètres qui constitue l'originalité de notre méthode, au niveau de la caractérisation du signal de parole.

3.3 Méthode de classification par l'arbre de décision

Traditionnellement, les méthodes statistiques telles que les modèles de Markov cachés *(HMM)* ou les modèles de mixtures de Gaussiennes *(GMM)* et leurs variantes sont utilisées en Traitement Automatique des Langues Naturelles (TALN). L'arbre de décision est une méthode classique d'apprentissage [FRA 99] s'utilisant de manière similaire : le processus entier se compose de deux phases séparées, apprentissage et test. L'apprentissage consiste à construire un modèle représentant un ensemble des éléments, alors que le test utilise ce modèle pour évaluer un nouvel élément inconnu. Ces dernières années, beaucoup de travaux en TALN ont adopté des solutions d'apprentissage. Cependant, depuis les années 2000, une tendance vers une utilisation mixte des 2 types d'algorithmes s'affirme [MAR 00]. L'arbre de décision est une approche *diviser-et-conquérir* pour le problème de l'apprentissage à partir d'un ensemble d'éléments indépendants (un élément concret est appelé une *instance*). Un nœud dans l'arbre consiste à tester une condition particulière qui, en général, compare la valeur d'un attribut avec une constante, ou compare ensemble deux attributs, ou encore utilise des fonctions mathématiques d'un ou plusieurs attributs. La feuille d'arbre donne soit une classification des éléments satisfaisant toutes les conditions menant à cette feuille, soit un ensemble de classification, ou soit une distribution probabiliste sur toutes les classifications possibles. Pour classifier une *instance* inconnue, l'algorithme teste les attributs dans les nœuds jusqu'à ce qu'il atteint une feuille. Là, cette instance est classifiée selon la classe attribuée à la feuille. L'implémentation de l'algorithme provient du logiciel *open-source* Weka (http://www.cs.waikato.ac.nz/~ml/weka/) qui comprend les algorithmes de *classification, régression, clustering, règles d'association* écrits en Java.

4 Résultats expérimentaux

Il y a 295 phrases *question* et 557 phrases *non question* dans le corpus. Nous appliquons la méthode *50-folds cross validation*, c'est-à-dire nous répétons 50 fois le processus de division aléatoire du corpus en deux parties : une pour l'apprentissage (200 *questions* et 200 *non questions*), une pour le test (le reste : 95

Question	Non Question	←classifier comme
184(92%)	16(8%)	Question
27(13%)	173(87%)	Non Question

Tableau 2. Matrice de confusion sur les données d'apprentissage.

Question	Non Question	←classifier comme
73(77%)	22(23%)	Question
93(26%)	264(74%)	Non Question

Tableau 3. Matrice de confusion sur les données de test (valeurs moyennes).

	Précision	Rappel	F_ratio
Moyenne	44,2%	76,5%	55,7%
Ecart-type	4%	7,2%	3,5%

Tableau 4. Les mesures moyennes sur les données test de la classe question.

questions et 357 *non questions*). On obtient alors 50 arbres, chacun présentant un *taux de bonne classification* différent. Il ne reste plus qu'à calculer le taux moyen de ces arbres. La construction de l'arbre se fait rapidement, l'évaluation de l'arbre sur les données de test est aussi rapide (en dixièmes de millisecondes). Avec un écart-type de 2,4%, le taux moyen de classification obtenu est de 84,5% (c'est-à-dire 84,5% des instances sont correctement classifiées). Le tableau 2 présente en plus de détails un des meilleurs cas de classification sur les données d'apprentissage réalisé par notre système.

Pour l'évaluation sur les données de test, du fait que les nombres de *questions* et de *non questions* ne sont pas égaux (97 vs 357), nous devons évaluer les notions de *précision*, *rappel*, *F_ratio* de la classe *question* (voir tableau 4), avec la matrice de confusion donné dans le tableau 3. L'indice *F_ratio* mesurant le taux de bonne classification, nous permet de conclure que le système est assez performant (55,7%). D'ailleurs, il est à la fois stable avec l'écart-type maximum ne dépassant pas 3,5% .

5 Conclusion

Nous avons présenté une nouvelle méthode de classification de parole en *question* et *non question* par l'utilisation d'un arbre de décision. Dans notre expérimentation, un seul paramètre prosodique F0 est calculé directement à partir du signal, et 12 autres paramètres sont dérivés de F0 pour construire l'arbre de décision. Ce résultat peut s'appliquer pour d'autres applications en parole telles que le résumé automatique, la navigation ou la recherche d'information, car les zones autour d'une question contiennent souvent des informations importantes à identifier. Afin d'augmenter la performance du système, davantage de paramètres sont à étudier. D'autres classes avec intonations peuvent aussi être analysées comme les exclamations, les ordres, etc.

6 Bibliographie

[LU 01] LU L., JIANG, H., ZHANG H.J., "A Robust Audio Classification and Segmentation Method", *9th ACM Int. Conf. on Multimedia*, 2001, pp.203-211.

[ZHA 98] ZHANG T., KUO C.C.J., "Content Based Classification and Retrieval of Audio", *SPIE's 43rd Ann. Meeting-Conf. on Advanced Signal Processing Algorithms, Architectures and Implementations VII*, SPIE Vol. 3461, San Diego, july 1998, pp. 432-443.

[FER 03] FERRER L., SHRIBERG E., STOLCKE A., "A Prosody-Based Approach to End-of-Utterance Detection That Does Not Require Speech Recognition", *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. I, Hong Kong, 2003, pp. 608-611.

[WAN 03] WANG D., LU L., ZHANG H.J., "Speech Segmentation Without Speech Recognition", *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol I, april 2003, pp. 468-471,.

[FRA 99] WITTEN I.H., FRANK E., *Data mining: Pratical machine learning tools and techniques with Java implementations*, Morgan Kaufmann, 1999.

[MAR 00] MARQUEZ L., Machine learning and Natural Language Processing, Technical Report LSI-00-45-R, Universitat Politechnica de Catalunya, 2000.

Une méthode graphique pour interpréter et représenter des classes

Ken Reed

Centre for Business Research,
Deakin University,
221 Burwood Highway
Burwood VIC 3125
Australia

RÉSUMÉ. Dans ce papier, une technique simple pour représenter et interpréter des classes est proposée. La méthode comprends deux phases et s'appuie sur un centrage-réduction des variables initiales et sur une analyse factorielle d'un tableau croisé qui contient les sommes des variables initiales par classe. Il en résulte un graphique qui met en évidence les différences entre les classes.

MOTS-CLÉS : Classification, représentation graphique des classes

1 Introduction

Une des plus importantes étapes d'une classification est l'attribution d'un nom aux classes ainsi que leur interprétation. La méthode habituelle pour identifier des classes est de faire une analyse discriminante *a posteriori*, en utilisant les variables initiales, ou des variables exogènes, pour prédire les classes.

Je propose une technique simple pour représenter les profils des classes sous une forme visuelle qui en facilite l'identification et l'interprétation par une audience non-initiée. Plusieurs recherches récentes ont montré la nécessité d'ordonner les données complexes dans les graphiques pour mieux les comprendre ([HUR 04], [CAR 96] et [FRI 03]). L'idée est que les représentations graphiques sont meilleures quand les variables, objets et catégories semblables sont présentées proches ou voisines. Par contre, les variables, objets et catégories différentes doivent être représentées éloignées.

La méthode comprend deux phases pour faciliter l'interprétation des classes et répondre à deux besoins: le besoin de représenter les profils des classes par rapport au profil de l'échantillon entier (le profil moyen), et le besoin de représenter les profils des classes d'une façon qui accentue les différences de base entre elles. La première phase consiste à construire un graphique qui représente les moyennes des variables normalisées pour chaque classe. La deuxième phase

consiste à modifier ce graphique à la lumière des résultats d'une analyse des correspondances d'un tableau croisé qui contient les sommes des variables initiales par classe.

2 Le méthode

2.1 Le contexte des exemples

Les exemples sont tirés d'un projet décrivant des profils caractéristiques d'emploi du temps. Les données sont extraites de journaux dans lesquels on a enregistré les activités effectuées pendant des périodes de dix minutes, pendant une ou deux journées [ABS 97]. Les individus et les journées sont sélectionnés de manière à obtenir un échantillon représentatif de toutes les journées d'une année (1997) en Australie. Chaque activité est catégorisée selon un niveau général et un niveau spécifique. La classification est faite à partir des niveaux spécifiques, mais les exemples sont illustrés à partir du niveau général. On a trouvé quatre classes.

2.2 Identification des classes

La première phase est la représentation des profils des classes par rapport au profil de l'échantillon entier. Cette étape est réalisée en utilisant en centrant et réduisant les variables initiales (scores z). La composition des classes est représentée par un graphique sur lequel figure la moyenne des scores z pour chaque catégorie d'emploi du temps et chaque classe. Le graphique (figure 1) montre l'avantage d'utiliser des scores z. La classe 4, par exemple, est surtout caractérisée par le temps consacré aux études, et peu (moins que la moyenne) par des activités relative à un emploi et des tâches domestiques.

La deuxième phase de la représentation des profils des classes est une manière d'accentuer les différences de base entre les classes. L'analyse des correspondances est utilisée pour faire cela [GRE 84]. Comme précédemment, un tableau est construit à partir des catégories d'emploi du temps et des classes, mais les valeurs sont les sommes des temps pour chaque catégorie et pour chaque classe. Le premier axe de l'analyse des correspondances de ce tableau représente les différences entre les classes quant aux catégories d'emploi du temps, et les différences entres les catégories quant à la composition des classes. Les classes et les catégories sont ensuite réarrangées selon le premier axe de l'analyse des correspondances et un graphique analogue à celui de la figure 1 est construit (figure 2).

Rappelons qu'une analyse des correspondances produit une échelle des lignes (ici les classes) en fonction des colonnes, et une échelle des colonnes (ici les activités) en fonction des lignes d'un tableau de données. Lorsque les classes sont rangées selon le premier axe d'une analyse des correspondances, on peut voir immédiatement le contraste entre les classes qui sont les plus différentes, soit ici les classes 3 et 4, c'est-à-dire les étudiants et les employés. De plus, sur la base de l'ordonnancement des activités, on peut mettre en évidence les différences à l'intérieur des classes.

Cet exemple ne comprend qu'un petit nombre de classes et de variables pour construire ces classes. Pour appliquer la méthode à des analyses plus complexes, il est préférable d'utiliser une suite de graphiques, chacun représentant un axe d'une analyse des correspondances. Dans ce cas, pour chaque graphique, on n'utilise que les classes et les variables qui contribuent fortement à l'axe.

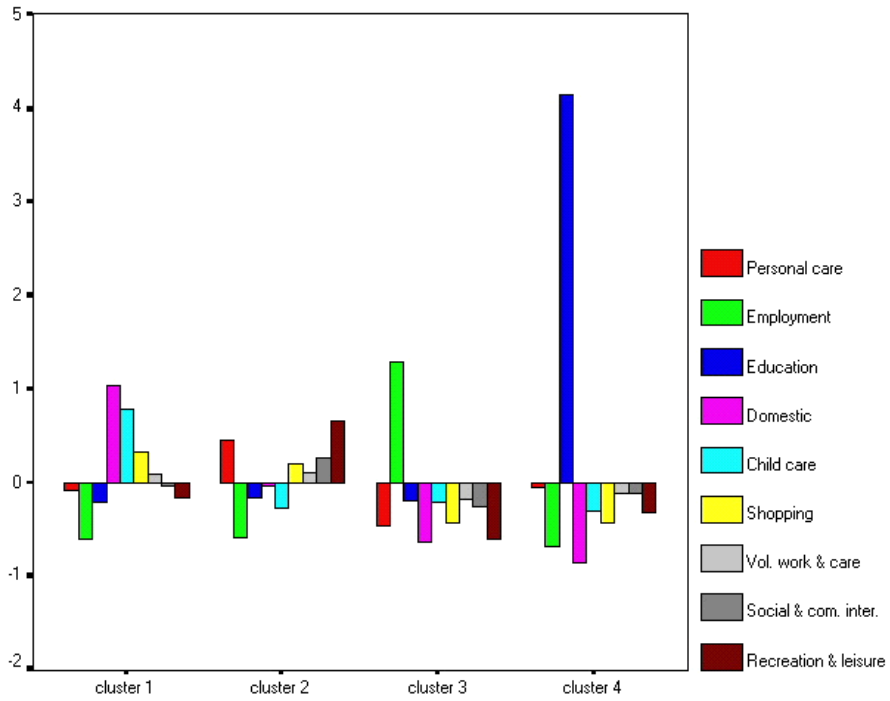


Figure 1. Les profils des classes selon l'emploi du temps

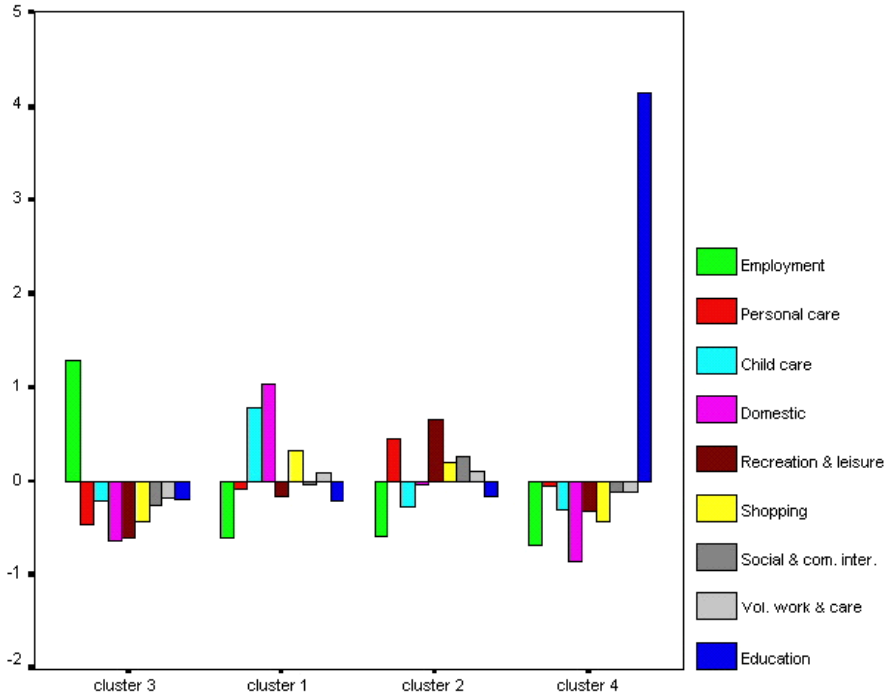


Figure 2. Les classes et les catégories réarrangés selon le premier axe d'une analyse des correspondances

3 Bibliographie

[ABS 97] AUSTRALIAN BUREAU OF STATISTICS "Time Use Survey, Australia 1997" [computer file].

[CAR 96] CARR, DB and OLSEN, AR ‘Simplifying Visual Appearance by Sorting’ Statistical Computing and Graphics Newsletter, vol 7, 1996, pp10-17.

[FRI 03] FRIENDLY, M and KWAN, E ‘Effect Ordering for Data Displays’, Computational Statistics and Data Analysis, vol. 43, 2003, pp 509-539.

[GRE 84] GREENACRE, M.J., Theory and Applications of Correspondence Analysis, Academic Press, London.

[HUR 04] HURLEY, CB ‘Clustering visualizations of multidimensional data’ Journal Of Computational and Graphical Statistics, vol. 13:4, 2004, pp 788-806.

Partition des centres mobiles pour données qualitatives

Maurice Roux

Faculté des Sciences et Techniques (Case 462)
Université Paul Cézanne (Marseille 3)
Avenue Normandie-Niemen
13397 Marseille Cedex 20, France

RÉSUMÉ. On présente un nouvel algorithme de partitionnement autour de centres mobiles (k-means) pour des données qualitatives, basé sur la métrique du Khi-carré. Cet algorithme est comparé à trois autres techniques similaires de partitionnement autour de centres mobiles en utilisant des jeux de données réelles et simulées. Les résultats sont évalués par le critère de l'inertie interclasse.

MOTS-CLÉS : agrégations autour de centres mobiles, métrique du Khi-carré, inertie interclasse, partitions.

1 Introduction

Un certain nombre de méthodes de classification ont été proposées pour traiter les données qualitatives [GOV 84, KAU 90, HUA 98]. Mais très peu ont utilisé la métrique du Khi-carré pourtant couramment utilisée avec succès, avec l'Analyse factorielle des Correspondances (AFC, [BEN 73]), bien adaptée à ce type de données. Reprenant une idée esquissée par Ralambondrainy [RAL 95] l'algorithme que nous exposons ici, applique le mécanisme usuel de réallocation-recentrage de la méthode des k-moyennes [FOR 65, MAC 67], sur les « profils » des objets, lesquels sont munis de poids, comme en AFC. Dans le paragraphe suivant on présente quelques algorithmes usuels pour traiter des variables qualitatives. Puis on décrit les étapes du nouvel algorithme. Ensuite on propose une évaluation de cet algorithme. Enfin on termine par une courte conclusion.

2 Quelques méthodes classiques pour traiter des données qualitatives

2.1 Pré-traitement par l'AFC [ROU 85]

La méthode consiste à traiter par l'AFC les données, mises sous forme disjonctive complète si nécessaire. On récupère ensuite les coordonnées factorielles pour les introduire comme variables quantitatives dans le programme de classification. La seule difficulté de cette méthode réside dans le choix du nombre d'axes factoriels à retenir pour définir les données soumises à la classification. Un certain nombre de règles empiriques peuvent aider l'utilisateur dans ce choix [SAP 93].

2.2 Utilisation des K-médoïdes [KAU 90]

Dans leur méthode PAM Kaufman et Rousseeuw travaillent directement sur une matrice de distances. Les représentants des classes sont les individus les plus centraux de ces classes, appelés "médoïdes", qui

minimisent la somme des distances aux autres objets de la classe. Nous avons adapté ce principe dans le cadre d'une procédure de réallocation-recentrage. Dans la phase de réallocation les objets sont affectés à la classe dont le médoïde est le plus proche. N'importe quelle distance peut être prise comme point de départ mais nous avons choisi la distance du Khi-carré en concordance avec les autres méthodes étudiées.

2.3 Méthode des k-modes [HUA 98]

Dans cette méthode les représentants des classes sont des objets artificiels, appelés k-modes, repérés par les mêmes variables que les objets réels. Leurs composantes sont les modalités de fréquence maximale dans leur classe. Cette définition est quelque peu ambigüe car il peut y avoir plusieurs modalités de même fréquence au sein d'une même classe ; dans ce cas l'une d'elle est choisie arbitrairement.

La distance d(i, k), entre un objet i et un objet modal, représentant la classe k, est égale au nombre de variables pour lesquelles les modalités de l'objet i et du représentant de k sont différentes. Il s'agit, en fait, de la distance L_1 entre les objets décrits par les indicatrices des modalités de variables. L'auteur montre que la méthode converge et minimise (localement) la somme des distances entre les individus et leurs objets modaux respectifs.

3 Le nouvel algorithme : méthode des k-profils.

On appelle x_{ij} les valeurs (zéro ou 1) du tableau X des données. La masse $x_{i.}$ d'un individu i est donnée par la somme des valeurs des variables pour cet individu ; elle est donc égale au nombre de 1 présents dans le vecteur décrivant l'individu. Le profil P(i) d'un objet i est donné par la suite des rapports de ses valeurs à sa masse :

$$P(i) = \{ x_{i1}/x_{i.}, ..., x_{ij}/x_{i.}, ...x_{ir}/x_{i.} \}$$

Le centre de gravité général G, du solide constitué par les profils des objets, munis des masses correspondantes, a pour j-ème coordonnée :

$$g_j = (1/x_{..}) \Sigma_i x_{i.} (x_{ij}/x_{i.}) = (1/x_{..}) \Sigma_i x_{ij} = x_{.j} / x_{..}$$

où $x_{..}$ représente la masse totale du solide et $x_{.j}$ désigne la somme des valeurs de la modalité de variable j, c'est à dire la fréquence de cette modalité. Un calcul analogue montre que le centre de gravité G_k d'une classe k est représenté également par un profil [JAM 78] :

$$G_k = \{x_{k1} / x_{k.} , ... x_{kj} / x_{k.} , ... x_{kr} / x_{k.} \}$$

où x_{kj} désigne la fréquence de la modalité j au sein de la classe k et $x_{k.}$ est la somme de toutes ces fréquences sur l'ensemble de toutes les modalités de variable. Et l'on vérifie facilement que le centre de gravité général G est bien égal à la moyenne pondérée des centres de gravité des classes de la partition.

Comme les autres méthodes d'agrégations autour de centres mobiles, notre algorithme se compose d'une phase de recentrage et d'une phase de réaffectation des objets aux classes. Une classe est représentée par son centre de gravité, tel qu'il a été défini ci-dessus, c'est à dire une sorte de profil moyen de la classe en question. Les objets sont ensuite réaffectés à la classe dont le centre de gravité est le plus proche au sens de la formule du Khi-carré :

$$d^2(i, k) = \Sigma_j (1/x_{.j}) [x_{ij}/x_{i.} - x_{kj}/x_{k.}]^2$$

dans laquelle chaque modalité de variable est pondérée par l'inverse de sa fréquence $x_{.j}$. Il est clair que cette procédure n'est qu'un cas particulier de la procédure générale d'agrégation autour de centres mobiles.

Donc cet algorithme converge et optimise le moment d'ordre 2, ou inertie inter-classe. Dans notre cas ce moment n'est autre (à un coefficient près) que le critère du Khi-carré de contingence entre la partition K et l'ensemble J des modalités de variables.

4 Evaluation du nouvel algorithme.

4.1 Application à des données connues (Critère externe).

Le premier jeu de données, que nous appelons PHYTOS (pour phytosociologie), est constitué de 16 relevés floristiques caractérisés par la présence ou l'absence d'un ensemble de 66 espèces [ROU 85]. De nombreux travaux sur ces données nous ont conduits à une partition en 4 classes que nous considérons comme « bonne ». Cette partition nous servira de référence dans les comparaisons ci-dessous.

Le second jeu de données, que nous appelons BOUCLES, décrit un ensemble de 59 plaques métalliques ornementées soutenant des boucles de ceintures. Ces boucles proviennent de fouilles archéologiques et sont d'époque médiévale (6-ème, 8-ème siècle). Elles sont décrites par 29 types de décorations en présence ou absence [LER 80]. Les auteurs de ce travail proposent plusieurs partitions, dont une en 5 classes qui nous servira de référence.

Un troisième jeu de données a été obtenu par simulation. Nous avons fabriqué une matrice de données en 0-1 constituée de blocs à prédominance de 1 (avec probabilité 0,8) et d'autres blocs à prédominance de zéros (avec probabilité 0,8 également) à la manière de Govaert [GOV 84]. Le tableau, que nous appelons BLOCS, comporte 100 objets repérés par 30 caractères. La classification porte sur les 100 objets.

4.2 Comparaison avec d'autres méthodes (Critère interne).

Les trois autres méthodes de classification évoquées au paragraphe 2 ci-dessus ont été mises en concurrence avec le nouvel algorithme. Les partitions obtenues par chacune des 4 méthodes sont évaluées par le critère de l'inertie interclasse, calculée selon la métrique du Khi-carré, et appliquée aux données initiales. Dans les trois jeux de données le tableau brut est traité directement, sans disjonction des modalités. Dans le cas du prétraitement par l'AFC, on a retenu les 6 premiers axes factoriels pour les données PHYTOS (représentant 72,4 % de la variation totale), 4 axes factoriels seulement pour les données BOUCLES (représentant 77,9 % de la variation totale) et 4 axes également pour les données artificielles BLOCS (représentant 43,1 % de la variation totale).

4.3 Résultats des comparaisons.

Les meilleures partitions obtenues avec chaque algorithme ont été comparées sur la base de l'inertie interclasse, calculée sur les données brutes avec la métrique du Khi-carré (Tableau 1). Ces partitions ont été obtenues après 500 tirages aléatoires initiaux pour tous les jeux de données.

	Prétraitement AFC	K-médoïdes	K-modes	K-profils	Partition de référence
PHYTOS	0,4003	0,3951	0,3922	0,3954	0,3857
BOUCLES	0,7189	0,6269	0,7132	0,7198	0,7119
BLOCS	0,3228	0,2170	0,3136	0, 3251	0,3089

Tableau 1. Valeurs des rapports inertie-inter/inertie-totale selon les algorithmes et les jeux de données.

Les qualités des partitions obtenues sont très voisines et, en général, meilleures que les partitions de référence. Le nouvel algorithme arrive au deuxième rang dans le premier cas et au premier rang dans les deux autres cas. Le résultat inattendu est la bonne tenue de la méthode utilisant le prétraitement par l'AFC.

5 Conclusion.

Une adaptation de l'algorithme classique des k-moyennes a été faite pour traiter des données qualitatives. Le nouvel algorithme repose sur la métrique du Khi-carré appliquée aux profils des individus et aux profils de leurs classes. Il converge rapidement vers un optimum local de l'inertie inter-classe, optimum dépendant de la partition initiale. Pour éviter cet inconvénient on réitère un grand nombre de fois des tirages au hasard de la partition initiale. Appliqué à diverses données le nouvel algorithme donne de bons résultats ; comparé à ses concurrents immédiats il obtient des résultats équivalents ou meilleurs que ceux-ci. Toutes les méthodes examinées nécessitent le choix préalable du nombre de classes ce qui est une opération délicate quand les données ne sont pas connues par ailleurs.

6 Bibliographie

[BEN 73] BENZÉCRI J.P. *L'Analyse des données. Tome 2: L'Analyse des Correspondances.* Dunod, Paris, 1973.

[FOR 65] FORGY E.W. "Cluster Analysis of Multivariate Data : Efficiency Versus Interpretability of Classifications", *Biometric Society Meetings*, Riverside, California (Abstract in *Biometrics* Vol. 21, no 3, p 768), 1965.

[GOV 84] GOVAERT G. "Classification simultanée de tableaux binaires". In *Data Analysis and Informatics III,* E. Diday, M. Jambu, L. Lebart, J. Pagès et R. Tomassone (Eds), Norh-Holland, Amsterdam, 1984, p. 223-236.

[HUA 98] HUANG Z. "Extensions to the k-means algorithm for clustering large data sets with categorical values". *Data Mining and Knowledge discovery*, vol. 2, 1998, p. 283-304.

[JAM 78] JAMBU M., LEBEAUX M.O. *Classification automatique pour l'Analyse des données. Tome 1.- Méthodes et Algorithmes,* Dunod, Paris, 1978.

[KAU 90] KAUFMAN L., ROUSSEEUW P.J. *Finding groups in data : an introduction to cluster analysis.* Wiley, 1984.

[LER 80] LEREDDE H., PERIN P. “Les plaques-boucles mérovingiennes”. *Les dossiers de l’archéologie*, no 42, 1980, p 83-87.

[MAC 67] MAC QUEEN J.B. "Some methods for classification and analysis of multivariate observations", *Proc. Symp. Math. Statist. and Probability, 5th,* Berkeley, AD 669871, Univ. of California Press, Berkeley, Vol. 1, 1967, p 281-297.

[RAL 95] RALAMBONDRAINY H. "A conceptual version of the k-means algorithm". *Pattern recognition letters,* vol. 16, 1995, p. 1147-1157.

[ROU 85] ROUX M. *Algorithmes de classification.* Masson, Paris. 1985.

[SAP 93] SAPORTA G. "Notions sur les méthodes factorielles". In *Traitement statistique des enquêtes,* D. Grangé et L. Lebart (Eds), Dunod, 1993, p. 75-89.

Prétraitement des séries temporelles microbiologiques en vue de la classification : Application à la détection des états physiologiques de la levure

N. Sadou[1], **L. Manyri**[1], **S. Régis**[2], **A. Doncescu**[1], **Jean-Pierre Asselin de Beauville**[3]

[1] *LAAS CNRS, 7 Av. du Col. Roche 31077 Toulouse Cedex 04,*
[2]*GRIMAAG Université Antilles-Guyane, Campus de Fouillole 97159 Pointe-à-Pitre*
[3]*Laboratoire d'Informatique de l'Université de Tours, 64 Av. J. Portalis 37200 Tours*

RÉSUMÉ. We present in this paper a model of fed-batch bioreactor states classification which it makes the differences between physiological states and the command action to maintain the nominal parameters. This method is based on the adaptive detection of non correlated bioreactor signals. A segmentation based on maxima of modulus of wavelets transform and Hölder exponent is used to before the principal component analysis PCA. The segmentation allows the detection of operator's action during the fed-batch fermentation and the principal component analysis allows to define the influence of those operator's actions on the physiological states.

MOTS-CLÉS : Transformée en Ondelettes, coefficient d'Hölder, état physiologique, Classification

1. Introduction

Pour produire de la biomasse ou du bioéthanol, les cellules ont besoin de substrat qui leur fournit l'énergie nécessaire. Le substrat doit être une source riche en carbone : le plus utilisé est le glucose. Une fois à l'intérieur de la cellule, le substrat principal (la source carbonée) peut emprunter deux voies métaboliques : la voie fermentaire et la voie oxydative (respiration). La fermentation alcoolique se définit, comme la transformation du glucose en éthanol. Cette voie métabolique suit l'enchaînement des réactions de la glycolyse, le pyruvate étant ensuite transformé en éthanol via l'acétaldéhyde. Cet ensemble de réactions est résumé par l'équation de Gay-Lussac :

$$C_6H_{12}O_6 \longrightarrow 2C_2H_5OH + 2CO_2 + [2ATP] \quad [1]$$

En pratique, les réactions de maintenance, les réactions de synthèse des infrastructures cellulaires et des composés secondaires (glycérol, acide acétique...) limitent le rendement de conversion à 80-90% de sa valeur théorique (0,511 g/g).
L'autre grande voie métabolique de dégradation des sucres, la respiration, comprend la glycolyse, le cycle de Krebs et la chaîne respiratoire. Dans ce cas, le glucose est dégradé en eau et gaz carbonique. Cependant, une part importante des intermédiaires de la réaction contribue à la synthèse de nouvelles cellules qui constituent en fait le produit majeur de la réaction : des rendements en biomasse de 52% sur glucose peuvent être obtenues.
La régulation entre métabolisme fermentaire et oxydatif est théoriquement sous le contrôle de l'oxygène. Mais le glucose, à forte concentration, inhibe la synthèse des cytochromes et contraint les levures à fermenter quelle que soient les conditions d'aération ; c'est la fermentation alcoolique ou l'effet Crabtree.

Les nombreuses études sur les mécanismes d'inhibition en fermentation alcoolique permettent de dégager deux grand types d'inhibition, l'un lié aux conditions initiales d'environement (taux de substrat, taux d'inoculation), l'autre dépendant de l'avancement de la réaction (production d'alcool et de biomasse).

1.1. *Les principaux paramètres biochimiques mesurés*

Comme nous venons de le voir, les paramètres biochimiques ont un rôle essentiel dans la fermentation. On distingue deux types de paramètres :

1. les paramètres non modifiables.
2. les paramètres modifiables ou régulés.

On entend par paramètres non modifiables des paramètres biochimiques observés qu'on ne peut modifier et qui traduisent l'état du système (le bioréracteur) et au moins dans une certaine mesure, l'état physiologique des micro-organismes. Certains paramètres non modifiables ne sont pas mesurés mais calculés à partir des variables mesurées (intégrales ou dérivées...). Les paramètres modifiables ou régulés sont des paramètres dont on connaît plus ou moins l'influence sur le bioréacteur et les levures et qui sont contrôlés directement par l'expert ou par l'ordinateur.

Parmi les paramètres non modifiables on trouve :
- les gaz : l'azote (N2), le dioxygène (O2) et le dioxyde de carbone (CO2). Ils sont mesurés en sortie du réacteur et traduisent l'état physiologiques des levures.
- le Kla ou coefficient de transfert de l'O2. Il traduit le transfert de l'O2 de l'air dans le liquide. Il doit être élevé pour que les micro-organismes vivent.
- le QR ou coefficient respiratoire. C'est le rapport de la vitesse de production de CO2 sur la vitesse de consommation de l'O2. S'il est supérieur à 1 alors les levures sont en état fermentaire, il y a production d'éthanol ; s'il est inférieur ou égal à 1, les levures sont en état oxydatif.
- la pO2 ou pression partielle en O2 dissous dans le liquide du réacteur. Elle donne une meilleur indication que la mesure directe de l'O2 ou le Kla sur l'oxygène présent dans le réacteur.
- Vferm. C'est le volume du fermenteur (i.e. bioréacteur).
- μCO2. C'est le taux de croissance calculé à partir du pourcentage de CO2.
- le carbone accumulé. Il donne une indication sur la biomasse du réacteur et égale à la différence : substrat - CO2.

Parmi les paramètres modifiables on trouve :
- le carbone ajouté. C'est le substrat (glucose). Il influence directement l'état des levures (fermentaire ou oxydatif).
- la masse de base. Elle est régulé automatiquement. Le changement de pente de cette variable peut servir à déterminer l'état physiologique des cellules.
- l'agitation. Elle dépend de la vitesse des palles du réacteur. Elle influence l'homogénéité du milieu et permet de diminuer la taille des bulles de gaz du milieu, permettant une meilleure absorption de ceux-ci par les micro-organismes.
- l'aération. C'est le débit d'air : il influence l'oxygénation du milieu.
- l'antimousse. C'est un détergent qui casse les mousses à l'intérieur du milieu. Il a ainsi une influence sur la pression du bioréacteur et sur le Kla.

Il existe de nombreux autres paramètres (plus d'une trentaine) et tous ces paramètres peuvent être utilisés pour la classification. L'un des problèmes principaux à résoudre est donc de déterminer les paramètres (modifiables ou non) les plus pertinents et les plus significatifs afin d'éviter le maximum de redondance.

2. Traitement des données pour la classification

2.0.1. *Evaluation de l'exposant de Hölder utilisant les maximums de la transformée en ondelettes*

Toutes les méthodes utilisant les maximums de la transformée en ondelettes ont une étape de recherche des maximums à travers les différents échelles. Il s'agit de trouver les maximums correspondant à une même singularité significative et qui se propagent dans toutes les échelles. On cherche donc les courbes connexes des maximums encore appelées lignes de maximums. Concrètement, lorsque l'on passe d'une échelle à l'autre, la recherche du maximum de l'échelle analysée correspondant au maximum de l'échelle précédente se fait à l'intérieur d'un cône

d'influence dépendant directement de la valeur de l'échelle [MAL 92b]. La difficulté majeure provient du fait que la propagation des maximums n'est pas rectiligne d'une échelle à l'autre (l'abscisse du maximum change souvent d'une échelle à l'autre) et cela d'autant plus que l'échelle analysée devient grossière.
Une fois que tous les maximums se propageant à toutes les échelles ont été récupérés il est possible d'évaluer pour chaque singularité correspondant à un maximum se propageant à toutes les échelles, l'exposant de Hölder. Plusieurs approches sont alors possibles. On distingue deux types de méthodes :

1) les méthodes graphiques

Cette approche a été proposée par Mallat et Whang [MAL 92a]. Pour chaque singularité on représente les maximums en ordonnée et l'échelle correspondante en abscisse dans une représentation logarithmique (il s'agit du logarithme de base 2 car on travaille sur des échelles dyadiques). Les maximums sont reliés par une droite et la pente de cette droite correspond à l'exposant de Hölder [MAL 92a]. Pour calculer le coefficient de la droite, plusieurs techniques ont été utilisées :

- Mallat et Hwang utilisent la régression linéaire classique pour calculer la pente de droite [MAL 92a].
- Hwang et Du proposent d'utiliser une autre technique appelé "régression médiane", plus robuste que la régression linéaire classique [DU 00]. La régression médiane a été proposée par Steele et Steiger [STE 86].
- Struzik propose d'utiliser le calcul direct de la pente de la droite en utilisant le maximum de l'échelle minimale et celui de l'échelle maximale [STR 99]. Il impose un seuil arbitraire pour l'échelle maximale dans le but de rendre la méthode plus robuste pour caractériser des signaux dont les singularités sont très proches les unes des autres [STR 98][STR 99]. Puis il remplace la valeur du maximum de l'ondelette correspondant l'échelle maximale par une moyenne de tous les maximums de cette échelle (voir [STR 99]) afin d'éviter une instabilité numérique des résultats. L'objectif de cette méthode, comme nous venons de le voir, est d'évaluer les exposants de Hölder des singularités proches les unes des autres.

2) les méthodes basées sur l'optimisation

Cette approche a été proposée par Mallat et Zhong [MAL 92b] qui proposent de minimiser la fonction suivante [MAL 92b] :

$$\sum_j \left(ln_2(|a_j|) - ln_2(C) - j - \frac{\alpha(x_0) - 1}{2} ln_2(\sigma^2 + 2^{2j}) \right)^2 \quad [2]$$

où a_j représente le maximum à l'echelle j, C est une constante dépendant de la singularité localisée en x_0, σ est l'écart-type d'une gaussienne approximant la singularité (voir [MAL 92b]) et $\alpha(x_0)$ l'exposant de Hölder. L'évaluation porte sur C, σ et bien sûr $\alpha(x_0)$.
Ces deux auteurs proposent d'utiliser la descente du gradient pour minimiser la fonction et évaluer ces paramètres [MAL 92b]. L'utilisation de la descente du gradient est simple dans ce cas, car le calcul des dérivées partielles de la fonction est direct et simple. Si l'on cherche à être précis au niveau des valeurs de l'exposant de Hölder et que le temps de calcul n'est pas l'objectif principal, la minimisation est plus adaptée. Il faut tout de même noter que même avec la méthode de minimisation, on n'est pas sûr de trouver la valeur exacte de l'exposant en raison des défauts de la descente du gradient évoqués ci-dessus.

2.0.2. *Evaluation du coefficient de Hölder par minimisation des algorithmes évolutionnaires*

Pour caractériser les exposants de Hölder des différents signaux biochimiques, la méthode utilisant la minimisation de fonction 2 semble la plus adaptée. En effet, les phénomènes biologiques sont souvent lents, le temps déxécution des méthodes de minimisation est tout à fait raisonnable par rapport à ce temps biologique. Il est plus intéressant de bien différencier les différents exposants de Hölder des singularités des signaux biologiques que de chercher une méthode d'évaluation rapide, car le "temps réel" d'un procédé biotechnologique est très lent et la précision prime sur la vitesse d'éxécution. En effet, la caractérisation précise des coefficients de Hölder est cruciale pour différencier les différents phénomènes survenant lors de l'expérience.
La méthode de minimisation de la fonction définie par l'équation 2 est donc la plus pertinente. Cependant, comme nous venons de le voir, la descente du gradient proposée par Mallat et Zhong [MAL 92b] posent plusieurs problèmes. Le principal défaut de la descente du gradient provient du fait que cette méthode peut stationner

au niveau d'un minimum local et considérer celui-ci comme la solution cherchée. Une méthode de minimisation plus robuste et moins sensible aux minimum locaux permettrait d'obtenir de meilleurs résultats.

La recherche d'une méthode plus robuste nous a conduit aux méthodes évolutionnaires de type algorithmes génétiques (AG).

La démarche suivante consiste à comparer cette nouvelle méthode aux méthodes classiques existantes.

3. Comparaison des méthodes d'évaluation de l'exposant de Hölder

La comparaison des méthodes est une étape indispensable pour évaluer celles-ci. La comparaison que nous nous proposons d'effectuer se fera sur un exemple simple et compréhensible. L'objectif est de comparer les résultats des différentes méthodes. Les méthodes testées sont les suivantes :

- la méthode graphique utilisant la regression linéaire [MAL 92a]
- la méthode graphique utilisant la regression médiane [DU 00]
- la méthode de minimisation utilisant la descente du gradient [MAL 92b]
- la méthode de minimisation utilisant les algorithmes évolutionnaires différentiels [MAN 03]
- la méthode hybride de minimisation utilisant les algorithmes évolutionnaires différentiels puis la descente du gradient

Les résultats montrent qu'aucune méthode ne donne les valeurs théoriques exactes. La méthode la moins bonne est la minimisation utilisant la descente du gradient (en dépit du fait que plusieurs initialisations des paramètres de cette méthodes ont été testées). Les méthodes graphiques fournissent des résultats acceptables pour le Dirac (valeurs négatives "proches" de -1) tandis que ceux-ci sont erronés pour l'exposant de Hölder des discontinuités du palier. Les meilleurs résultats sont obtenus en utilisant les algorithmes génétiques. Par ailleurs la méthode hybride combinant les AG et la méthode du gradient n'apporte aucune amélioration significative.
Des tests supplémentaires ont été effectués sur un autre signal de synthèse. Ces tests ont été réalisés dans le cadre d'une application concernant la détection de fautes. Cette fois, l'ondelette utilisée est une LOG. Les résultats du tableau 1 montrent encore une fois que les AG donnent les meilleurs résultats (dans cet exemple la régression médiane et la méthode hybride n'ont pas été utilisée car ces méthodes donnent des résultats similaires respectivement à la régression médiane et aux AG). Des tests complémentaires montrent que même en changeant le nombre de voix par octave, les AG donnent toujours les mêmes résultats.

Singularités	Dirac	Singularité Palier	Singularité "cups"	Rampe
Coef. de Hölder théorique	-1	0	0,5	1
Coef. de Hölder par la Régression Linéaire (Méthode Graphique)	-1,13	0,16	0,61	0,84
Coef. de Hölder par Descente de Gradient (Optimisation)	-0,24	0,39	0,74	1,20
Coef. de Hölder par les AG (Optimisation)	-1,02	0,02	0,52	1,0007

TAB. 1. *Résultats des méthodes utilisées pour le calcul de l'exposant de Hölder avec 16 voix par octave*

4. Clustering pour les bio-procédés fed-batch

La première utilisation du maximum de la transformée en ondelettes a été la détection des points d'inflexion de certains paramètres biochimique dans un procédé de type fed-batch. La caractéristique principale de bio-procédés

de type fed-batch est la régulation de certains paramètres biochimiques, contrairement aux bio-procédés de type batch où quasiment aucune modification n'est faite. En plus de ces régulations sur certains paramètres, il est également possible en fed-batch de rajouter de nouveaux produits en cours d'expérience. Le clustering que nous proposons pour détecter ces interventions est assez proche de la méthode présentée ci-dessus. En effet, on cherche encore les singularités significatives en utilisant les ondelettes et les AG, mais cette fois-ci, tous les paramètres sont analysés (et pas seulement les paramètres régulés). Les singularités détectées et sélectionnées définissent les bornes d'intervalles temporels. On obtient ainsi une segmentation temporele du bio-procédé en plusieurs intervalles de temps. Nous proposons de caractériser chaque intervalle par les signes des différents coefficients de corrélations linéaires calculés deux à deux entre tous les paramètres. Ainsi chaque intervalle est caractérisé par une série de signes positifs ou négatifs. Les intervalles ayant la même série de signe sont regroupés dans la même classe.

5. Conclusion

Cette approche bien qu'assez simple à la base, semble être prometteuse. En effet, les tests que nous avons effectués nous ont permis de classer dans une seule et même classe *toutes* les interventions de même nature effectuées à des moments différents.

6. Bibliographie

[DU 00] DU C.-L., HWANG W.-L., Singularity Detection and Characterization with Complex-Valued Wavelets and their Applications, rapport n^{o}IIS-00-010, 2000, Institute of Information Science, Academia Sinica.

[MAL 92a] MALLAT S., HWANG W.-L., Singularity Detection and Processing with Wavelets, *IEEE Transaction on Information Theory*, vol. 38, n^o 2, 1992, p. 617-643.

[MAL 92b] MALLAT S., ZHONG S., Characterization of Signals from Multiscale Edges, *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 14, n^o 7, 1992, p. 710-732.

[MAN 03] MANYRI L., REGIS S., DONCESCU A., DESACHY J., URRIBELAREA J., Holder Coefficient Estimation by Differential Evolutionary Algorithms for Saccharomyces Cerivisiae Physiological States Characterisation, *ICPP-HPSECA*, Kaohsiung, Taiwan, Octobre 2003.

[STE 86] STEELE J., STEIGER W., Algorithms and complexity for least median of squares regression, *Discrete Applied Mathematics*, , 1986.

[STR 98] STRUZIK Z., *Fractals and beyonds- Complexities in the Sciences*, Chapitre Removing Divergences in the Negative Moments of the Multi-Fractal Partition Function with the Wavelet Transformation, p. 351-352, World Scientific, 1998.

[STR 99] STRUZIK Z. R., *Fractals : Theory and Application in Engineering*, Chapitre Local Effective Hölder Exponent Estimation on the Wavelet Transform Maxima Tree, p. 93-112, Springer Verlag, June 1999.

Analyse des données incomplètes avec l'application aux expériences biopuces

Basavanneppa Tallur

IRISA-IFSIC,
Université de Rennes l,
Avenue Général Leclerc, Campus de Beaulieu,
35042 RENNES CEDEX, France
mél :tallur@irisa.fr

RÉSUMÉ. La technologie des bio-puces permet aux biologistes d'effectuer des expériences sur plusieurs milliers de gènes simultanément dans des conditions variées. Mais souvent il manque certaines mesures à cause des limitations matérielles, et il arrive que un certain nombre de mesures, n'étant pas de qualité suffisante, sont considérées comme manquantes. Dans ces conditions la solution consiste souvent à remplacer de telles données par des valeurs estimées ou même, dans certains cas, à supprimer les gènes ou les échantillons (« arrays ») comportant des données manquantes. D'autre part, les travaux récents dans le domaine de la bioinformatique ([ALT 00], [LEE 03], [MAO 03], [GIR 04], ...) montre que le modèle factoriel permet une interprétation en terme des processus biologiques. Nous proposons ici, une méthode d'estimation itérative des données manquantes, optimale en vue d'une analyse factorielle, qui peut aussi être utilisée préalablement à une classification hiérarchique, notamment par la méthode A.B.C ([TAL 88]), qui utilise une représentation des données par le nuage des profils à la manière de l' AFC (Analyse Factorielle des Correspondances).

MOTS-CLÉS : biopuces, expression génomique, données manquantes, analyse factorielle, classification

1 Introduction

L'énorme progrès réalisé dans les domaines de la technologie bio-puces à ADN (DNA microarray) et le séquençage de génome permet aujourd'hui aux biologistes de mesurer le niveau d'expression à l'échelle d'un génome entier. L'analyse de ces données est précieuse pour la connaissance fondamentale de la vie au niveau moléculaire depuis la régulation de l'expression des gènes et leurs fonctions aux mécanismes cellulaires. Parmi les outils statistiques utilisés pour l'analyse de telles données, on peut citer la classification (cluster analysis) rendue « populaire » par les travaux d'Eisen ([EIS 98]). On peut citer parmi de nombreux articles de synthèses sur l'application de l'analyse classificatoire aux données de l'expression, celui de W. Shannon [SHA 03] qui contient une riche bibliographie. Nous avons souligné divers problèmes que l'on est amené à résoudre avant d'utiliser de telles méthodes ([TAL 03]). Il s'agit du prétraitement des données issues des expériences bio-puces consistant notamment en la « normalisation », la « standardisation » et le traitement des données manquantes. Des travaux récents montrent l'intérêt des méthodes factorielles telles que l'ACP (Analyse en Composantes Principales) ([ALT 00]), l'ACP probabiliste (Probabilistic PCA) ([TIP 97]) et une version généralisée de l'ACP - Analyse en composantes indépendantes (Independent Component Analysis ou ICA) - ([ROB 03]). Les « composantes » ou facteurs identifieraient les processus biologiques qui contribuent aux profils

d'expression observés, en rendant les facteurs biologiquement interprétables. Mark Girolami propose une approche variante de l'ICA ([GIR 04]) en considérant que chaque profil d'expression est une combinaison linéaire de plusieurs profils d'expression « prototypes » ou des processus physiologiques. Ces derniers travaux justifient suffisamment l'emploi des méthodes factorielles telles que l'ACP et l'AFC (après transformation des données) conjointement avec la classification. Les expériences bio-puces comportent plus ou moins de données manquantes pour diverses raisons : résolution de l'image insuffisante, image polluée, ou simplement à cause de la poussière ou égratignure des plaques. Quelle que soit la méthode d'analyse employée, les données manquantes posent un problème sérieux car toutes ces méthodes nécessitent les données complètes. Il existe des solutions plus ou moins raffinées à ce problème telles que par exemple : remplacer les données manquantes par des zéros, par les moyennes (ou la médiane) de la ligne (ou de la colonne), ou par la moyennes des k plus proches voisins, etc.. Dans cet article nous abordons le problème des données manquantes et proposons une méthode itérative d'estimation (ou d'imputation) afin de pouvoir obtenir des facteurs ou une classification hiérarchique aussi fidèlement que possible.

2 Représentation des données et notations

Les données de l'expression se présentent sous forme d'un tableau croisant les n gènes (lignes) et p échantillons (ou conditions expérimentales appelés « arrays »). Dans l'ACP normée, les données sont centrées et réduites et chaque gène est représenté par un vecteur de l'espace de dimension p muni de la distance euclidienne usuelle. Analyse des correspondances, bien que développée dans le cadre d'un tableau de contingence, est tout à fait applicable au tableau des valeurs positives, surtout lorsqu'elles sont toutes exprimées dans la même unité. On considère alors, le nuage des profils lignes et celui des profils colonnes, munis d'une distance du Chi-deux.

Soit x_{ij} le niveau d'expression du gène i dans l'échantillon j ($i=1, \ldots, n$; $j=1, \ldots, p$). Avec les notations habituelles, on notera les sommes marginales $x_{i.}$ et $x_{.j}$

La formule de reconstitution des données, bien connue, à l'ordre s (c.à.d. avec les s premiers facteurs) est

$$x_{ij} = \frac{x_{i.}x_{.j}}{\sum_{i,j} x_{ij}} \left[1 + \sum_{\alpha=1}^{s} \lambda_\alpha^{-1/2} F_\alpha(i) G_\alpha(j) \right]$$

où λ_α sont des valeurs propres, F_α et G_α sont des facteurs lignes et facteurs colonnes, respectivement.

3 Méthode d'estimation itérative

Nous avons proposé une méthode itérative d'estimation des données manquantes en vue d'une AFC basée sur la formule de reconstitution des données à partir des facteurs ([TAL 73]). En fait, le problème est apparu lors d'une expérience en biologie sur les récepteurs olfactifs. Il était impossible, pour des raisons matérielles et techniques de mesurer toutes les réponses en chacun des points du récepteur à tous les stimulii. Nous avons donc simulé les « trous » dans un tableau des mesures connues de deux façons différentes :

- Les trous sont répartis de façon aléatoire
- Les trous sont planifiés suivant un plan d'expérience

En faisant varier le taux de données « manquantes », on a reconstitué les facteurs après avoir estimé les données par la méthode proposée. Dans un tableau de dimension 121 X 20, on a pu pratiquement reconstituer les 5 premiers facteurs (avec une corrélation de plus de 0.90) jusqu'à 20% des données manquantes aléatoirement et jusqu'à 25% des données manquantes planifiées.

L'algorithme peut se résumer comme ci-dessous :

1. Initialisation: on remplace les valeurs inconnues par des zéros et on fixe la valeur maximum s_{max} du paramètre s, l'ordre d'estimation

2. On pose *s=0 ;* on fait une estimation des données manquantes à l'ordre 0 (qui consiste à appliquer la formule de reconstitution ci-dessus avec *s=0*, càd, sans aucun facteur); les valeurs inconnues sont remplacées par leurs estimations
3. On répète (2) jusqu'à convergence
4. on incrémente s $\rightarrow$ s+1
5. On calcule les s premiers axes de l'AFC du tableau précédent (calcul de λ_α, F_α et G_α, pour α=1, .., *s*)
6. On remplace les données manquantes par leurs estimations à l'ordre s
7. On répète (4), (5) et (6) jusqu'à convergence
8. Si $s = s_{max}$ on arrête, sinon on retourne à l'étape 4.

On constate expérimentalement que la convergence pour chaque valeur de *s* est très rapide. La qualité de l'estimation est évaluée en comparant les facteurs estimés avec les facteurs « réels » (obtenus avec les données complètes). Cette méthode est tout à fait applicable aux données de l'expression et donne des résultats comparables à ceux obtenus sur des données des récepteurs olfactifs.

4 Validation des résultats

En vue d'évaluer la performance de la méthode d'estimation proposée en fonction du nombre de données manquantes, nous avons réalisé plusieurs expériences en supprimant, dans un tableau des données complètes, un certain pourcentage d'observations au hasard. Le taux des valeurs manquantes a été progressivement augmenté jusqu'à 20% des données. Les facteurs issus du tableau complet ont été comparés à ceux obtenus à partir des données estimées à l'aide des coefficients de corrélation linéaire entre les facteurs de même rang. Dans tous les cas testés, le coefficient de corrélation reste très forte pour les 5 premiers facteurs (par exemple, nous avons trouvé une corrélation supérieure à 0,95 pour un tableau comportant une centaines de lignes et une vingtaine de colonnes).

5 Conclusion et perspectives

La méthode pour estimer les données manquantes, que nous avons proposée et expérimentée sur des données d'une expérience en biologie est simple et efficace. Elle est tout à fait utilisable non seulement en vue de l'analyse factorielle (ACP ou AFC), mais aussi pour la classification hiérarchique par l'Agrégation Basée sur la Corrélation (ABC, [TAL 88]). Il reste à l'évaluer massivement sur des données plus importantes. Nous avons également implémenté une méthode d'estimation adaptée aux données standardisée (par exemple, celles de l'expression génomique), qui est basée sur la formule de reconstitution des données à partir des axes principaux d'inertie et de composantes principales de l'ACP normée.

6 Bibliographie

[ALT 00] ALTER O., BROWN P., BOTSTEIN D., "Singular value decomposition for genome-wide expression data processing and modeling", *PNAS*, vol. 97, n° 18, 2000, p. 10101-10106

[GIR 04] GIROLAMI M., BREITLING R., « Biologically valid linear factor models of gene expression », *Bioinformatics,* vol. 20, n° 17, 2004, p. 3021-3033.

[EIS 98] EISEN M., SPELLMAN P., BROWN PO. *et al*., « Cluster analysis and display of genome-wide expression patterns », *PNAS,* vol. 95, 1998, p. 14863-14868.

[LEE 03] LEE S.I., BATZOGLOU S., « Application of independent component analysis to microarrays», *Genome Biol.*, vol. 4, R76.

[MAO 03] MAO R., ZIELKE CL., ZIELKE HR., PEVNSER J., "Global upregulation of chromosome 21 gene expression in the developing Down syndrome brain", *Genomics*, vol. 81, p. 457-467.

[ROB 01] ROBERTS S., EVERSON R. (eds), *«Independent component analysis Principles and practice »*, 2001, Cambridge university press, Cambridge.

[SHA 03] SHANNON W., CULVERHOUSE R., DUNCAN J., « Analyzing microarray data using cluster analysis », *Pharmacogenomics,* vol. 4, n° 1, 2003, p. 41-51.

[TAL 03] TALLUR B., « Analyse des données de l'expression génomique par la classification : pourquoi et comment ?», *Méthodes et perspectives en classification*, 2003, Dodge Y., Melfi G. (eds), Press académique de Neuchâtel.

[TAL 73] TALLUR B., « Analyse des correspondances en cas de données manquantes: application en biologie», Thèse Doctorat de 3ème cycle, Université de Paris 6, 1973.

[TAL 88] TALLUR B., « Contribution à l'analyse exploratoire de tableaux de contingence par la classification », Thèse, Doctorat ès sciences Mathématique, Université de Rennes1, 1988.

[TRO 01] TROYANSKAYA O., CANTOR M., SHERLOCK G., BROWN P., HASTIE T., TIBSHIRANI R., BOTSTEIN D., ALTMAN R., «Missing value estimation methods for DNA microarrays», *Bioinformatics,* vol. 17, n° 6, 2001, p. 520-525.

Etude expérimentale du coût subjectif en théorie bayésienne de la décision

Gilles Verley, Jimmy Edouard, Jean-Pierre Asselin de Beauville

Laboratoire d'informatique,
Université François Rabelais deToursl,
64 avenue Jean Portalis
37200 Tours, France

RÉSUMÉ. Dans le cadre de la théorie bayésienne de la décision dans l'incertain, on étudie comment le coût subjectif intervient concrètement dans une expérience cognitive simple de reconnaissance de formes avec apprentissage supervisé. Si on admet que le coût subjectif est la perception cognitive de l'information indicative de coût fournie en consigne aux sujets lors de l'expérience et que nous appelons ici coût indicatif, alors les résultats expérimentaux sont en accord avec le modèle théorique en ce sens qu'une variation du coût indicatif induit bien une variation du coût subjectif qui modifie significativement la décision de classement. Le rapport du coût subjectif au coût indicatif caractérise la sensibilité du groupe testé à la consigne donnée au cours de l'expérience.

MOTS-CLÉS : coût subjectif, apprentissage, règle de bayes, théorie de la décision,

1 Introduction

Il y a longtemps que les probabilités subjectives ont fait l'objet de nombreux travaux en psychologie [BRE 65] et que l'on sait que la théorie bayésienne de la décision peut être interprétée de manière subjectiviste comme une modélisation de la décision humaine dans l'incertain [MAT 67]. Si, depuis cette période lointaine, de nombreux autres modèles de la décision ont vu le jour [DEN 99] (logique floue, théorie des possibilités, méthodes à base de cas, d'agents, approches biomimétiques, théorie de Dempster et Shafer, etc.), l'approche bayésienne se situe parmi celles qui prennent le plus naturellement en compte le concept de coût subjectif associé à une décision dans l'incertain. Les expérimentations sur l'homme étant rares dans ce domaine, il nous a semblé pertinent de tenter une expérience simple de reconnaissance de formes avec apprentissage supervisé pour observer le coût subjectif prévu par la théorie bayésienne. Pour cela, après avoir rappelé ce qu'est un classifieur bayésien, nous présentons une expérience par la pensée qui tend à montrer le rôle du coût subjectif dans une décision humaine de reconnaissance de formes. Nous décrivons ensuite un dispositif expérimental issu de la psychologie cognitive permettant d'observer le coût subjectif dans une situation contrôlée simple. Il s'agit de faire apprendre des classes de sons à des sujets puis de leur faire reconnaître ces sons en contrôlant les coûts indicatifs associés aux classes. Dans une dernière partie, nous présentons et discutons les résultats expérimentaux.

2 Le classifieur bayésien

Le classifieur de Bayes est basé sur la connaissance totale des lois de probabilité gouvernant les échantillons. Le rôle du classifieur bayésien est d'affecter une forme anonyme $\underline{x}$ à une des classes données. Les formes sont modélisées par un vecteur aléatoire $\underline{X}$ de dimension d. On soumet les formes observées $\underline{x}$ au classifieur qui, suivant une règle de décision va affecter la forme à une classe. Le principe général de cette règle consiste à choisir la décision $\hat{\omega}(\underline{x})$ dont le coût moyen de décision (ou risque de Bayes) sera le plus faible :

$$\hat{\omega}(\underline{x}) = \omega_i \text{ si } l_i(\underline{x}) \le l_j(\underline{x}) \ \forall j = 1,2..n_c$$

avec $l_i(\underline{x}) = \sum_{j=1}^{n_c} \lambda(\omega_i / \omega_j) P(\omega_j / \underline{x})$

avec $P(\omega_j / \underline{x}) = \dfrac{f(\underline{x} / \omega_j) P(\omega_j)}{f(\underline{x})}$

et $f(\underline{x}) = \sum_{j=1}^{c} f(\underline{x} / \omega_j) P(\omega_j)$

avec :

n_c : Nombre de classes de l'échantillonω_i : $i^{ème}$ classe i = 1, 2,...., n_c

$\underline{x}$: Observation d'un individu (réalisation du vecteur aléatoire $\underline{X}$)

$P(\omega_i / \underline{x})$: Probabilité à posteriori de la classe ω_i sachant $\underline{x}$

$P(\omega_i)$: Probabilité à priori de la classe ω_i

$f(\underline{x} / \omega_i)$: Densité de probabilité conditionnelle de $\underline{x}$ sachant ω_i

$\lambda(\omega_i / \omega_j)$ le coût du classement dans la classe i d'un élément de la classe j

3 Une expérience par la pensée de l'influence d'un coût indicatif sur un coût subjectif bayésien

Dans l'approche subjectiviste de la théorie bayésienne, la probabilité devient une propriété d'un individu et non plus d'un événement comme dans l'approche fréquentiste. On en déduit les conséquences suivantes [SAV 54]:

- il existe une probabilité numérique, associée aux états de la nature et satisfaisant les règles du calcul des probabilités.
- on peut associer à chaque conséquence une valeur numérique qui représente son "utilité" pour l'individu.
- enfin, l'ordre de préférence sur les actes est celui de leurs espérances mathématiques, calculées à partir des probabilités et des utilités qu'on vient d'introduire. C'est le principe de l'inférence statistique.

Illustrons ces principes par une expérience par la pensée [VER 94]. Soit un barrage de police tenu par un policier qui a pour mission d'arrêter un malfaiteur. Sa connaissance de ce malfaiteur se limite à quelques photos (apprentissage supervisé de la forme à reconnaître à partir d'exemples). Pour certaines personnes parmi l'ensemble de celles qui vont passer le barrage, ce policier va décider d'effectuer une vérification approfondie par des moyens appropriés. Il aura donc reconnu ces personnes comme étant a priori suspectes, c'est-à-dire que ces personnes ont une probabilité subjective plus importante d'être la personne recherchée que d'autres qui vont être rejetées. On a donc un problème de reconnaissance à une classe avec une classe de rejet. Si maintenant, on informe le policier que le ministre de l'Intérieur est très attaché à ce que ce malfaiteur soit arrêté et qu'une prime substantielle sera attribuée à celui qui l'aura arrêté, il est raisonnable de penser que le groupe de personnes classées par le policier comme suspectes sera plus important. En somme, en plus des personnes classées comme suspectes dans la première expérience, vont s'ajouter de nouvelles personnes qui ne vont devenir suspectes qu'au regard de la prime promise. En termes bayésiens, cela revient à dire que le coût subjectif de classement d'un suspect dans la classe de rejet a augmenté. Néanmoins, quelque soit le désir de toucher la prime, tout le monde ne pourra pas être classé comme suspect compte tenu des capacités limitées de vérification et d'encadrement des personnes contrôlées sans parler de la mauvaise humeur du public… Inversement, si on avertit maintenant le policier qu'une haute personnalité en vacances doit aussi passer incognito cette route et qu'il serait très dommageable qu'elle soit inquiétée, on peut s'attendre à ce que la taille de l'ensemble des suspects se rétrécisse substantiellement. C'est maintenant le coût subjectif de classement dans la classe des suspects qui a augmenté.

Cette expérience mentale nous permet de saisir l'influence de la seule information de coût indicatif sur le coût subjectif de la décision d'affectation d'un élément de l'échantillon dans une classe ou dans une autre et ceci en termes bayésiens. Ici, le coût indicatif de classement est fourni soit quantitativement

par la valeur de la prime en cas de succès dans la première expérience, soit qualitativement par les remontrances dans la seconde.

4 Une expérience concrète de reconnaissance de sons appris dans un contexte de coûts controlés

Pour observer l'existence du coût subjectif prévu par le modèle théorique, on imagine une expérience cognitive très simple [EDO 97] conforme aux protocoles expérimentaux recommandés en psychologie cognitive [ROS 96]. Il s'agit dans une première phase de faire apprendre à un groupe d'une trentaine de sujets deux classes de sons générés selon deux lois distinctes de telle manière qu'il existe une certaine ambiguïté entre les deux classes. A chaque classe est associée une forme visuelle géométrique, un cercle ou un carré, qui représente symboliquement la loi et donc la classe de sons qui en est issue. Dans une deuxième phase, on demande aux sujets (c'est la consigne) de reconnaître la classe d'appartenance de sons qui leur sont présentés selon différentes modalités de coûts indicatifs associés aux classes. On peut alors mesurer les proportions empiriques de formes bien classées.
Les résultats expérimentaux sont présentés sous la forme de courbes de probabilités a posteriori des classes sachant la valeur de la variable aléatoire (la fréquence sonore).ou/et de frontières de décision On peut alors analyser ces courbes selon les modalités expérimentales.

5 Résultats expérimentaux

5.1 Apprentissage des probabilités subjectives sans coût associé

Dans la situation de contrôle où le coût n'intervient pas (aucune consigne de coût indicatif), nous observons sur la figure 1 ci-dessous la bonne adéquation entre la courbe théorique des probabilités a posteriori et la courbe expérimentale de la proportion empirique de bien classés de la classe 1 en fonction de la variable aléatoire. Il en est de même symétriquement pour la classe 2 que l'on n'a pas représenté. Cette adéquation montre que les probabilités subjectives apprises lors de l'apprentissage sont bien corrélées avec les probabilités théoriques des classes, ce que confirme la mesure de la corrélation linéaire entre les deux séries (probabilité théorique et proportion empirique) qui est de 0,97. On observe également sur la figure 2 que les frontières théorique et empirique de décision sont presque identiques, ce qui est conforme au modèle.

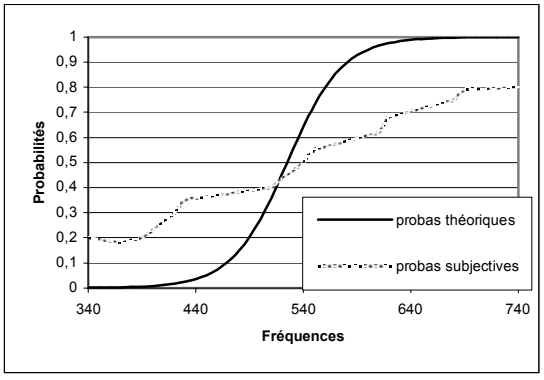


fig. 1 Courbes de probabilités de la classe 1

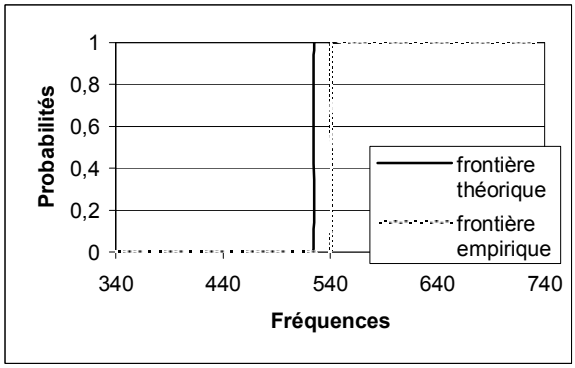


fig. 2 Frontières de décision d'appartenance à la classe 1

5.2 Coûts subjectifs

Dans la situation où on mesure les proportions expérimentales selon deux coûts indicatifs bien distincts, on observe (fig. 3) un déplacement de la courbe des proportions de bien classés dans la classe 1. Sur la même figure, on a représenté le déplacement de la frontière empirique de décision. Les courbes théoriques correspondantes ont été représentées sur la figure 4. Elles ont été calculées en considérant que le coût subjectif est exactement égal au coût indicatif fourni dans la consigne, ce qui représente une situation idéale. On observe que le déplacement des courbes expérimentales (proportion des bien classés et frontière de décision) est orienté dans le même sens que le déplacement des courbes théoriques correspondantes.

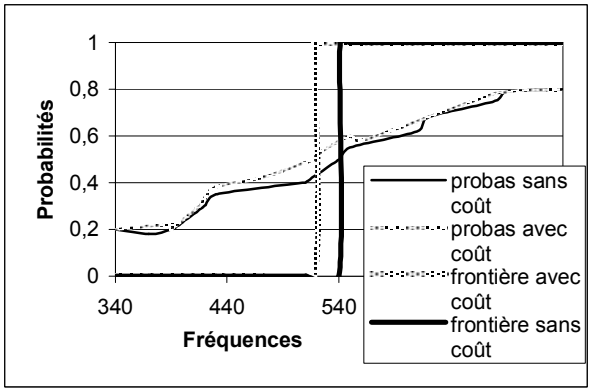


fig. 3 Courbes des proportions empiriques de bien classés dans la classe 1 selon deux situations de coût et frontières de décision correspondantes

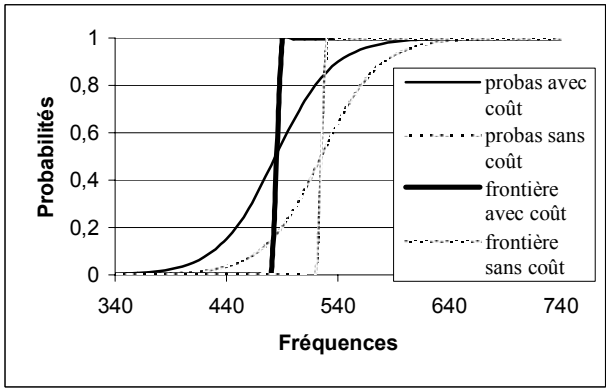


fig. 4 Courbes théoriques

Le rapport entre le déplacement empirique (fig. 3) et théorique (fig. 4) est, en réalité, le rapport du coût subjectif moyen induit chez le groupe de sujets sur le coût indicatif fourni dans la consigne. Ce taux est ici important (supérieur à 50%). Il fournit une estimation grossière de l'incidence quantitative du coût indicatif sur le coût subjectif dans cette situation expérimentale. L'étude et l'affinement de cette mesure nécessiterait un plan d'expériences beaucoup plus lourd en termes de nombres de sujets, de modalités testées, etc., que celui que nous avons réalisé.

6 Conclusion

Le travail présenté ici constitue une approche empirique de l'étude du coût subjectif dans le processus de décision de l'homme dans l'incertain. Il contribue à montrer que le modèle bayésien constitue encore un paradigme pertinent pour ce type d'études même si l'expérience présentée ici reste limitée. Il apporte également un cadre méthodologique pour valider des modèles de la décision humaine dans l'incertain.

7 Bibliographie

[BRE 65] BRESSON F., Les décisions, *Traité de psychologie expérimentale,* vol. 8, 1965, Paris.

[DEN 99] DENUX T., *Modélisation de l'imprécis et de l'incertain en apprentissage supervisé,* IFA99, Berlin, 1999.

[EDO 97] EDOUARD 97, *L'apprentissage de formes chez l'homme*, Rapport interne du L.I. de Tours, Tours, 1997.

[MAT 67] MATALON B., Epistémologie des probabilités, *Logique et connaissance scientifique,* p. 526-552, 1967, Gallimard, Paris.

[ROS 89] ROSSI J.P et al, *La méthode expérimentale en psychologie*, Paris, Dunod, 1989.

[SAV 54] SAVAGE L.J., *The Foundations of Statistics*, New-York, 1954.

[VER 04] VERLEY G., Contribution à la validation des réseaux connexionnistes en reconnaissance des formes, *Thèse de doctorat de l'Université de Tours,* p.19-20, 1994.

Compétition de colonies de fourmis pour l'apprentissage supervisée : CompetAnts

Gilles Verley, Nicolas Monmarché

Laboratoire d'Informatique de l'Université François Rabelais,
64 avenue Jean Portalis
37200 Tours, France

RÉSUMÉ. Nous présentons une modélisation originale de l'utilisation des phéromones par les fourmis afin de résoudre des problèmes d'apprentissage supervisé. Ce cadre biomimétique permet d'illustrer simplement le principe général suivant : plus les fourmis sont attirées par une zone de l'espace de recherche, plus elles y consommeront de la nourriture et plus elles déposeront des phéromones qui augmenteront ainsi la fréquentation de la zone. L'évaporation et la diffusion des phéromones permettent d'étendre l'influence des découvertes des fourmis. Les résultats obtenus sur un jeu de test relativement simple sont encourageants et promettent des études futures.

MOTS-CLÉS : apprentissage supervisé, fourmis artificielles

1 Introduction

Le problème de la classification non supervisée a été abordé à de nombreuses reprises en s'inspirant du paradigme des fourmis artificielles. Par exemple en donnant aux fourmis les données à classer puis en leur laissant le loisir d'agréger ces données pour former des classes [LUM 94, MON 99, ABR 03]] comme elles le font pour leur couvain dans la nature, en associant à chaque fourmi une donnée puis en simulant les échanges d'odeur leur permettant de se construire une identité coloniale [LAB 02] ou en simulant leur agrégation sous forme de grappes [AZZ 03]. D'autres travaux se sont inspiré des capacités de déplacement collectif plus généralement rencontrées dans le monde animal [AUP 03] comme le vol des oiseaux ou le déplacement de bancs de poissons. Dans [PAR 02], on peut trouver une approche basée sur l'utilisation de phéromones pour l'apprentissage supervisé de règles de classification. Cependant, si la capacité des fourmis à construire des chemins de phéromones pour exploiter des sources de nourriture a été largement utilisé en optimisation combinatoire [BON 99], cette voie a été moins explorée pour ce qui est de la classification automatique en particulier. De plus, dans la plupart des cas, la méthode proposée était non supervisée.

Dans cet article nous montrons comment adapter simplement le principe des phéromones déposées par les fourmis pour résoudre un problème d'apprentissage supervisée, c'est-à-dire pour lequel un étiquetage des données est connu.

2 Description algorithmique

2.1 Notations et définition du problème

On considère un ensemble de n données (ou individus) décrits par m paramètres. Pour chaque individu i, on connaît sa classe d'appartenance, notée c_i ($i \in \{1,...,k\}$). On cherche à découvrir pour chaque classe une fonction de densité de probabilité d'appartenance à cette classe : $f_i(x)$ $(i=1...k)$ où x est un point de l'espace des paramètres.

2.2 Retour sur les fourmis artificielles

Les fourmis utilisent les phéromones (mélange d'hydrocarbures) pour construire des chemins entre leur nid et une source de nourriture. Ces phéromones servent alors à mémoriser le chemin mais également à recruter d'autres ouvrières pour exploiter la source de nourriture. Ces substances chimiques s'évaporent avec le temps et le chemin disparaît s'il n'est pas renforcé par le passage des fourmis et le dépôt de nouvelles phéromones. Les phéromones déposées sur le sol constituent donc une forme de mémoire collective et permet aux fourmis d'apprendre collectivement, et sans communication directe, tout en ayant la possibilité d'oublier des chemins qui auraient perdu de leur intérêt.

Dans notre approche, les fourmis se déplacent dans l'espace des paramètres à la recherche de nourriture représentée par les données que l'on considère. A chaque fois qu'une fourmi découvre de la nourriture, elle dépose une certaine quantité de phéromones. Quand une fourmi recherche de la nourriture, elle est à la fois attirée par l'odeur de la nourriture ainsi que par les phéromones présentes dans son voisinage. La nourriture trouvée est consommée par la fourmi qui poursuit ensuite sa quête. La quantité de nourriture totale disponible étant fixée au départ en fonction du nombre d'objets, on peut considérer que les fourmis sont en compétition pour se nourrir : les plus aptes à détecter les phéromones seront les mieux nourries. Une fois toute la nourriture consommée, les traces de phéromones constituent un marquage dans l'espace des paramètres qui peut être interprété comme une densité de probabilité de présence de nourriture et nous donner ainsi une estimation de la fonction f.

Similairement à ce que l'on peut observer en milieu naturel, plusieurs espèces de fourmis peuvent cohabiter, avoir des régimes alimentaires différents et utiliser des « bouquets » de phéromones différents. De la même façon, nous utilisons une espèce différente pour chaque classe d'objets. Une espèce i construit par conséquent une densité f_i en interagissant indirectement avec les autres espèces : bien qu'une espèce soit attachée à une classe, on peut autoriser une certaine plasticité dans l'attrait des fourmis pour un type de nourriture afin de prendre en compte un bruit possible dans l'étiquetage des objets.

Contrairement à un dépôt déterministe de phéromones sur les positions des objets dans l'espace des paramètres, l'utilisation des fourmis et leur attirance pour la nourriture et les phéromones introduit une intensification de la recherche sur les zones denses en objets. De plus, l'évaporation lente des phéromones nous permet de donner une importance décroissante avec le temps aux objets isolés.

2.3 L'algorithme CompetAnt

Le principe général est résumé par la figure suivante :

1. Initialiser la nourriture aux positions des individus de l'échantillon d'apprentissage
2. Tant qu'il reste de la nourriture, faire :
    a. Pour chaque fourmi faire :
        i. Absorber une partie de la nourriture et calculer la quantité de phéromone déposée par la fourmi
    b. Pour chaque position faire :
        i. Calculer l'évaporation des phéromones
        ii. Calculer la diffusion des phéromones sur les positions adjacentes

Figure 1 : algorithme général

La quantité $\tau_{i,t}$ de phéromones déposée par une fourmi à la position i au temps t, est donnée par :

$$\tau_{i,t} = \tau_{i,t-1}(1-\rho) + \delta_{i,t} \times \omega_{i,t} + \theta(\Delta_{i,t} - \Delta'_{i,t}) .$$

Où :

- ρ est un paramètre d'évaporation ;

- $\delta_{i,t}$ permet de prendre en compte la présence d'une fourmi et est obtenue par :

$$\delta_{i,t} = \begin{cases} 1 \text{ si } q < 1/\alpha + \beta\tau_{\text{i,t}} \\ 0 \text{ sinon} \end{cases},$$

avec q, un nombre aléatoire uniformément généré dans [0,1], α est un paramètre d'appétit (appétence) et β est un paramètre permettant de contrôler l'influence des phéromones sur l'attirance des fourmis ;

- $\omega_{i,t}$ représente la quantité de phéromone déposée par la fourmi.
- $\Delta_{i,t}$ (resp. $\Delta'_{i,t}$) représente la diffusion des phéromones provenant du (resp. vers le) voisinage de *i*.
- θ représente un paramètre de volatilité de la phéromone qui caractérise sa capacité de diffusion.

3 Expérimentations

Nous considérons dans un premier temps le cas particulier où l'espace des paramètres ne comporte qu'une seule dimension. Dans ce cas, nous pouvons expliciter la composante de diffusion des phéromones en limitant le voisinage à deux positions de chaque coté de *i* :

$$\Delta_{i,t} = 0.1 \times (\tau_{i-2,t-1} + \tau_{i+2,t-1}) + 0.4 \times (\tau_{i-1,t-1} + \tau_{i+1,t-1})$$

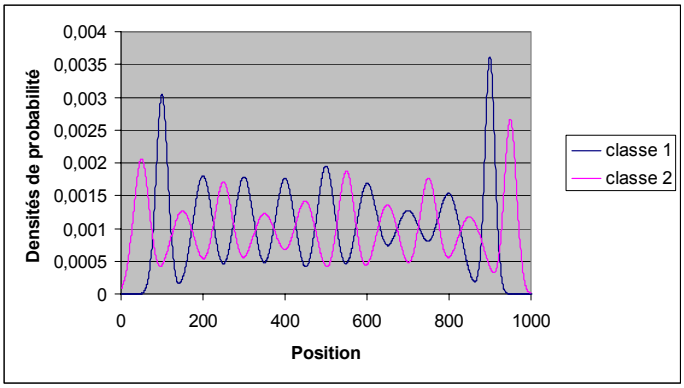


Figure 2 : jeu de test.

Dans une première phase, nous nous sommes concentrés sur les paramètres d'appétence et de diffusion tout en simplifiant le modèle en fixant les paramètres ρ et β à 0 et $\omega_{i,t}$ à 1.

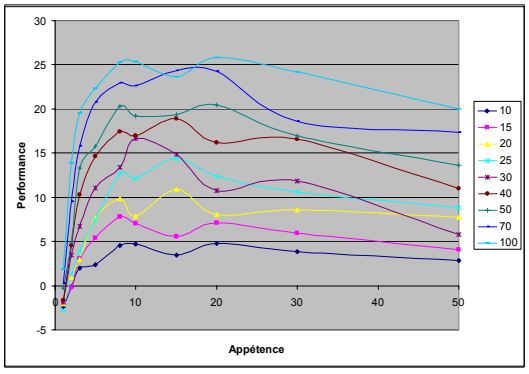


Figure 3 : performances obtenues en fonction de l'appétence (α)

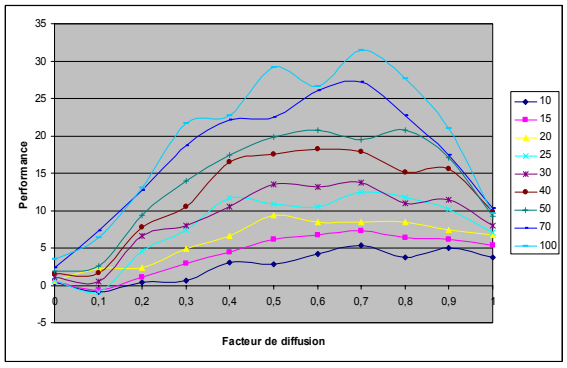


Figure 4 : performances obtenues en fonction de la diffusion (θ)

Sur la figure 3, la performance (taux de bien classés moins taux de mal classés) dépend de l'appétence d'une manière que l'on peut expliquer ainsi ; si l'appétence est trop grande, alors la nourriture est mangée trop rapidement et la phéromone déposée par les fourmis n'a pas le temps de se répandre et, inversement si l'appétence est trop faible.

Sur la figure 4, la performance dépend de la capacité de la phéromone à se diffuser spontanément dans son voisinage (volatilité). Cela peut s'expliquer d'une manière relativement analogue à l'appétence. Dans les deux figures, l'influence des facteurs étudiés est d'autant plus importante que l'échantillon est grand, la taille de l'échantillon améliorant les performances dans tous les cas comme on pouvait l'espérer. Dans d'autres expériences que nous n'avons pas la place de présenter graphiquement ici, nous avons fait varier la complexité du problème. On a pu remarquer des courbes semblables avec un déplacement des optima corrélé à la complexité.

4 Conclusion

Nous avons présenté dans cet article une première approche de l'utilisation du paradigme des fourmis pour l'apprentissage supervisé. On constate des résultats préliminaires encourageants. Toutefois, il reste à montrer l'intérêt pratique de cette approche par rapport à d'autres modèles qui comportent de réelles analogies avec le modèle des fourmis et qui sont mieux étayés sur le plan mathématique quant à leur convergence théorique [VER 94].

5 Bibliographie

[ABR 03] A. ABRAHAM and V. RAMOS. Web usage mining using artificial ant colony clustering. In Proceedings of IEEE Congress on Evolutionary Computation (CEC 2003), pages 1384-1391, Canberra, Australia, december 9-12 2003. IEEE Press.

[AUP 03] S. AUPETIT, N. MONMARCHÉ, M. SLIMANE, C. GUINOT, and G. VENTURINI. Clustering and Dynamic Data Visualization with Artificial Flying Insect. In Erick Cantu-Paz, editor, Genetic and Evolutionary Computation Conference, volume 2723 of Lecture Notes in Computer Science, pages 140-141, Chicago, july 12-16 2003. Springer-Verlag Telos.

[AZZ 03] H. AZZAG, N. MONMARCHE, M. SLIMANE, G. VENTURINI, et C. GUINOT. Classification arborescente de données par auto-assemblage de fourmis artificielles. In Y. Dodge and G. Melfi (éditeurs), Société Francophone de Classification (SFC), pages 55-58, Neuchatel, Suisse, 9-12 Septembre 2003. Presses académiques neuchatel.

[BON 99] E. BONABEAU, M. DORIGO, and G. THERAULAZ. Swarm Intelligence: From Natural to Artificial Systems. Oxford University Press, 1999.

[LAB 02] N. LABROCHE, N. MONMARCHÉ, and G. VENTURINI. A new clustering algorithm based on the chemical recognition system of ants. In F. van Harmelen, editor, Proceedings of the 15th European Conference on Artificial Intelligence, pages 345-349, Lyon, France, july 2002. IOS Press.

[LUM 94] E.D. LUMER and B. FAIETA. Diversity and adaptation in populations of clustering ants. In D. Cliff, P. Husbands, J.A. Meyer, and Stewart W., editors, Proceedings of the Third International Conference on Simulation of Adaptive Behavior (SAB), pages 501-508. MIT Press, Cambridge, Massachusetts, 1994.

[MON 99] N. MONMARCHE, M. SLIMANE, and G. VENTURINI. On improving clustering in numerical databases with artificial ants. In D. Floreano, J.D. Nicoud, and F. Mondala, editors, 5th European Conference on Artificial Life (ECAL'99), volume 1674 of Lecture Notes in Artificial Intelligence, pages 626-635, Swiss Federal Institute of Technology, Lausanne, Switzerland, 13-17 September 1999. Springer-Verlag.

[PAR 02] R.S. PARPINELLI, H.S. LOPES, and A.A. FREITAS. Data mining with an ant colony optimization algorithm. IEEE Transactions on Evolutionary Computation, 6(4). Pages 321-332, 2002.

[VER 94] G. VERLEY. Contribution à la validation des réseaux connexionistes en reconnaissance des formes. Thèse de doctorat de l'Université de Tours, 1994.

FaUR : Méthode de réduction unidimensionnelle d'un tableau de contingence

Erray Walid

Laboratoire ERIC,
Université Lumière Lyon 2,
Bâtiment L, 5 Avenue Pierre Mendès-France,
Code Postal 69500, Bron France

RÉSUMÉ. Dans cet article nous proposons une méthode de réduction unidimensionnelle d'un tableau de contingence appelée FaUR. A l'aide des propriétés des critères utilisés, nous donnons une version optimisée de FaUR, qui permet de rendre la méthode applicable aux grands jeux de données. La méthode FaUR permet de discrétiser des variables continues. Elle permet également de traiter des variables discrètes. Ainsi, les utilisations de FaUR peuvent être très diverses. Elle peut être utilisée aussi bien pour la construction d'arbres de décisions que pour la classification de textes (Déterminer l'ensemble de mots qui décrivent le mieux les documents).

MOTS-CLÉS : Tableau de contingence, Réduction unidimensionnelle, Discrétisation

1 Introduction

Le tableau de contingence *T* est un moyen particulier de représenter simultanément deux variables X_v et X_w sur une même population $\prod$, qu'elles soient discrètes ou bien continues. Dans le domaine de l'apprentissage supervisé, il sert principalement à représenter une variable exogène X_k et la variable endogène Y. Par la suite, nous noterons T_k : le tableau de contingence issu de ce croisement.
Afin d'évaluer l'interaction entre la variable X_k et la variable Y à l'aide du tableau de contingence, plusieurs types de mesures existent [ZIG 00]. Le tableau de contingence (ou tableau croisé) est un outil très fréquemment utilisé en apprentissage à des fins très diverses. En apprentissage supervisé, la plupart des algorithmes basés sur les arbres de décisions se basent sur l'utilisation de tableaux de contingence [BRE 84][QUI 93][ZIG 96][KAS 80]. La plus grande différence entre ces algorithmes réside dans les traitements qu'ils effectuent sur les tableaux de contingence et les mesures qu'ils utilisent pour apprécier leurs qualités. Dans cet article, nous proposons une méthode qui permet de réduire la taille du tableau de contingence. Cette réduction est unidimensionnelle : elle est effectuée seulement sur la variable X_k. Dans une première partie, nous donnons le principe général de la version initiale de l'algorithme. Ensuite, nous exposons des équations et théorèmes liés aux critères utilisés et qui nous permettrons d'aboutir à une version optimisée de l'algorithme FaUR.

2 Méthode FaUR

2.1 Présentation

La méthode que nous proposons est une méthode de réduction unidimensionnelle valable aussi bien pour les variables quantitatives que qualitatives. Le principe général est le suivant : Nous partons de la partition la plus fine, nous chercherons par la suite, à chaque itération, les deux meilleures colonnes à fusionner. Ces deux colonnes sont celles dont la fusion maximise la valeur globale de *Tschuprow* [TSC 21].

L'algorithme s'arrête quand aucune fusion ne peut augmenter la valeur t_s de la mesure de *Tschuprow*. Afin de contrôler la procédure de réduction, nous introduisons un paramètre θ qui représente un gain minimum à obtenir. Dans le cas d'une variable quantitative, nous commençons par trier les valeurs de la variable. Aussi, dans ce cas, seuls les colonnes adjacentes peuvent être fusionnés.

2.2 Algorithme Initial

La version initiale de *FaUR* est donnée par l'algorithme 1 :

Algorithme 1 FaUR

Arrêt $\Leftarrow$ FAUX
tant que Arrêt = FAUX **faire**
  $t_{s_{min}} = t_{s_0} = t_s(T_k)$: t_s du tableau de contingence T_k
  pour tout couple $(x_z, x_{z'})$ **faire**
    Calculer $t_{s_{z,z'}}$: t_s de T_k après fusion des deux colonnes z et z'
    si $t_{s_{z,z'}} > (t_{s_{min}} \times \theta)$ **alors**
      $t_{s_{min}} \leftarrow t_{s_{z,z'}}$, $z_1 = z$, $z_2 = z'$
    fin si
  fin pour
  si $t_{s_{min}} > (t_{s_0} \times \theta)$ **alors**
    Fusionner les deux colonnes z_1 et z_2
  sinon
    Arrêt = $VRAI$
  fin si
fin tant que

Algorithme 1. FaUR Initiale

2.3 Complexité Algorithmique

Nous noterons l et m respectivement les nombres de modalités des variables X_k et Y. Nous allons dans cette étude distinguer deux cas :

2.3.1 X_k est une variable quantitative :

Nous allons donner une écriture simplifiée de l'algorithme avec la complexité qui correspond à chaque étape :

- Tri des l valeurs de X_k : $l \log l$ (Utilisation d'un arbre binaire),
- Répéter (au maximum l fois)
  - Calculer t_s $(l \times m)$ pour l fusions possibles : $(l \times m) \times l$,
  - Retenir la meilleure fusion : l
  - Test d'arrêt : 1

La complexité algorithmique sera donc en $O(l^3 m)$.

Remarque : Dans le cas où la variable X_k est quantitative et où Y est la variable endogène, FaUR se rapporte à une discrétisation. Or, une complexité en $O(l^3 m)$ rend notre méthode de discrétisation inutilisable surtout dans le cas où l est très élevé (Nombre d'individus N très grand).

2.3.2 X_k est une variable qualitative :

Nous allons donner une écriture simplifiée de l'algorithme avec la complexité qui correspond à chaque étape :

- Répéter (au maximum l fois)

- Calculer t_s *(l × m)* pour *l(L+1)/2* fusions possibles : *(l × m) × l(l+1)/2*,
- Retenir la meilleure fusion : l
- Test d'arrêt : 1

La complexité algorithmique sera donc en $O(l^4m)$.

Remarque : La complexité algorithmique est très importante. Il est impossible d'utiliser la version initiale de l'algorithme FaUR par exemple pour construire des arbres de décision comme pour la méthode ChAid [KAS 80].

2.4 Conclusion

Quelque soit la nature de la variable X_k, la complexité de FaUR reste très élevée. Elle rend l'algorithme inapplicable avec des données réelles où *l* et *m* peuvent être très grands. Il est donc nécessaire de trouver une optimisation qui réduit la complexité de FaUR.

3 FaUR optimisée

3.1 Recherche de la meilleure fusion

A une étape donnée, le meilleur couple de colonne est celui dont la fusion maximise le gain en t_s.

Théorème : Parmi deux couples de colonnes (z_1,z_2) et (z_3,z_4), le meilleur est dont la fusion donne la plus grande valeur du Chi2. $\Delta ts(z_1,z_2) > \Delta ts(z_3,z_4)$ ssi $\chi^2(z_1,z_2) > \chi^2(z_3,z_4)$.

Ainsi, l'évaluation de la valeur de t_s revient à évaluer la valeur de χ^2. χ^2 étant la mesure du Chi2[PEA, 04]

3.2 Test d'arrêt

FaUR s'arrête quand aucun gain dans la mesure de Tschuprow n'est possible.

Théorème : FaUR s'arrête quand la perte dans la valeur globale du Chi2 ne dépasse pas $\sqrt{1-\frac{1}{l'-1}}$

$$\Delta t_{s_{(z,z')}} > 0 \quad ssi \quad \chi^2_{(z,z')} > \chi^2\sqrt{1-\frac{1}{l'-1}}$$

Nous rappelons qu'aucune fusion n'est effectuée que si le nombre de colonnes est supérieur ou égal à 2. Nous supposons donc que *l'* est strictement supérieure à 2. Ainsi la valeur de $\sqrt{1-\frac{1}{l'-1}}$ est inclus dans l'intervalle] 0; 1[. Nous pouvons donc dire que cette valeur correspondant bien à une perte dans la mesure du Chi2.

3.3 Ecritures simplifiées

3.3.1 Mesure du Chi2

Nous rappelons la mesure du Chi2 par l'équation suivante :

$$\chi^2 = \sum_{j=1}^{l} \chi^2_{C_j}$$

Où χ^2_{Cj} est la contribution de la $j^{ème}$ colonne à la mesure globale du Chi2. χ^2_{Cj} peut être écrit sous la forme suivante :

$$\chi^2_{C_j} = n_{.j}\left(\sum_{i=1}^{m}\frac{\left(\frac{n_{ij}}{n_{.j}}\right)^2}{\frac{n_{i.}}{N}} - 1\right)$$

3.3.2 Gain dans la mesure du Chi2

Le Chi2 issu de la fusion des deux colonnes *z* et *z'* est décrit par l'équation suivante :

$$\chi^2_{(z,z')} = \chi^2 + \Delta\chi^2_{C_{(z,z')}}$$

A l'aide de cette équation, nous pouvons affirmer que la recherche de la fusion qui maximise le *Chi2* revient à chercher la fusion qui maximise le $\Delta\chi^2$ qui lui correspond. Il faut également retenir, que la valeur du $\Delta\chi^2$ est calculée de manière locale. Ainsi, après la fusion de deux colonnes *z* et *z'*, le $\Delta\chi^2$ qui correspond à un couple (z_1, z_2) reste inchangé. Grâce à ces expressions nous pouvons proposer une première simplification de FaUR. Ainsi, d'une manière intuitive, nous pouvons penser à calculer les $\Delta\chi^2$ qui correspondent à chaque couple de colonnes. Ensuite, la fusion qui correspond au plus grand $\Delta\chi^2$ est effectuée. Si (*z*, z') est le couple de colonnes fusionnées, alors seuls les $\Delta\chi^2$ qui incluent l'une de ces deux colonnes est recalculé.

3.3.3 Simplification de l'expression du DeltaChi2

Nous pouvons écrire l'expression de $\Delta\chi^2$ local de la fusion de *z* et *z'* sous la forme suivante :

$$\Delta\chi^2_{C_{(z,z')}} = -\frac{N}{n_{.z} + n_{.z'}} \sum_{i=1}^{m} \frac{n_{iz}^2 \frac{n_{.z'}}{n_{.z}} + n_{iz'}^2 \frac{n_{.z}}{n_{.z'}} - 2n_{iz}\, n_{iz'}}{n_{i.}}$$

3.3.4 Mise à jour de la liste des Chi2

Le χ^2 local de la fusion des colonnes z_1, z_2 et z_3 peut s'écrire sous la forme suivante :

$$\chi^2_{C_{((z_1,z_2),z_3)}} = (n_{.z_1} + n_{.z_2} + n_{.z_3}) \left(\frac{N}{(n_{.z_1} + n_{.z_2} + n_{.z_3})^2} ((\Gamma_4 + \Gamma_5 + \Gamma_6) - (\Gamma_1 + \Gamma_2 + \Gamma_3)) - 1 \right)$$

Où Γ_4, Γ_5 et Γ_6 sont les χ^2 locaux correspondant aux fusions (z_1,z_2), (z_1,z_3) et (z_2,z_3) et où Γ_1, Γ_2 et Γ_3 les χ^2 locaux des colonnes z_1, z_2 et z_3. Ainsi, nous pouvons mettre à jour les χ^2 locaux à l'aide des χ^2 déjà calculés sans avoir à lire des données sur le tableau de contingence ce qui permet d'obtenir un gain important en temps de calcul.

3.3.5 Mise à jour de la liste des $\Delta\chi^2$

Le $\Delta\chi^2$, correspondant à la fusion des colonnes z_1, z_2 et z_3, peut être écrit sous la forme suivante :

$$\Delta\chi^2_{C_{((z_1,z_2),z_3)}} = \chi^2_{C_{((z_1,z_2),z_3)}} - \chi^2_{C_{(z_1,z_2)}} - \chi^2_{C_{z_3}}$$

A l'aide de cette équation la mise à jour d'un $\Delta\chi^2$ ne coûte que 1 instruction puisque les autres parties de l'équation sont déjà calculés.

3.4 Algorithme FaUR optimisé

A l'aide de ces écritures simplifiées et de ces équations, nous pouvons proposer une version optimisée de l'algorithme FaUR. Il suffit de calculer la première valeur du χ^2, les χ^2 correspondent aux fusions de deux colonnes et les $\Delta\chi^2$ qui en résultent. Nous trions les $\Delta\chi^2$ et nous effectuant la fusion des deux colonnes qui ont le plus grand $\Delta\chi^2$. Ensuite, la liste des χ^2 et par conséquent la liste des $\Delta\chi^2$ sont mises à jour à l'aide des équations présentées précédemment avec un coût très faible.

Cette version représente une complexité très acceptable. Elle est en $O(l \log l)$ pour les variable continue (Comparable aux méthodes Chimerge [KER 91], Fusinter [ZIG 98] etc.) et en $O(l^2 m)$ pour les variables discrètes.

La version optimisée de FaUR est décrite par l'algorithme 2.

Algorithme 2 FaUR Optimisé

Arrêt $= FAUX$
$l' = l$
Calculer le χ^2 initial
Calculer les χ^2_C
Calculer les $\Delta\chi^2_C$
Trier la liste des $\Delta\chi^2_C$
tant que $((l' > l)$ et (Arrêt $= Faux$)) **faire**
  Soit $\Delta\chi^2_{C_{(z_1,z_2)}}$ le meilleur $\Delta\chi^2_C$
  si $\chi^2_{(z_1,z_2)} > \chi^2\sqrt{1 - \frac{}{l'-1}}$ **alors**
    Fusionner z_1 et z_2
    Mettre à jour la liste des χ^2_C
    Mettre à jour la liste des $\Delta\chi^2_C$
    $l' = l' - 1$
  sinon
    Arrêt $= VRAI$
  fin si
fin tant que

Algorithme 2. FaUR Optimisé

4 Conclusion

Le gain apporté par la version optimisée de FaUR est très important puisque nous passons d'une complexité en $O(l^4m)$ à une complexité en $O(l^2m)$ pour les variables discrète par exemple. Ainsi, l'utilisation de FaUR dans des arbres de décisions devient réalisable et peu coûteuse. Egalement les premiers tests de la méthode FaUR en temps que méthode de discrétisation étaient assez satisfaisants.

Bibliographie

[BRE 84] BREIMAN L., FRIEDMAN J., OLSHEN R., STONE C., *Classification and regression trees*, Chapman & Hall, 1984.

[KAS 80] Kass G.V. An exploratory technique for investigating large quantities of categorical data, Applied Statistics, 119-127, 1980.

[KER 91] Kerber R., Chimerge discretization of numeric attributes, Proceedings of the 10th International Conference on Artificial Intelligence, 123-128, 1991.

[QUI 93] Quinlan J.R., C4.5: Programs for machine learning, Morgan Kaufman, 1993.

[TSC 21] Tschuprow A.A., On the mathematical expectation of moments of frequency distribution. Biometrika, 185-210, 1921.

[ZIG 00] Zighed D.A., Rakotomalala R., Graphes d'induction, HERMES Science Publications, 2000.

[ZIG 96] Zighed D.A., Rakotomalala R., Sipina-W for Windows: User's Guide, Laboratory ERIC, University of Lyon 2, 1996.

[ZIG 98] Zighed D.A., Rabaseda S., Rakotomala R., Fusinter: a Method for discretization of continuous attributes for supervised learning, International Journal of Uncertainty, Fuzzinzss and Knowledge-Based Systems, 307-326.

[BOU 04] Boulle M., Khiops: A statistical discretization method of continuous attributes, Machine Learning, 53-69, 2004.

Prix Simon Régnier

Graphes de rigidité et structuration d'un système de classes

Christophe Osswald

Laboratoire E^3I^2 – EA 3876
ENSIETA
2, rue François Verny, 29206 Brest Cedex 9

L'un des objectifs de la classification est d'expliciter les relations entre éléments et classes. Les modèles usuels sont les hiérarchies, les pyramides et les systèmes de classes arborés, classiques en biologie. Ils nécessitent de faire le choix d'une structure pour approcher les données, et la plupart mènent à des problèmes NP-difficiles.

Dans le cadre de l'analyse de la similitude (Flament *et al.*, 1962-1981), il n'y a pas d'approximation, et l'interprétation des données passe par un *graphe de rigidité* le plus petit possible, tel que les classes d'un système donné en soient des classes connexes. Il existe plusieurs méthodes pour engendrer un système de classes à partir d'une dissimilarité : cliques maximales (classes naturelles), boules, 2-boules, réalisations (Brucker, 2003). Le problème est NP-difficile dans le cas général et pour les trois premières méthodes ; il existe un algorithme en $O(n^4)$ pour les réalisations.

Il est aisé d'identifier les hyperarbres comme étant les systèmes de classes rigides sur un arbre. Ceci s'étend aux *hypercycles*, dont un graphe de rigidité est un cycle, en $O(n^4)$ opérations, ainsi que les systèmes de classes dont un graphe de rigidité a n arêtes.

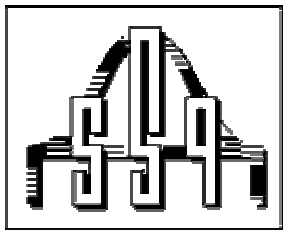
Pour une dissimilarité arboricole (dont les classes admettent un arbre de rigidité), les arbres de rigidité des boules sont aussi ceux des 2-boules, des classes et des réalisations. Au-delà de ce modèle, l'équivalence entre les graphes de rigidité de ces systèmes de classes se limite à une chaîne d'inclusion. Le nombre d'arêtes d'un graphe de rigidité minimum des réalisations d'une dissimilarité quelconque nous permet ainsi de construire une mesure de la structuration d'icelle. Nous confrontons cette mesure à plusieurs modèles de dissimilarité aléatoire, afin notamment d'analyser des dissimilarités réelles et graphiques.

Ces méthodes ont été appliquées à des données issues de la psychologie de la mémoire, de l'analyse de données textuelles, de l'imagerie sonar et de la génétique.

[BRU 79] FLAMENT C., DEGENNE A., VERGES P., “Analyse de similitude ordinale”, *Informatique et Sciences Humaines*, vol. 40-41, 1979, pp 223-231.

[FLA 62] FLAMENT C., “L'analyse de similitude”, *Cahiers du Centre de Recherche Opérationnelle*, vol. 4, 1962, pp 63-97.

[BRU 03] BRUCKER F., “Réalisations de dissimilarités”, *Rencontres de la Société Francophone de Classification*, 2004, pp 7-10.



Fondation
"La Science Statistique"

Laboratoires
universitaires Bell