XVIèmes Rencontres

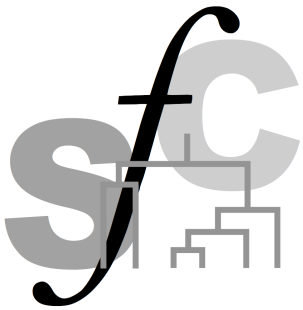
de la Société Francophone de Classification

Du 2 au 4 septembre 2009 Grenoble

Président
Gérard d'Aubigny

Actes

Actes des 16èmes Rencontres de la Société Francophone de Classification



2-4 septembre 2009, Grenoble, France

Gérard d'Aubigny, président du comité de programme

http://sfc-2009-grenoble.imag.fr/

Préface

Construire le programme scientifique d'un congrès francophone de classification tient actuellement de la gageure tant il est vrai que les contours des thèmes éligibles a désormais du mal à endosser l'habit du classificateur taillé dans une pièce tissée par le dictionnaire. Les définitions de ce dernier sont sans doute désormais datées. Hélas pour le dictionnaire, le nombre des domaines d'application a explosé, et chacun tente d'arbitrer entre la référence à un vocabulaire commun trans-disciplinaire et la nécessité de redire ou de ré-inventer sous des formes parfois imagées compréhensibles par le chercheur moyen de sa discipline support. Ces redites sont parfois fort productives !

Ajoutons à cela que le vocabulaire devient aussi ambigu à cause d'emprunts souvent inconscients et non maîtrisés à la langue anglaise. On comprend alors que le travail des arbitres devient essentiel pour simplement reconnaître la question traitée par certaines soumissions.

Il est sans doute sage de promouvoir un esprit d'ouverture en cette période transitoire et par conséquent de laisser ouvertes toutes les ambiguïtés et toutes les interprétations possibles du terme même de classification. Cette utilisation générique présente en particulier l'avantage d'adapter le programme aux centres d'intérêt exprimés tout à la fois par les communications soumissent qui trouvent leur motivation dans l'exploration continuée de pistes réputées classiques au vu des éditions précédentes mais aussi aux tentatives plus audacieuses qui tentent d'attirer l'attention sur des sujets neufs ou plus rarement explorés.

Dans les deux cas, le comité de programme subit plus qu'il n'oriente le contenu des journées, parce qu'il est contraint par les thématiques qu'imposent ceux qui se sentent légitimement classificateurs. Cela n'est pas sans risque : effets de mode et sclérose de communautés fonctionnant en milieu fermé sont les plus aisément repérables. Il est alors indispensable de se poser la question suivante : quels thèmes actuellement importants ou fortement étudiés dans d'autres communautés francophones ou par des groupes non francophones ont été semble-t-il oubliés dans les communications soumises ?

Ce type d'omission me semble le plus souvent dû aux dangers de toute nomenclatures : ces sujets ne sont pas proposés parce qu'ils ne semblent pas rentrer dans le champ de compétences légitimement attribuable aux classificateurs.

Nous avons tenté de limiter ce risque en oeuvrant de trois façons. La première repose sur le choix des communications invitées. Mais ses effets sont limités car comme la sélection des communications libres, ce choix doit intervenir avec des contraintes de dates. La deuxième a consisté à solliciter des communications libres afin de donner une place à certains thèmes initialement non abordés. Là encore l'exercice atteint vite ses limites. La troisième méthode consiste à jouer de façon différenciée sur les évaluations de la qualité de certaines soumissions : plutôt que de dire non, nous avons préféré motiver certains auteurs à faire évoluer leur proposition pour permettre une présentation portant sur ce thème, améliorée pendant la période estivale. Ces trois démarches n'ont certes pas suffit à effacer certaines frustrations du comité de programme. Après tout, il est aussi dans ses fonctions de rêver un peu au contenu du congrès idéal.

La dure réalité revenant au galop, les fruits de nos efforts ont parfois été mis à mal par les contraintes de découpage en sessions. Cela est classique, mais plus apparent dans des congrès

de taille limitée. Il ne me paraît pas possible de défendre le compromis retenu, sans présenter en même temps nos excuses aux auteurs dont la communication apparaîtrait à leurs yeux, placée dans une session inadaptée. L'intitulé des sessions est lui aussi un compromis, destiné seulement à annoncer une thématique majoritaire et non à couvrir tous les exposés ....

On doit enfin rajouter au chapitre des difficultés de planification, le positionnement des communications sur des applications. Il nous a semblé inadapté de prévoir des sessions applications, du fait de l'importance que nous attribuons à l'existence de telles soumission. Nous avons donc souvent utilisé ces communications pour équilibrer les sessions, en partant du principe que tous les congressistes seront intéressés par des applications.

Au bilan, nous avons souhaité organiser chaque créneau horaire consacré aux communications libres en deux sessions en parallèle. Une exception à cette règle concerne le jeudi après midi. Nous avons en effet souhaité offrir à tous les participant la possibilité d'assister à la session CLADAG proposée et organisée par nos collègues italiens. Nous avons été très sensibles à cette proposition amicale et prometteuse. Nous tenions à en faire profiter le plus grand nombre de participants.

Le présent document fournit les textes de présentation des communications soumises. Quelque soient les difficultés techniques du travail du comité de programme évoquées plus haut, l'essentiel est ailleurs : Il me semble utile de noter que la diversité et la qualité de bien des textes reçus ont rendu très intéressantes et parfois passionnantes les lectures du comité de programme. Il me revient donc, en guise de conclusion, de remercier en son nom tous les auteurs de soumissions. Nous ne pouvons qu'espérer que notre bilan optimiste soit partagé par tous les participants au congrès SFC'2009.

Gérard d'Aubigny, Président du CP

Comité de programme

Président : Gérard D'Aubigny	LJK, UPMF
• Antoniadis Anestis	UJF, Grenoble 1
• Azzag Hanene	LIPN, Paris 13
• Bock Hans-Hermann	RWTH Aachen, Allemagne
• Brito Paula	Université de Porto, Portugal
• Bruckner François	Université de Metz
• Celeux Gilles	INRIA, Paris
• Chavent Marie	IMB, Bordeaux
• Cleuziou Guillaume	LIFO, Orléans
• Crémilleux Bruno	GREYC, Caen
• Demongeot Jacques	UJF, Grenoble 1
• Drouet d'Aubigny Gérard	UPMF, Grenoble 2
• Girard Stéphane	INRIA, Montbonnot
• Govaert Gérard	UTC, Compiègne
• Guénoche Alain	CNRS, Marseille
• Hardy André	Université de Namur, Belgique
• Hérault Jeanny	INPG, Grenoble 1
• Nédellec Claire	INRA, Jouy en Josas
• Lebba Mustapha	LIPN, Paris 13
• Lechevallier Yves	INRIA, Rocquencourt
• Poncelet Pascal	Université Montpellier 2
• Rouveirol Céline	Institut Galilée, Paris Nord
• Sebag Michèle	LRI, Orsay, Paris Sud
• Vichi Maurizio	La sapienza Roma1, Italie
• Zighed Djamel	Université Lumière Lyon 2

Comité de d'organisation

Présidents : Ahlame Douzal	TIMC-IMAG, UJF
Gilles Bisson	TIMC-IMAG, CNRS
• Anne Guérin	GIPSA-lab, Grenoble-INP
• Eric Gaussier	LIG, UJF
• Cécile Amblard	TIMC-IMAG, UJF
• Jérôme Gensel	LIG, UPMF

Partenaires de la conférence

Table des matières

POSTERS

Réduction non-linéaire de dimension et visualisation

Michel Verleysen

Université catholique de Louvain
Département DICE - Microelectronics laboratory,
3 place du Levant,
B-1348 Louvain-la-neuve, Belgium

michel.verleysen@uclouvain.be

RÉSUMÉ : La réduction de dimension a pour ambition de produire des représentations en faible dimension d'ensembles de données en haute dimension. Un des objectifs principaux de la réduction de dimension est la visualisation de données (en dimension 2 ou 3). Si les méthodes linéaires comme l'analyse en composantes principales et le MDS (Multi-Dimensional Scaling) ont longtemps été les seuls outils disponibles, des méthodes non-linéaires sont apparues plus récemment. Elles permettent en général de mieux projeter des ensembles de données dont la distribution se trouve sur ou autour d'une hyper-surface non plane de l'espace initial ; dans ce cas en effet, une méthode linéaire résulte généralement en un écrasement.

De nombreuses méthodes non-linéaires de réduction de dimension ont été proposées récemment. Une grande partie d'entre elles se basent sur l'optimisation d'un critère de respect de distances entre paires de points. Le critère peut être simple (souvent quadratique, éventuellement après transformation non-linéaire des données), permettant une optimisation de type algébrique. D'autres critères plus pertinents face aux objectifs de la réduction de dimension ont également été définis, permettant par exemple de sous-pondérer les paires de données éloignées, dont la distance est moins importante à prendre en considération pour un objectif de visualisation. Dans ce cas, l'optimisation du critère requiert en général des méthodes itératives basées sur le gradient.

L'abondance des méthodes proposées pose la question de la mesure de leur qualité. De façon surprenante, peu de mesures de qualité tiennent comptent des objectifs des méthodes de projection, objectifs qui peuvent refléter des points de vues très différents. Par exemple, pour projeter une sphère sur un plan, vaut-il mieux l'écraser, ou la déchirer ?

L'objectif de cet exposé est de présenter un panorama, non exhaustif, des méthodes de projection non-linéaires les plus courantes, ainsi que des mesures de qualité permettant leur comparaison objective, tout en tenant compte du but poursuivi par l'utilisateur.

MOTS-CLÉS : Réduction de dimension, projection non linéaire, visualisation.

*I*nférence de langages stochastiques rationnels

Conférence invitée (présentée au congrès SFC2009)

François Denis

Laboratoire d'Informatique Fondamentale de Marseille (LIF)
UMR 6166 - CNRS - Univ. de la Méditerranée - Univ. de Provence
Centre de Mathématiques et d'Informatique (CMI) 39, rue F. Joliot-Curie
13453 Marseille Cedex 13
francois.denis@lif-univ-mrs.fr

RÉSUMÉ. *Un* langage stochastique *est une distribution de probabilités définie sur l'ensemble des mots Σ^* que l'on peut former sur l'alphabet fini Σ. Un problème classique en* inférence grammaticale probabiliste *consiste à tenter d'inférer un langage stochastique p, dans une certaine classe de modèles probabilistes, à partir d'un échantillon i.i.d. de mots produits selon p. La classe de modèles la plus couramment utilisées est celle des Modèles de Markov Cachés (HMM), par ailleurs équivalente à celle des automates probabilistes (PA). Nous décrirons quelques résultats classiques puis nous montrerons l'avantage qu'il y a, du point de vue de l'inférence, à se placer dans la classe plus expressive des* langages stochastiques rationnels, *c'est-à-dire calculable par un automate pondéré (WA). A chaque langage stochastique p, on peut associer l'ensemble de ses langages résiduels $\dot{u}p$ définis par $\dot{u}p(v) = p(uv)$ et l'on peut montrer que p est rationnel ssi ses langages résiduels engendrent un espace vectoriel de dimension fini. Inférer un langage stochastique rationnel revient donc à identifier un espace vectoriel de dimension fini dans l'espace des fonctions, ou séries formelles, définies de Σ^* dans $\mathbb{R}$, à partir des approximations de ses résiduels calculées à partir de l'échantillon d'apprentissage : cela peut être réalisé de manière incrémentale, ou globale via une Analyse en Composantes Principales (PCA). Plusieurs algorithmes seront présentés, accompagnés de résultats de convergence asymptotique et de résultats expérimentaux.*

MOTS-CLÉS : *Inférence, langages stochastiques rationnels.*

Approximations en norme du supremum: simplicité et complexité

Bernard Fichet

LIF, 163, avenue de Luminy - Case 901
F-13288 Marseille Cédex 9, France, bernard.fichet@lif.univ-mrs.fr

Summary. Dans cet exposé, nous dressons un vaste panorama des approximations en norme du supremum pour nombre de structures classiques de la classification et de l'analyse des données. Seront en particulier concernées, les structures ultramétriques et hiérarchiques, leurs extensions Robinsoniennes et arborées, que ces dernières soient métriques ou de dissimilarité, les métriques linéaires, ainsi que les régressions, soit convexes, soit monotones sur un ensemble partiellement ordonné.

Tous ces résultats seront situés dans un cadre très général, tel qu'il fut décrit dans [2], et où les concepts de sous-dominante et de sur-dominée jouent un rôle crucial. Dans un tel cadre, toute structure sera identifiée à un sous-ensemble K d'un espace vectoriel E, p-dimensionnel, et muni de la norme du supremum, relativement à une base fixée.

Le problème d'optimisation est alors le suivant: donnés un ou n vecteurs $u_1, \cdots, u_n$ de E, trouver $\hat{y}$ de K, solution (si elle existe) de:
$inf_{y \in K} max_{j=1,\cdot,n} ||y - u_j||_\infty$. Un vecteur $u_*(u^*)$ de K est appelé sous-dominante (sur-dominée) dans K de $u \in E$, s'il est le plus grand (petit) vecteur de K inférieur (supérieur) à u (au sens de l'ordre partiel défini par les coordonnées). L'existence pour tout u d'une sous-dominante (sur-dominée) revient à dire que K est un sup- (inf-) demi-treillis. On montre alors que sous cette condition, et pourvu que K soit invariant par translation le long de la diagonale principale de E, une solution du problème général est donnée par translation de la sous-dominante (sur-dominée) de l'infimum (supremum) des $u_1, \cdots, u_n$. Ces conditions sont remplies pour toutes les structures sus-évoquées, avec un calcul analytique ou algorithmique aisé d'une sous-dominante et/ou d'une sur-dominée, à l'exception des structures arborées, Robinsoniennes et linéaires, pour lesquelles la NP-difficulté de calcul d'une solution a été prouvée [1],[3],[7]. On notera en particulier, l'extrême simplicité théorique et algorithmique offerte par l'existence d'une sous-dominante dans le cas ultramétrique, par rapport à la première approche de [5].

La NP-difficulté de l'approximation par une distance arborée, peut être levée si l'on se restreint aux dissimilarités préservant les distances à un point pivot donné. En effet, une transformation bien connue, dite de Farris [6] , établit une bijection linéaire entre dissimilarités et leurs transformées, assurant qu'il y a équivalence entre la condition dite des quatre points, base des dissimilarités arborées, et ultramétricité de la transformée [8]. Sous ces conditions, il y a existence d'une sous-dominante via

l'ultramétricité, et ainsi une solution au problème. En outre, comme établi dans [1], une simple retouche faite sur les arêtes de l'arbre fournit pour le problème général un algorithme d'approximation de facteur 3. L'usage d'un pivot s'avère également fructueux pour les métriques linéaires. Dans [7], les auteurs proposent un algorithme de facteur 2 pour celles-ci. L'existence d'un algorithme à facteur constant pour les structures Robinsoniennes a été montrée récemment par une démarche toute autre [4]. L'algorithme et sa preuve sont très sophistiqués et reposent en partie sur le fait que l'erreur optimale appartient à une famille finie bien spécifiée, d'erreurs potentielles.

Nous discutons in fine du consensus en norme du supremum, qui non seulement bénéficie lui aussi d'un cadre où existe une sous-dominante (sur-dominée), mais s'en trouve renforcé puisqu'alors le supremum (infimum) des données $u_1, \cdots, u_n$ appartient à la structure K souhaitée. C'est ainsi que dans [2], les auteurs pallient la non-unicité de la solution, en proposant inductivement sur les solutions précédentes par récurrence finie polynomiale, un unique consensus dit universel, rejoignant ainsi dans le cas de la régression isotone la solution proposée par [9].

Key words: ultramétrique, distance d'arbre, régréssion monotone, norme du supremum, NP-difficulté, approximation à facteur constant.

References

1. R. Agarwala, V. Bafna, M. Farach, B. Narayanan, M. Paterson, M. Thorup, On the approximability of numerical taxonomy: Fitting distances by tree metrics. In *Proceedings of the 7th Annual ACM-SIAM Symposium on Discrete Algorithms*, 1996.
2. V. Chepoi, B. Fichet, l_∞-Approximation via Subdominants. *Journal of Mathematical Psychology*, **44**: 600–616, 2000.
3. V. Chepoi, B. Fichet, M. Seston, Seriation in the presence of errors: NP-hardness of l_∞-fitting Robinson structures to dissimilarity matrices. *Journal of Classification*, in press.
4. V. Chepoi, M. Seston, Seriation in the presence of errors: a factor 16 approximation algorithm for l_∞-fitting Robinson structures to distances. *Algorithmica*, in press.
5. M. Farach, S. Kannan, T. Warnow, A robust model for finding optimal evolutionery trees. *Algorithmica*, **13**: 155–179, 1995.
6. J.S. Farris, A.G. Kluge, M.J. Eckardt, A numerical approach to phylogenetic systematics. *Systematic Zoology*, **19**: 172–189, 1970.
7. J. Hastad, L. Ivansson, J. Lagergren, A numerical approach to phylogenetic systematics. *Systematic Zoology*, **19**: 172–189, 1970.
8. B. Leclerc, Minimum spanning trees for tree metrics: abridgment and adjustements. *Journal of Classification*, **12**: 207–241, 1995.
9. V.A. Ubhaya, Isotone optimization, I, II. *Journal of Approximation Theory*, **12**: 146–159, 1974.

Ordonnancement et optimisation de la courbe ROC

Nicolas Vayatis

ENS Cachan
Centre de mathématiques et de leurs applications,
61 avenue du président Wilson,
94235 Cachan cedex, France

Nicolas.Vayatis@cmla.ens-cachan.fr

RÉSUMÉ. On considère la question de l'ordonnancement de données étiquetées de labels binaires. Dans ce contexte, l'objectif de l'apprentissage est de retrouver l'ordre induit sur l'espace des données par la fonction de régression inconnue.Les éléments optimaux pour ce problème sont des fonctions à valeurs réelles, appelées règles de scoring, ayant une courbe ROC optimale. Une idée naturelle est alors d'estimer les règles de scoring optimales en optimisant la version empirique sur l'échantillon de la courbe ROC. Dans ce cadre, le critère à optimiser prend des valeurs fonctionnelles et son optimisation requiert simultanément d'approcher et d'estimer la fonction cible. Dans l'exposé, seront présentées trois stratégies différentes pour l'ordonnancement optimal : (1) par sélection de règles à base d'histogrammes réguliers parmi les partitions de l'espace d'entrée, (2) par superposition des solutions de problèmes de classification avec contrainte de masse, (3) par partitionnement récursif dans l'esprit des arbres

MOTS-CLÉS : ordonnancement, apprentissage, scoring, courbe ROC.

Forêts aléatoires : importance et sélection de variables

Jean-Michel Poggi

Université d'Orsay, Lab. de Mathématiques
bat. 425, 91405 Orsay, France
&
Université Paris Descartes, IUT Dept. STID
143 avenue de Versailles, 75016 Paris, France

Jean-Michel.Poggi@math.u-psud.fr, Jean-Michel.Poggi@parisdescartes.fr

RÉSUMÉ. Introduite par Leo Breiman en 2001, la méthode des forêts aléatoires est désormais largement utilisée tant en classification qu'en régression avec un succès spectaculaire. On s'intéresse au comportement du score d'importance des variables basé sur les forêts aléatoires et on examine deux problèmes classiques de sélection de variables. Le premier est de dégager les variables importantes à des fins d'interprétation tandis que le second, plus restrictif, vise à se restreindre à un sous-ensemble suffisant pour la prédiction. La stratégie générale procède en deux étapes : le classement des variables basé sur les scores d'importance suivie d'une procédure d'introduction ascendante séquentielle des variables.

Cet exposé est tiré d'un travail en collaboration avec R. Genuer et C. Tuleau. Pour en savoir plus :
Robin Genuer, Jean-Michel Poggi, Christine Tuleau. "Random Forests: some methodological insights". Rapport de Recherche INRIA, Nov. 2008, 32 p. http://hal.inria.fr/inria-00340725/fr/

MOTS-CLÉS : Random Forests, Classification, régression, sélection de variables, scores d'importance.

Adaptation des modèles d'auto-organisation pour la classification recouvrante

Guillaume Cleuziou

Laboratoire d'Informatique Fondamentale d'Orléans (LIFO)
Université d'Orléans
45067 ORLEANS Cedex 2
Guillaume.Cleuziou@univ-orleans.fr

RÉSUMÉ. *Les méthodes de classification sont en constante évolution du fait de l'explosion des besoins d'analyse décisionnelle, tant en terme de multiplication des données disponibles que de complexité de ces données. Les modèles de classification recouvrante (ou empiétante) ont par exemple montrés leur capacité à générer une organisation plus fidèle aux données d'entrée tout en conservant la simplification attendue par une structuration en classes strictes (partition ou hiérarchie). Par ailleurs les modèles neuronaux non-supervisés (cartes de kohonen) sont plébicités lorsqu'il s'agit de proposer une visualisation de la structure de classes (partition). L'étude proposée dans cet article vise à étendre l'approche de kohonen aux classes recouvrantes. La méthode proposée s'appuie alors sur une version recouvrante de l'algorithme des nuées dynamiques (*OKM*) proposée récemment et en dérive une variante pour la construction de cartes topologiques recouvrantes.*

MOTS-CLÉS : *Classification automatique, classification recouvrante, cartes auto-organisatrices.*

1. Introduction

La problématique de la classification recouvrante s'inscrit dans le processus général d'extraction de connaissances à partir de données ; elle consiste à faire émerger une organisation synthétique d'un ensemble d'individus à l'aide d'une structure de classes (hiérarchique ou non) dans laquelle chaque individu peut appartenir à plusieurs classes. Souvent plus riches et plus adaptés que leurs analogues non recouvrants, les modèles recouvrants accompagnent les avancées réalisées dans le domaine de la classification non-supervisée en général ; à titre d'exemples on notera les adaptations suivantes : les pyramides [DID 84] généralisent les hiérarchies, OKM [CLE 08] généralise l'algorithme des k-moyennes et MOC [BAN 05] correspond à une variante recouvrante des modèles de mélanges.

Outre les problématiques récurrentes inhérentes aux méthodes de classification (paramétrage, évaluation, passage à l'échelle, etc.) la classification recouvrante présente des problématiques propres telles que :

- **le choix du nombre de classes :** dans le cas des modèles recouvrants non hiérarchiques, l'ampleur des recouvrements entre classes est une caractéristique incontournable voire déterminante sur la décision du nombre approprié de classes ; cependant il reste difficile de donner une préférence *a priori* entre une organisation constituée de peu de classes avec de forts recouvrements ou une organisation avec d'avantage de classes moins recouvrantes. On peut illustrer cette problématique de manière pratique en Recherche d'Information en se demandant si un sous-ensemble de textes qui traitent de deux thèmatiques constitue à lui seul une nouvelle classes ou correspond à l'intersection des deux classes thématiques associées.
- **la cohérence topologique des recouvrements :** ce que nous appelons ici "cohérence topologique" correspond au fait que des classes ne puissent se recouvrir que si elles sont proches au sens de la topologie induite par la métrique utilisée. Celle-ci est presque toujours admise dans le processus même de construction de la structure classificatoire : les modèles hiérarchiques recouvrants limitent de fait les chevauchements à des classes voisines qui correspondent à des zones[1] adjacentes sur l'espace de projection des individus, induit par la classification ; dans le cas des méthodes de types réallocations dynamiques ([DAT 68],[CLE 08],[BEZ 81]), les affectations multiples d'un individu sont souvent réalisées parmi les

1. Ces zones peuvent être des intervalles dans le cas de hiérarchies 2D ou des surfaces pour des hiérarchies 3D, etc.

classes les plus proches de l'individu en question, ce dernier assurant ainsi une continuité topologique entre les classes qu'il permet de se faire recouvrir.
– **le coût induit par l'élargissement de l'espace des possibilités :** les hiérarchies (resp. partitions) ne sont que des cas particuliers de pyramides (resp. recouvrements) ; l'espace des solutions possibles étant fortement élargi lorsque l'on considère des structures classificatoires recouvrantes, les algorithmes d'exploration peuvent dans certains cas s'avérer plus complexes. Si la complexité théorique de la variante recouvrante (OKM) des k-moyennes reste linéaire sur le nombre d'individus, l'algorithme de construction d'une pyramide est de complexité polynomiale d'ordre au moins trois alors que son analogue non recouvrant est au plus quadratique.

Le travail que nous proposons dans cet article vise à répondre aux problématiques spécifiques évoquées ci-dessus par l'utilisation du concept des cartes auto-organisatrices [KOH 84], adapté à la problématique de la classification recouvrante. En effet, la structure classificatoire proposée par ce type de carte permettra :

– de ne plus décider à l'avance du nombre de classes souhaité mais de laisser le processus organiser les individus sur une carte dont le nombre de neuronnes est généralement très supérieur au nombre de classes finales ;
– d'assurer de manière simple la cohérence topologique des recouvrements par le biais d'une affectation des individus à plusieurs neurones vérifiant une structure particulière de sous-graphe sur la carte (e.g. cliques) ;
– une pré-organisation des individus dans un formalisme facilitant un processus ultérieur de classification hiérarchique recouvrante, et ainsi de réduire le coût de traitement de ce type de modèles recouvrants.

Par ailleurs, les recouvrements permettant à chaque individu d'être associé à plusieurs neuronnes sur la carte, l'existance ou non de chevauchements entre neuronnes peut être facilement visualisé et apporte une information capitale sur l'interprétation visuelle (et éventuellement la ségmentation) de la carte topologique résultante.

2. Le modèle de carte auto-organisatrice recouvrante

2.1. *Les modèles de base :* SOM *et* OKM

Les cartes auto-organisatrices consistent à représenter un ensemble d'individus $\mathcal{X} = \{x_i\}_i$ par un nombre réduit de représentants $\{m_c\}_c$ (appelés neuronnes) organisés sur une carte topologique (génralement une grille 2D) telle que deux représentants proches seront voisins sur la carte. L'algorithme SOM [KOH 84] permet de construire une carte topologique en itérant un processus de correction de la carte qui consiste, à l'itération t, à

1. tirer aléatoirement un individu $x^t \in \mathcal{X}$,
2. repérer le neuronne vainqueur m_v le plus proches selon une métrique d pré-établie :

$$m_v = \arg\min_{m_c} d(x^t, m_c)$$

3. adapter le neuronne m_v ainsi que ses voisins sur la carte de manière à intégrer l'individu x^t dans leur définition

$$\forall c, \quad m_c^{(t+1)} = m_c^{(t)} + \alpha(t).N(c,v).\left(x^t - m_c^{(t)}\right) \tag{1}$$

Dans la mise à jour des neuronnes (1) le terme $\alpha(t)$ désigne le coefficient d'apprentissage qui décroît à mesure des itérations[2] et $N(c,v)$ est une fonction de voisinage de valeur 1 lorsque $c = v$ et diminuant à mesure que le neuronne m_c est éloigné du neuronne vainqueur m_v sur la carte.

Du point de vue de la classification recouvrante, [CLE 08] a proposé l'algorithme OKM (*Overlapping k-means*) comme une variante généralisant l'algorithme bien connu des k-moyennes. La généralisation porte à la fois sur le critère objectif guidant la recherche et sur le processus algorithmique lui-même.

2. $\alpha(t) \to 0$ ce qui permet d'assurer la convergence de l'algorithme.

– le critère objectif évalue le décalage (ou inertie) des individus (x_i) par rapport à leur représentant ($\phi(x_i)$) dans la classification, il s'exprime ainsi

$$\mathcal{J}(\{\pi_c\}_{c=1}^k) = \sum_{x_i \in \mathcal{X}} \|x_i - \phi(x_i)\|^2 \quad (2)$$

Dans (2), $\{\pi_c\}_{c=1}^k$ désigne les k classes en construction (k fixé) et derrière les représentants $\{\phi(x_i)\}_i$ se cachent les paramètres du modèle à savoir : les affectations et les centres de classes[3] $\{m_c\}_c$. On peut alors montrer qu'en limitant chaque individu à n'appartenir qu'à une seule classe, le critère (2) se réécrit comme le critère des moindres carrés utilisé dans k-moyennes (d'où la généralisation).

– l'algorithme OKM de minimisation du critère (2) diffère de son équivalent non recouvrant dans les deux étapes itérées : un individu x_i est affecté à autant de classes que nécessaire pour améliorer d'autant son représentant $\phi(x_i)$ et la mise à jour des centres de classes $\{m_c\}_c$ tient compte des affectations multiples. Une nouvelle fois, le fait de contraindre chaque individu à n'être affecté qu'à une seule classe reviendrait à exécuter l'algorithme initial k-moyennes.

2.2. *Extension de* SOM *pour la classification recouvrante*

Nous proposons une extension des cartes auto-organisatrices au domaine de la classification recouvrante ; pour cela nous nous inspirons à la fois du modèle initial SOM et de la formalisation des recouvrements introduite dans OKM *via* la notion de représentant défini comme combinaison de centres m_c. Dans ce nouveau modèle, la présentation d'un individu x_i dans le réseau de neurones (formalisé par $\{m_c\}_c$) conduit à rechercher non plus "le" neurone vainqueurs mais un ensemble $M(i)$ de neurones vainqueurs tel que $M(i)$ correspond à un sous-graphe particulier sur la carte (e.g. composante connexe ou clique) permettant de satisfaire la contrainte de cohérence évoquée plus haut.

Définition 2.1 *Soit x_i un individu de $\mathbb{R}^p$ et $M(i) \subset \{m_c\}_c$ l'ensemble des neuronnes vainqueurs, le représentant de x_i (noté $\phi(x_i)$) est défini par l'application $\phi : \mathbb{R}^p \to \mathbb{R}^p$ suivante*

$$\phi(x_i) = \frac{\sum_{m_c \in M(i)} m_c}{|M(i)|}$$

Dans ce cadre, nous proposons un processus de construction de la carte de kohonen recouvrante qui à chaque itération t consiste à :

1. tirer aléatoirement un individu $x^t \in \mathcal{X}$,

2. initialiser la recherche des neurones vainqueurs $M(t)$ par un premier neuronne m_v^1 le plus proches selon une métrique d pré-établie :

$$m_v^1 = \arg\min_{m_c} d(x^t, m_c)$$

3. étendre $M(t)$ par une heuristique de recherche naïve[4] de manière à minimiser $d(x^t, \phi(x^t))$,

4. adapter l'ensemble des neuronnes vainqueurs dans $M(t)$ ainsi que leurs voisins sur la carte de manière à intégrer l'individu x^t dans leur définition

$$\forall c, \quad m_c^{(t+1)} = m_c^{(t)} + \alpha(t).N(c, M(t)).\left(x^t - m_c^{(t)}\right) \quad (3)$$

Dans la dernière expression (3) de mise à jour des neurones, la fonction de voisinage $N(.,.)$ dépend cette fois de la longueur du chemin entre le neurone m_c et le sous-graphe $M(t)$ (notée $\delta(c, M(t))$) sur la carte. Dans l'expérience à venir nous proposerons les formes suivantes pour $\alpha(.)$ et $N(.,.)$:

$$\alpha(t) = \frac{a}{b+t} \text{ avec a et b constantes} \qquad N(c, M(t)) = exp\left(\frac{-\delta(c, M(t))}{2.\sigma^2(t)}\right)$$

3. Le représentant $\phi(x_i)$ est défini par combinaison (e.g. centre de gravité) des centres des classes auxquelles x_i appartient.
4. Dans le cas d'une recherche de cliques on pourra par exemple commencer par rechercher si l'affectation supplémentaire de x^t à un voisin de m_v^1 sur la carte améliore $\phi(x_i)$ et ainsi de suite.

2.3. *Illustration sur un exemple*

Une expérimentation préliminaire sur la base Iris (UCI repository) nous permet un premier aperçu de l'intérêt des recouvrements pour l'exploitation visuelle de la carte.

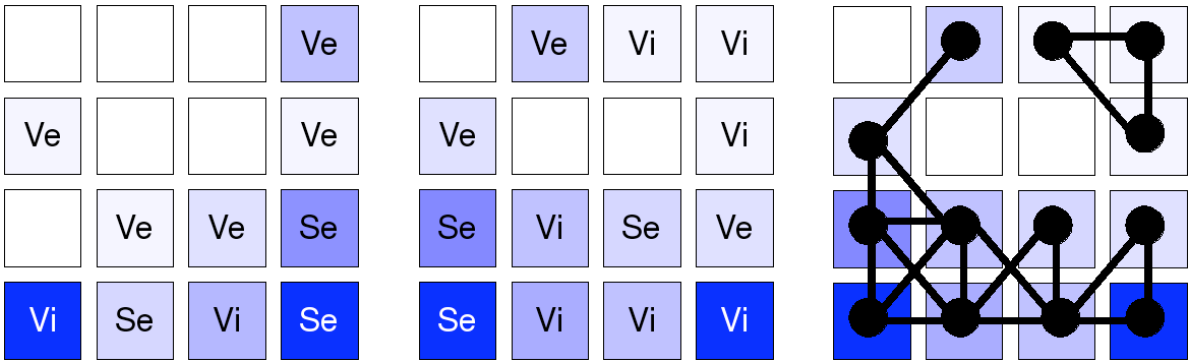


FIG. 1. *Cartes de Kohonen (4×4) obtenues sur Iris ; de gauche à droite :* SOM, SOM *avec recouvrement, visualisation des recouvrements.*

Dans les cartes présentées en Figure 1, l'intensité de chaque case modélise la quantité d'individus affectés au neurone associé[5] et les étiquettes Se, Ve et Vi illustrent le label de classe majoritaire (Setosa, Versicolour et Virginica respectivement) dans le neurone. On se contentera sur cet exemple de faire observer l'aide apportée par le graphe des intersection dans la lecture de la carte, notamment pour dissocier plus facilement les cellules étiquetées Vi des cellules Ve dans la partie supérieure droit.

3. Conclusion

Nous avons exposé dans ce travail une approche de classification recouvrante utilisant le formalisme des cartes auto-organisatrices. Nous avons justifié ce travail par une double profitabilité : d'une part les solutions apportées par les cartes de Kohonen pour résoudre certaines difficultés en classification recouvrante (choix du nombre de classes, cohérence topologique des recouvrements entre autres) et d'autre part l'information supplémentaire que peuvent apporter les recouvrements pour l'exploitation visuelle des cartes topologiques.

Les bases du modèle étant posées, il s'agira dans l'avenir de confirmer les intuitions énoncées en analysant plus finement le comportement des modèles auto-adaptatifs recouvrants sur des jeux de données adaptés. D'autres travaux moins expérimentaux pourront venir compléter l'étude, notamment l'intégration de pondérations locales jugées particulièrement intéressantes pour la classification recouvrante ou encore la comparaison de ces cartes recouvrantes avec les treillis supports des hiérarchies recouvrantes (e.g. pyramides 2D, 3D, etc.).

4. Bibliographie

[BAN 05] BANERJEE A., KRUMPELMAN C., GHOSH J., BASU S., MOONEY R. J., Model-based overlapping clustering, *KDD '05 : Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, New York, NY, USA, 2005, ACM Press, p. 532–537.

[BEZ 81] BEZDEK J. C., Pattern Recognition with Fuzzy Objective Function Algorithms, *Plenum Press, New York*, , 1981.

[CLE 08] CLEUZIOU G., An extended version of the k-means method for overlapping clustering, *ICPR*, IEEE, 2008, p. 1-4.

[DAT 68] DATTOLA R., A fast algorithm for automatic classification, rapport, 1968, Report ISR-14 to the National Science Foundation, Section V, Cornell University, Department of Computer Science.

[DID 84] DIDAY E., Une représentation visuelle des classes empiétantes : Les Pyramides, rapport, 1984, INRIA num.291, Rocquencourt 78150, France.

[KOH 84] KOHONEN T., Self-Organization and Associative Memory, *Springer*, , 1984.

5. Case claire pour un neurone vide, case foncée pour un neurone avec plusieurs dizaines d'individus

Kohonen Approach for Assisted Living Services Construction

T.B.T. Truong, F. Saïd-Hocine*, F. Frizon de Lamotte, J-Ph. Diguet

*Lab-STICC, LMAM**
UBS-CNRS
56100 Lorient, France

ABSTRACT. *The social cost for the increasing older population is motivating new research, designed to support the ability and to favorite the autonomy of the disabled as well as the elderly. This paper describes an original proactive application of a Kohonen classification for home monitoring. Our approach attempts not to use any additional sensor and is only based on the observation of existing home automation events as non vectorial data. An analysis of user habits through the dissimilarity of services with a Kohonen algorithm is proposed for offering automatic scenarii in adequacy with the user capabilities. The simulation results, issued from real data from previous experiences, show the interest of our algorithm.*

KEYWORDS: *Clustering, assisted living service, dissimilarity, DSOM*

1. Introduction

According to the french poll "HID Handicap : Incapacités-Dépendances" in 2001, 13,7% of the french people suffers from disabilities. Furthermore, the life expectancy for both men and women rose during the last century and continues increasing, leading to more elderly people in society. A french official report [1] forecasts that the ratio of people being 60 years old and more will be more than 30% in 2040. As the problem of dependency is strongly related to population age, these ratios clearly show how challenging the question of autonomy is. Social cost for older population is motivating new research, designed to support the ability and the autonomy of the disabled as well as the elderly. Taking in this context, our work is following our previous project realized at Kerpape Center, called QUATRA [F.D 08, ker], and includes two steps : i) providing the user with new services, based on the analysis of his habits, in order to help him in his everyday tasks ; ii) providing a low-level and non-intrusive personal supervision based on the above analysis in order to detect the anomalies and rise the warning. Having presented the first step, the rest of this paper is organized as follows : section 2 presents the data and the dissimilarity measure used to compare them. Section 3 presents a self organizing map algorithm for learning user's habits and results are shown in section 4. Finally, we draw up some conclusions in section 5.

2. Data and dissimilarity measure

A lot of disabled or elderly people meet difficulties to monitor multimedia and home automation services they are supplied with. One of the ways to help them is to group similar services in personalized scenarii with automatic activation. This needs knowledge of the users' habits, ability to measure similarity (or dissimilarity) between services and a clustering procedure.

We observe a user's behaviors, related to multimedia and home automation services, during n days. The collected data consist in a set of activities assorted with their occurrence times. Given two services i and j, we define the pairwise frequency f_{ij} as

$$f_{ij} = \frac{n_{ij} + n_{ji}}{2n} \quad (1)$$

1. Cours des comptes, 2005

where n_{ij} is the number of successive observations of i and j during the observation period. f_{ij} measures the observed frequency of the pair of services $\{i, j\}$ independently of their occurrence order.

A high frequency f_{ij} indicates a high occurrence of the pair $\{i, j\}$ and so a high similarity between the two services. Needing to take into account time in scenario construction, we measure the dissimilarity between services i and j with :

$$d(i,j) = \begin{cases} \frac{\overline{\Delta_{ij}}}{\overline{\Delta_i}+\overline{\Delta_j}}.(1-f_{ij}) & \text{if} \quad i \neq j \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where $\overline{\Delta_{ij}}$ is the mean time lapse between services i and j and $\overline{\Delta_i}$ is the mean time lapse between service i and other existing services. With this dissimilarity measure, two time close services are more similar than two services of same pairwise frequency occurring in longer time lapse.

3. Clustering with DSOM

The Kohonen's SOM algorithm ([T.K 01]) is a nonlinear projection technique for visualizing the underlying structure of high dimensional vectorial data. Input vectors are represented by prototypes arranged along a regular low dimensional map (usually 1D-2D) in such a way that similar vectors in input space become spatially close in the map (or grid of neurons). The SOM algorithm proceeds iteratively in two steps : First, a quantization phase that allows to represent the input dataset by a finite set of prototypes P. Next, the prototypes are organized in the grid by minimizing the quantization error :

$$E(P) = \sum_r \sum_{x \in V_r} \sum_s h(r,s) d(x, p_s) \tag{3}$$

where p_s denotes the prototype associated to neuron s in the input space and V_s the corresponding Voronoi region. d is the Euclidian distance in the input space and h is a topological neighborhood function that performs as a smoothing kernel over the grid.

An adaptation of Kohonen's SOM to dissimilarity data, e.g. data that are only described by pairwise comparisons of observations, was proposed in [T.K 98]. The main difference with the original SOM lies in the quantization phase and in the quantization error expression where d holds for dissimilarity measure rather than Euclidian distance. Before running the DSOM algorithm, we performed an hierarchical clustering of the data using complete linkage ; that helped us to choose the grid size. The DSOM algorithm we implemented proceeds as follows :

- *Initialization step : We assign a prototype to each neuron of the grid by sampling over the clusters obtained by hierarchical clustering.*
- *Learning step : affectation and representation phases are iterated until a stable clustering is reached. At iteration l :*
  - *Affectation phase : Each service x is assigned to its closest prototype p_k^{l-1} : $x \in V_k^{l-1}$.*
  - *Representation phase : The new prototype p_j^l of the neuron j is solution of the minimization problem :*

$$p_j^l = \arg\min_{y \in V_j^{l-1}} \sum_i \sum_{x \in V_i^{l-1}} h^l(i,j) d(x,y) \tag{4}$$

where h^l is a gaussian kernel which varies during the learning.

4. Results

During the Quatra project and OT advise at Kerpape center, real life observations concerning disabled people living in Kerpape center were collected. These observations included their wished services and their use of supplied assisted devices such as "Turn on light", "open door", "turn on TV".... The offered services are given in table1.

Our purpose is now to save users' efforts and time by learning their habits and by proposing personalized scenarii. Learning a user's habits requires to observe him for a long period. As we presently have not such information, we selected a user and basing on his daily activity schedule for a given day, we simulated a user's profile by generating 100 daily activity schedules. The offered services are supposed to respect given laws (Gaussian distribution, uniform distribution...) and supposed to be randomly affected by rare events (sickness, delay...).

Label	Occurence time	Daily service	Label	Occurence time	Daily service
E1	08 :00 :00	Switch on light	E13	15 :00 :00	Turn on computer
E2	08 :05 :00	Open shutter	E14	19 :00 :00	Switch on light
E3	08 :10 :00	Turn on TV	E15	20 :00 :00	Turn on TV
E4	08 :15 :00	Turn on hot water*	E16	20 :30 :00	Watch DVD
E5	08 :30 :00	Unlock door	E17	21 :00 :00	Turn on out light
E6	08 :45 :00	Turn off TV	E18	21 :15 :00	Hang on telephone
E7	08 :55 :00	Open door	E19	21 :30 :00	Hang up telephone
E8	09 :00 :00	Switch off light	E20	21 :50 :00	Close shutter
E9	09 :05 :00	Close door	E21	22 :00 :00	Locate bed
E10	13 :00 :00	Open door	E22	22 :15 :00	Turn off TV
E11	13 :10 :00	Close door	E23	22 :30 :00	Switch off out light
E12	13 :25 :00	Locate bed	E24	22 :40 :00	Switch off in light

* wished service, but not yet available

Table 1. *User's schedule*

Hierarchical clustering using complete linkage showed a reasonable clustering of the daily services into 8 classes. The grid size in the DSOM algorithm was consequently chosen to be 3×3. We performed DSOM clustering with 50 different initializations and they all needed about 20 iterations to reach local minima of the quantization error. We extracted stable subsets from the resulting clusters ; they are listed in table 2 and they correspond to strong scenarii.

Scenarii	Frequency (%)
[E7,E8,E9]	94
[E10,E11]	92
[E21,E22,E23,E24]	86
[E12,E13]	86
[E1,E2,E3,E4,E5]	76
[E16,E17,E18]	64
[E19,E20]	64
[E14,E15]	62

Table 2. *Stable scenarii*

The user is proposed 8 scenarii : a "Wake up" scenario ([E1,E2,E3,E4,E5]) consisting in a set of services the user is used to do when getting up in the morning : turn on the light, open shutter, turn on TV, turn on hot water, unlock the door ; a "Go out" scenario [E7,E8,E9] in the morning ; a "Go in" scenario [E10,E11] in the afternoon and so on.

The high correlation between DSOM scenarii and real ones in Quatra project proves the effectiveness of our approach.

The execution time was about 15 seconds on a machine having a 3.2 GHz Pentium IV processor and 2GB RAM. However, much less than 50 initializations are needed to extract strong scenarii from the data.

5. Conclusions

The automation procedure we propose in this paper consists in two steps :

1. learning step : scenarii construction by DSOM,

2. automation step : when a service is activated, the following services in the corresponding scenario are automatically proposed at suitable times and activated if confirmed by the user.

The clusters number and the DSOM initialization were determined following an hierarchical clustering. About the number of clusters, we intend to try other approaches using distances evaluation between clusterings. In what concerns DSOM initialization, several strategies such as the one proposed in ([A.F 08]) are to be tested.

The automation step is also to be improved, especially concerning the time at which a service is proposed.

6. References

[A.F 08] A.FIANNACA, R.RIZZO, A.URSO, S.GAGLIO, "A new SOM Initialization algorithm for nonvectorial data", *KES*, vol. 5177, 2008, p. 41–48.

[F.D 08] F.DE LAMOTTE AND J-P.DEPARTE AND F.LE SAOUT AND J-PH.DIGUET AND J-L.PHILIPPE, "Quadra : Rapport Final", report , juin 2008, Lab-STICC CNRS / UBS, Lorient, France.

[ker] "Kerpape mutualistic functional reeducation and rehabilitation center", http ://www.kerpape.mutualite56.fr/page_uk/index_uk.htm.

[T.K 98] T.KOHONEN, P.J.SOMERVUO, "Self-organizing maps of symbol strings", *Neurocomputing*, vol. 21, 1998, p. 19–30.

[T.K 01] T.KOHONEN, *Self Organizing Maps*, Springer, 3 edition, 2001.

Auto-organisation d'une structure neuronale arborescente

Chebira Abdennasser, Budnyk Ivan, Madani Kurosh

LISSI,
Université de PARIS-EST / PARIS 12 - Val de Marne,
IUT de Sénart-Fontainebleau
Avenue Pierre Point
77127 Lieusaint, France
{chebira, budnyk, madani}@univ-paris12.fr

RÉSUMÉ. Nous allons présenter un outil de construction par apprentissage de structures neuronales arborescentes ouvrant de nouveaux horizons dans la résolution des problèmes de classification réputés difficiles. La structure logicielle de cet outil appelé T-DTS (Tree Divide To Simplify) est basée sur le paradigme diviser pour régner. Deux originalités caractérisent cet outil : la première est que le processus de décomposition est guidé par les données ; quant à la seconde, elle puise son origine dans l'agilité de la structure à s'auto-organiser afin d'optimiser les ressources engagées dans la solution construite. Une analyse du problème permet de quantifier sa complexité. Celle ci est par la suite exploitée pour guider le processus d'auto-organisation de l'arbre neuronal. Ce processus utilise des critères tels que la maximisation ou la minimisation des taux d'apprentissage et de généralisation, le nombre de nœuds et de feuilles de l'arbre et le temps d'exécution. Nous présentons une méthode d'estimation de complexité s'appuyant sur le processeur neuronal massivement parallèle. Plus particulièrement, nous allons développer la procédure d'auto-organisation qui permet de trouver les seuils de décision, qui est basée sur l'estimation de la distribution de la complexité au niveau des feuilles d'un arbre issu d'une décomposition totale.

MOTS-CLÉS : Classification, Auto-organisation, Structures Arborescentes, Réseaux de Neurones Artificiels, Complexité.

1. Introduction

Nous présentons dans cet article, une méthode d'optimisation d'une structure neuronale arborescente que nous avons développée et mise au point. Cette structure de traitement est particulièrement adaptée pour résoudre des problèmes de classification difficiles et de taille réelle [CHE 09]. L'approche que nous proposons est de type "diviser pour régner" et se démarque des approches similaires par le fait que le processus d'auto-organisation de l'arbre neuronal s'appuie sur des techniques d'estimation de complexité. Les problèmes abordés étant de type classification, nous cherchons à maximiser ou minimiser les taux de généralisation, d'apprentissage, les temps d'exécution etc. Dans la section deux, nous allons présenter brièvement l'approche T-DTS (Tree Divide To Simplify) et on présentera par la suite la structure logicielle ouverte que nous avons développée et qui permet de trouver la structure arborescente la mieux adaptée à un problème de classification donné. Dans la section trois, afin de valider notre approche, nous allons présenter deux expérimentations : la première de type académique, la deuxième est un benchmark utilisé par la communauté à des fins de comparaison de résultats. Nous analyserons les résultats obtenus dans le cas du traitement du problème des deux spirales et dans le cas d'un problème de classification de séquence d'ADN. Nous finirons cet article par une conclusion et des perspectives.

2. Structure neuronale arborescente

Nous nous basons sur le paradigme "diviser pour régner" pour construire une arborescence neuronale de traitement de données. Le traitement est principalement de type classification supervisée. L'approche que nous proposons ne se limite pas uniquement à ce type de problèmes. A partir d'un ensemble de données d'apprentissage, décrivant un problème de classification particulier, une structure neuronale arborescente est

construite dynamiquement et sans l'intervention de l'utilisateur. Le processus de construction est entièrement guidé par les données. La figure 1 représente la structure logicielle de T-DTS, [CHE 09]. L'unité de contrôle va construire un arbre neuronal, dont les nœuds sont constitués d'unités de décomposition et les feuilles d'unités de traitement.

Cette unité de contrôle va exploiter un ensemble d'estimateurs de complexité afin d'adapter la structure arborescente produite par T-DTS. Au niveau des unités de décomposition, nous mettons en œuvre des algorithmes de segmentation : Cartes auto-organisatrices de Kohonen, Réseaux de neurones à compétition, Learning Vector Quantization. Au niveau des feuilles, nous exploitons des techniques d'apprentissage neuronales : Perceptrons Multicouches, réseaux de neurones à régression (GRNN), réseaux de neurones à base radiale (RBF), réseaux de neurones probabilistes (PNN) et les réseaux de neurones linéaires [TRE 01]. Les techniques d'estimation de complexité sont basées sur les mesures de Mahalanobis, Bhattacharyya, Divergence de Kullback-Leibler, PRISM (Entropie collective, pureté et k plus proches voisins), Jeffries-Matusita, Fisher ou écart type maximal [FIS 23], [SIN 03], [TAK 87]. En particulier nous avons développé un estimateur de complexité obtenu au moyen d'un réseau de neurones artificiel implanté sur une carte électronique massivement parallèle le ZISC®-036 d'IBM© [BUD 08].

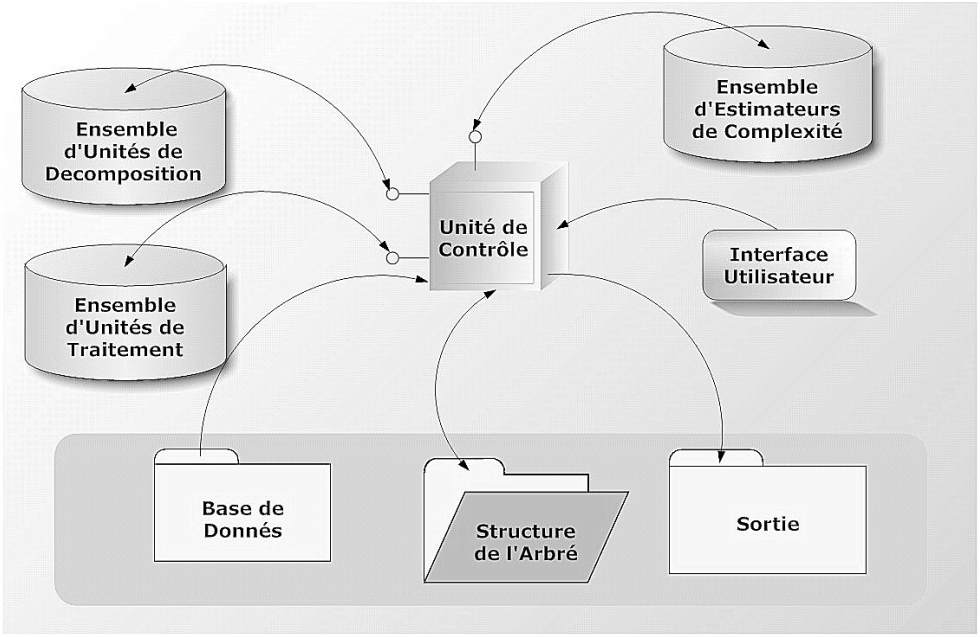


***FIGURE 1** – Architecture logiciel de l'outil T-DTS*

Cette structure logicielle est très ouverte dans le sens où nous disposons de plusieurs techniques de décomposition, d'estimation de complexité et d'apprentissage. Une difficulté supplémentaire réside dans le choix d'un seuil utilisé dans la procédure de décomposition : une technique d'estimation de complexité fourni une mesure dans l'intervalle [**0**: problème très complexe, **1** très simple]. L'unité de control intègre un module de décision. Ce dernier exploite la mesure de complexité pour décider si le problème doit être décomposé en sous-problèmes ou pas. La notion de complexité dépend du contexte, de la nature des données manipulées, du type de problèmes traités et même de la définition même du mot complexité. Cette notion est donc de nature subjective. D'où la difficulté d'interpréter la mesure de complexité, valeur comprise entre **0** et **1**, et de lui associé une sémantique, c'est-à-dire simple, complexe ou très complexe par exemple. Nous avons développé un algorithme d'auto-organisation multicritères capable de produire une solution, c'est-à-dire de construire un arbre neuronal dont la structure est adaptée à un problème de classification particulier. Cet algorithme permet d'associer une sémantique à la mesure de complexité et de trouver dans les ensembles d'unités de traitement, d'unités de décomposition et d'estimateurs de complexité les techniques les plus adaptées au problème en cours de traitement. Les critères que nous exploitons pour apporter une solution au problème de l'adéquation

structure/problème sont les suivants : maximisation du taux d'apprentissage, maximisation du taux de généralisation, minimisation du temps de traitement. Pour des fins de validation de notre approche, nous allons analyser les résultats que nous obtenons dans le cas du traitement d'un problème académique et dans le cas du traitement d'un problème de traitement de séquences d'ADN.

3. Deux spirales et classification de séquences d'ADN

Pour valider notre approche, nous avons traité deux problèmes de classification. Le premier est de type académique: deux classes séparées par une frontière en forme de quatre spirales. Nous maîtrisons l'ensemble des paramètres paramètre dans ce cas : nombre et répartition statistique des échantillons, bruit, données manquantes etc. L'analyse des résultats dans ce cas est plus facilitée. Le deuxième consiste à classer des séquences d'ADN est utilisée par la communauté comme benchmark à des fins de comparaison de performances. Dans le premier cas, la base de données est constituée de 500 prototypes pour l'apprentissage et autant pour la généralisation. Les résultats obtenus sont représentés dans le tableau 1.

TABLEAU 1 *– Résultats obtenus pour le cas des deux spirales*

Exp.	UD	EC	UT	G & ET	A & ET	Nbr feuilles	*Sopt*
1	CNN	Ent	Elman's BN	79.1 +/- 0.4	96.4+/-0.1	104	0,27
2	SOM	Ent	Elman's BN	77.7 +/- 1.6	97.6 +/- 0.2	144	0,33
3	CNN	Fisher	PNN	80.4 +/- 0.4	95.8 +/- 0.2	176	0,72

La première colonne du tableau 1, représente l'indice d'une expérimentation. Les colonnes suivantes l'algorithme de décomposition (UD) et d'estimation de complexité (EC) utilisés au niveau des nœuds de l'arbre, la technique d'apprentissage utilisée au niveau des feuilles (UT), les taux de généralisation et d'apprentissage (G et A) avec l'écart type (ET) associé, le nombre de feuilles et la valeur du seuil de décision relatif à la mesure de complexité (Sopt). La signification des abréviations est la suivante : Competitive Neural Network (CNN), cartes de Cohonen (SOM), mesure d'entropie (Ent), Fisher mesure de Fisher, BN Back propagation Network, PNN Probabilistic Neural Network.

Pour la première expérimentation (Exp. 1), le critère ciblé est le temps de traitement. Dans ce cas, l'arbre généré a une structure plus simple (nombre de feuilles plus petit) et des taux de généralisation et d'apprentissage de l'ordre de 79 et 96 %. Dans la deuxième expérimentation (Exp. 2), le critère ciblé est le taux d'apprentissage. On peut remarquer que dans ce cas, nous obtenons le taux le plus élevé avec une structure de l'arborescence plus complexe. Finalement pour la troisième expérimentation (Exp. 3), le critère ciblé est le taux de généralisation. Dans ce cas, la structure est encore plus complexe, le temps de traitement est plus important et le taux d'apprentissage est plus faible.

Pour le cas du traitement d'un problème de classification de séquence d'ADN, les données pour l'expérimentation sont prises de la base de données du centre de systèmes intelligents et apprentissage artificiel de l'université d'Irvin, Californie *(*ftp://genbank.bio.net*)*. Les principales caractéristiques de cette base de données sont les suivantes : nombre de prototypes 3190, nombre d'attributs 62, pas de données manquantes, trois classes avec une répartition de 25%, 25% et 50%. Nous avons exploité 50% de cette base de données pour l'apprentissage et 50% pour la généralisation. Les résultats obtenus sont présentés dans le tableau 2.

TABLE 2 *– Classification de séquences d'ADN*

Exp.	UD	EC	UT	G & ET	A & ET	Nbre Feuilles	*Sopt*
1	aucun	Aucun	Elman's BN	94,6+/-0.04	99,9+/-0.01	aucun	Aucun
2	CNN	Pureté	Elman's BN	93.5+/-0.4	99.8+/-0.02	2	0,003
3	CNN	Pureté	Elman's BN	93.3+/-0.4	99.8+/-0.05	4	0,03

Dans ce cas nous obtenons un résultat qui parait étonnant: une structure de traitement classique (Exp 1) sans aucune décomposition permet d'obtenir les meilleurs taux d'apprentissage, de généralisation et le temps de traitement le plus court. Si nous forçons la décomposition et fixons comme critère pour le processus d'auto-

organisation le taux d'apprentissage (Exp 2) ou encore le taux de généralisation (Exp 3), le temps de traitement augmente et le taux de généralisation se dégrade légèrement. Le taux d'apprentissage reste quasiment inchangé. Il est à noter que nous obtenons de meilleurs résultats que ceux obtenus par Makal & all [MAK 08], arrivant à un taux de généralisation de l'ordre de 91%. Des résultats de même ordre de grandeur sont obtenus par Malousi & all en exploitant des SVM [MAL 08].

4. Conclusion

Nous avons présenté dans cet article, une méthode de construction d'arborescence neuronale. La souplesse offerte par cette technique, c'est-à-dire, la possibilité d'exploiter différents algorithmes au niveau des nœuds et des feuilles de cet arbre nous amène à faire face à un problème d'optimisation. En effet, nous devons trouver la structure qui optimise un ou plusieurs critères. Les premières expérimentations que nous avons menées dans le cas d'un problème de classification académique et dans le cas d'un problème de classification de séquences d'ADN donnent des résultats très encourageants. En fonction de critères choisis, nous obtenons des performances très élevées. Ce travail sera complété en approfondissant l'analyse des résultats obtenus, en traitant d'autres types de problèmes de classification et en comparant nos résultats à ceux obtenus par d'autres approches.

5. Bibliographie

[BUD 08] BUDNYK I., BOUYOUCEF E. K., CHEBIRA K., MADANI K., "Neurocomputer based complexity estimator optimizing a hybrid multi neural network structure", *International Scientific Journal of Computing (ISJC)*, vol. 7, n° 3, 2008, p. 122-129.

[CHE 09] CHEBIRA K., MELLOUK A., MADANI K., "Neural machine learning approaches: *q*-learning and complexity estimation based information processing system", *Machine Learning*, (Chebira A., Mellouk A., eds.), 2009, ISBN: 9783902613561, p. 1-38, I-Tech Education and Publishing.

[FIS 23] FISHER A., *The mathematical theory of probabilities application to frequency curves and statistical methods*, 1923, ASIN: B0008CQD9M, John Wiley, New York.

[LUM 06] LUMINI A., NANNI L., "Identifying splice-junction sequences by hierarchical multiclassifier", Pattern Recognition Letters archive, vol. 27, n° 12, 2006, p. 1390-1396.

[MAK 08] MAKAL S., OZYILMAZ L., PALAVAROGLU S., "Neural network based determination of spliced junctionsby ROC Analysis", *Proceedings of World Academy of Science, Engeneering and Technology,* vol. *33,* 2008, ISSN: 20703740.

[MAL 08] MALOUSI A., CHOUVARDA I., KOUTKIAS V., , KOUIDOU S., MAGLAVERAS N., "Variable-length positional modeling for biological sequence classification", *American Medical Informatics Association Annual Symposium Proceedings,* 2008, p. 91–95.

[SEN 08] SENE M., CHEBIRA A, MADANI K., "A hybrid multi-experts approach for mechanical defects' detection and diagnosis", *Proceedings of the 7th International conference on Computer Information Systems and Industrial Management (IEEE-CISIM'08)*, ISBN: 9780769531847, 2008, p. 59-64, Prague.

[SIN 03] SINGH S., "Multiresolution estimates of classification complexity", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, n° 12, 2008, p. 1534 – 1539.

[TAK 87] TAKESHITA T., KIMURA K, MIYAKE Y., On the estimation error of Mahalanobis distance, *The Transactions of the Institute of Electronics, Information and Communication Engineers (Academic Journal),* J70-D/3, 1987, p.567-573.

[TRE 01] TRESP V., "Committee Machine", *Handbook for Neural Network Signal Processing*, (Hu Y.H., Hwang J.N., eds.), 2001, CRC Press.

A Latent Logistic Model to Uncover Overlapping Clusters in Networks

Pierre Latouche, Etienne Birmelé, Christophe Ambroise

Laboratoire Statistique et Génome
UMR CNRS 8071-INRA 1152-UEVE
La Genopole, Tour Evry 2,
523 place des Terrasses, 91000 Evry
pierre.latouche@genopole.cnrs.fr

RÉSUMÉ. *It is now widely accepted that knowledge can be learnt from networks by clustering their vertices according to connection profiles. Many deterministic and probabilistic methods have been developed. Given a network, almost all them partition the vertices into disjoint clusters. However, recent studies have shown that these methods were too restrictive and that most of the existing networks contained overlapping clusters. To tackle this issue, we present in this paper a latent logistic model, that allows each vertex to belong to multiple clusters, as well as an efficient approximate inference procedure based on global and local variational techniques. We show the results that we obtained on a transcriptional regulatory network of yeast.*

MOTS-CLÉS : *Networks, Clustering methods, Overlapping clusters, Global and local variational approaches.*

1. Introduction

Networks are used in many scientific fields such as biology, social science, and information technology. In this context, a lot of attention has been paid on developing models to learn knowledge from the presence or absence of links between pairs of objects. Both deterministic and probabilistic strategies have been proposed. Among these techniques, random graph models [HAN 07, LAT 08], based on mixture distributions, have recently received a growing interest. In particular, they have been shown capable of characterizing the complex topology of real networks, that is, a majority of vertices with none or very few links and the presence of hubs which make networks locally dense.

A drawback of such methods is that they all partition the vertices into disjoint clusters, while lots of objects in real world applications typically belong to multiple groups or communities. For instance, many genes are known to participate in several functional categories, and actors might belong to several groups of interests. Thus, a graph clustering method should be able to uncover overlapping clusters.

This issue has received growing attention in the last few years, starting with an algorithmic approach based on small complete sub-graphs developed by Palla and al. [PAL 05]. More recent works [AIR 08] proposed a mixed membership approach. In this paper, we present a new mixture model [LAT 09] for generating networks, depending on $(Q+1)^2+Q$ parameters, where Q is the number of components in the mixture. A latent $\{0,1\}$-vector of length Q is assigned to each vertex, drawn from products of Bernoulli distributions whose parameters are not vertex-dependent. Each vertex may then belong to several components, allowing overlapping clusters, and each edge probability depends only on the components of its endpoints.

Please note that due to the limited size of this paper, we did not include the results we obtained on toy data sets. We did not include either the proof that the model is generically identifiable within classes of equivalence. These results can be presented during SFC.

2. Model and Notations

We consider a directed binary random graph $\mathcal{G}$, where V denotes a set of N fixed vertices and $\mathbf{X} = \{X_{ij}, (i,j) \in \mathrm{V}^2\}$ is the set of all the random edges. We assume that $\mathcal{G}$ does not have any self loop, and therefore, the variables X_{ii} will not be taken into account.

For each vertex $i \in \mathrm{V}$, we introduce a latent vector $\mathbf{Z}_i$, of Q independent Boolean variables $Z_{iq} \in \{0,1\}$, drawn from Bernoulli distributions :

$$\mathbf{Z}_i \sim \prod_{q=1}^{Q} \mathcal{B}(Z_{iq};\, \alpha_q) = \prod_{q=1}^{Q} \alpha_q^{Z_{iq}} (1-\alpha_q)^{1-Z_{iq}}, \tag{1}$$

and we denote $\boldsymbol{\alpha} = \{\alpha_1, \ldots, \alpha_Q\}$ the vector of class probabilities. Note that in the case of a usual mixture model, $\mathbf{Z}_i$ would be generated according to a multinational distribution with parameters $(1, \boldsymbol{\alpha})$. Therefore, the vector $\mathbf{Z}_i$ would see all its components set to zero except one such that $Z_{iq} = 1$ if vertex i belongs to class q. The model would then verify $\sum_{q=1}^{Q} Z_{iq} = \sum_{q=1}^{Q} \alpha_q = 1,\ \forall i$. In this paper, we relax these constraints using the product of Bernoulli distributions (1), allowing each vertex to belong to multiple classes. We point out that $\mathbf{Z}_i$ can also have all its components set to zero.

Given two latent vectors $\mathbf{Z}_i$ and $\mathbf{Z}_j$, we assume that the edge X_{ij} is drawn from a Bernoulli distribution :

$$X_{ij} |\, \mathbf{Z}_i, \mathbf{Z}_j \sim \mathcal{B}\big(X_{ij};\, \mathrm{g}(a_{\mathbf{Z}_i,\mathbf{Z}_j})\big) = e^{X_{ij} a_{\mathbf{Z}_i,\mathbf{Z}_j}}\, \mathrm{g}(-a_{\mathbf{Z}_i,\mathbf{Z}_j}),$$

where

$$a_{\mathbf{Z}_i,\mathbf{Z}_j} = \mathbf{Z}_i^{\intercal}\, \mathbf{W}\, \mathbf{Z}_j + \mathbf{Z}_i^{\intercal}\, \mathbf{U} + \mathbf{V}^{\intercal}\, \mathbf{Z}_j + W^*, \tag{2}$$

and $\mathrm{g}(x) = (1+e^{-x})^{-1}$ is the logistic sigmoid function. $\mathbf{W}$ is a $Q \times Q$ matrix whereas $\mathbf{U}$ and $\mathbf{V}$ are Q-dimensional vectors. The first term in (2) describes the interactions between the vertices i and j. If i belongs only to class q and j only to class l, then only one interaction term remains ($\mathbf{Z}_i^{\intercal}\, \mathbf{W}\, \mathbf{Z}_j = W_{ql}$). However, the interactions can become much more complex if one or both of these two vertices belong to multiple classes. Note that the second term in (2) does not depend on $\mathbf{Z}_j$. It models the overall capacity of vertex i to connect to other vertices. By symmetry, the third term represents the global tendency of vertex j to receive and edge. Finally, we use W^* as a bias, to model sparsity.

3. Variational Approximations

The log-likelihood of the observed data set is defined through the marginalization : $p(\mathbf{X} \,|\, \boldsymbol{\alpha}, \tilde{\mathbf{W}}) = \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z} \,|\, \alpha, \tilde{\mathbf{W}})$. This summation involves 2^{NQ} terms and quickly becomes intractable. To tackle this issue, the Expectation-Maximization (EM) algorithm has been applied on many mixture models. However, the E-step requires the calculation of the posterior distribution $p(\mathbf{Z} \,|\, \mathbf{X}, \boldsymbol{\alpha}, \tilde{\mathbf{W}})$ which can not be factorized in the case of networks. In order to obtain a tractable procedure, we propose some approximations based on global and local variational techniques.

3.1. *The q-transformation (Variational EM)*

Given a distribution $q(\mathbf{Z})$, the log-likelihood of the observed data set can be decomposed using the Kullback-Leibler divergence (KL) :

$$\ln p(\mathbf{X} \,|\, \boldsymbol{\alpha}, \tilde{\mathbf{W}}) = \mathcal{L}\Big(q;\, \boldsymbol{\alpha}, \tilde{\mathbf{W}}\Big) + \mathrm{KL}\Big(q(\mathbf{Z}) \,||\, p(\mathbf{Z} \,|\, \mathbf{X}, \boldsymbol{\alpha}, \tilde{\mathbf{W}})\Big). \tag{3}$$

By definition $\mathrm{KL}(. \,||\, .)$ is always positive and therefore $\mathcal{L}$ is a lower bound of the log-likelihood :

$$\ln p(\mathbf{X} \,|\, \boldsymbol{\alpha}, \tilde{\mathbf{W}}) \geq \mathcal{L}\Big(q;\, \boldsymbol{\alpha}, \tilde{\mathbf{W}}\Big)\ , \forall q(\mathbf{Z}). \tag{4}$$

The maximum $\ln p(\mathbf{X} \mid \boldsymbol{\alpha}, \tilde{\mathbf{W}})$ of $\mathcal{L}$ is reached when $q(\mathbf{Z}) = p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\alpha}, \tilde{\mathbf{W}})$. Thus, if the posterior distribution $p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\alpha}, \tilde{\mathbf{W}})$ was tractable, the optimizations of $\mathcal{L}$ and $\ln p(\mathbf{X} \mid \boldsymbol{\alpha}, \tilde{\mathbf{W}})$, with respect to $\boldsymbol{\alpha}$ and $\tilde{\mathbf{W}}$, would be equivalent. However, in the case of networks, $p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\alpha}, \tilde{\mathbf{W}})$ can not be calculated and $\mathcal{L}$ can not be optimized over the entire space of $q(\mathbf{Z})$ distributions. Thus, we restrict our search to the class of distributions which satisfy :

$$q(\mathbf{Z}) = \prod_{i=1}^{N} q(\mathbf{Z}_i) = \prod_{i=1}^{N} \prod_{q=1}^{Q} \mathcal{B}(Z_{iq};\, \tau_{iq}). \tag{5}$$

Each τ_{iq} is a variational parameter and corresponds to the posterior probability of node i to belong to class q. Note that we do not constrain the vectors $\boldsymbol{\tau}_i = \{\tau_{i1}, \ldots, \tau_{iQ}\}$ to lay on the $Q-1$ dimensional simplex, and thereby, each node can belong to multiple clusters.

The decomposition (3) and the factorization (5) lead to a variational EM algorithm. During the E-step, the parameters $\boldsymbol{\alpha}$ and $\tilde{\mathbf{W}}$ are fixed ; and by optimizing the lower bound with respect to the τ_{iq}s, the algorithm looks for the best approximation of the posterior distribution. Then, during the M-step, $q(\mathbf{Z})$ is used to optimize the lower bound and to find new estimates of $\boldsymbol{\alpha}$ and $\tilde{\mathbf{W}}$.

3.2. *The ξ-transformation*

The lower bound of (3) is given by

$$\mathcal{L}\Big(q;\, \boldsymbol{\alpha}, \tilde{\mathbf{W}}\Big) = \mathrm{E}_{\mathbf{Z}}[\ln p(\mathbf{X} \mid \mathbf{Z}, \tilde{\mathbf{W}})] + \mathrm{E}_{\mathbf{Z}}[\ln p(\mathbf{Z} \mid \boldsymbol{\alpha})] - \mathrm{E}_{\mathbf{Z}}[\ln q(\mathbf{Z})], \tag{6}$$

where the expectations are calculated according to the distribution $q(\mathbf{Z})$. The first term of (6) is given by :

$$\mathrm{E}_{\mathbf{Z}}[\ln p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\alpha}, \tilde{\mathbf{W}})] = \sum_{i \neq j}^{N} \Big\{ X_{ij} \tilde{\boldsymbol{\tau}}_i{}^{\intercal} \tilde{\mathbf{W}} \tilde{\boldsymbol{\tau}}_j + \mathrm{E}_{\mathbf{Z}_i, \mathbf{Z}_j}[\ln \mathrm{g}(-a_{\mathbf{Z}_i, \mathbf{Z}_j})] \Big\}. \tag{7}$$

Unfortunately, since the logistic sigmoid function is non linear, $\mathrm{E}_{\mathbf{Z}_i, \mathbf{Z}_j}[\ln \mathrm{g}(-a_{\mathbf{Z}_i, \mathbf{Z}_j})]$ can not be computed analytically. Thus, we need a second level of approximation to carry out the variational E and M steps described previously (Sect. 3.1).

We use the bound $\ln \mathrm{g}(x, \xi)$ on the log-logistic sigmoid function introduced by [JAA 00]. When applied on $\mathrm{E}_{\mathbf{Z}_i, \mathbf{Z}_j}[\ln \mathrm{g}(-a_{\mathbf{Z}_i, \mathbf{Z}_j})]$, it leads to :

$$\mathrm{E}_{\mathbf{Z}_i, \mathbf{Z}_j}[\ln \mathrm{g}(-a_{\mathbf{Z}_i, \mathbf{Z}_j})] \geq \ln \mathrm{g}(-a_{\mathbf{Z}_i, \mathbf{Z}_j}, \xi_{ij}) = \ln \mathrm{g}(\xi_{ij}) - \frac{(\tilde{\boldsymbol{\tau}}_i{}^{\intercal} \tilde{\mathbf{W}} \tilde{\boldsymbol{\tau}}_j + \xi_{ij})}{2} - \lambda(\xi_{ij}) \Big(\mathrm{E}_{\mathbf{Z}_i, \mathbf{Z}_j}[(\tilde{\mathbf{Z}}_i{}^{\intercal} \tilde{\mathbf{W}} \tilde{\mathbf{Z}}_j)^2] - \xi_{ij}^2 \Big), \tag{8}$$

where $\lambda(\xi) = \frac{1}{4\xi} \tanh(\frac{\xi}{2}) = \frac{1}{2\xi} \big\{ \mathrm{g}(\xi) - \frac{1}{2} \big\}$. Thus, for each edge (i, j) in the graph, we have introduced a lower bound which depends on a variational parameter ξ_{ij}. By optimizing each function $\ln \mathrm{g}(-a_{\mathbf{Z}_i, \mathbf{Z}_j}, \xi_{ij})$ with respect to ξ_{ij}, we obtain the tightest bounds to the functions $\mathrm{E}_{\mathbf{Z}_i, \mathbf{Z}_j}[\ln \mathrm{g}(-a_{\mathbf{Z}_i, \mathbf{Z}_j})]$. These bounds are then used during the variational E and M steps to optimize an approximation of $\mathcal{L}$ defined in (6).

4. Experiments

We consider the yeast transcriptional regulatory network described in [MIL 02] and we focus on a subset of 192 vertices connected by 303 edges. Nodes of the network correspond to operons, and two operons are linked if one operon encodes a transcriptional factor that directly regulates the other operon. Such networks are known to be relatively sparse which makes them hard to analyze. In this Section, we aim at clustering the vertices according to their connection profile. Using $Q = 6$ clusters, we apply our algorithm and we obtain the results in Table 1.

cluster	size	operons
1	2	STE12 TEC1
2	33	YBR070C MID2 YEL033W SRD1 TSL1 RTS2 PRM5 YNL051W PST1 YJL142C **SSA4** YGR149W SPO12 YNL159C SFP1 YHR156C YPS1 YPL114W HTB2 MPT5 SRL1 DHH1 **TKL2** PGU1 YHL021C RTA1 WSC2 GAT4 YJL017W TOS11 YLR414C BNI5 YDL222C
3	2	MSN4 MSN2
4	32	CPH1 **TKL2** HSP12 SPS100 MDJ1 GRX1 SSA3 ALD2 GDH3 GRE3 HOR2 ALD3 SOD2 ARA1 HSP42 YNL077W **HSP78** GLK1 DOG2 HXK1 RAS2 **CTT1** HSP26 TPS1 TTR1 HSP104 GLO1 **SSA4** PNC1 MTC2 YGR086C **PGM2**
5	2	YAP1 SKN7
6	19	YMR318C **CTT1** TSA1 CYS3 ZWF1 HSP82 TRX2 GRE2 SOD1 AHP1 YNL134C **HSP78** CCP1 TAL1 DAK1 YDR453C TRR1 LYS20 **PGM2**

TAB. 1. *Classfication of the operons into $Q = 6$ clusters. Operons in bold belong to multiple clusters.*

First, we notice that the clusters 1, 3, and 5 contain only two operons each. These operons correspond to hubs which regulate respectively the nodes of clusters 2, 4, and 6. More precisely, the nodes of cluster 2 are regulated by STE12 and TEC1 which are both involved in the response to glucose limitation, nitrogen limitation and abundant fermentable carbon source. Similarly, MSN4 and MSN2 regulate the nodes of cluster 4 in response to different stress such as freezing, hydrostatic pressure, and heat acclimation. Finally, the nodes of cluster 6 are regulated by YAP1 and SKN7 in the presence of oxygen stimulus. In the case of sparse networks, one of the clusters often contains most of the vertices having weak connection profiles, and is therefore not meaningful. Conversely, with our approach, the vectors $\mathbf{Z}_i$ can have all their components set to zero, corresponding to vertices that do not belong to any cluster. Thus, we obtained 85 unclassified vertices. Our algorithm was able to uncover two overlapping clusters (operons in bold in Table. 1). Thus, SSA4 and TKL2 belong to cluster 2 and 4. Indeed, they are co-regulated by (STE12, TEC1) and (MSN4 and MSN2). Moreover, HSP78, CTT1, and PGM2 belong to cluster 4 and 6 since they are co-regulated by (MSN4, MSN2) and (YAP1, SKN7).

5. Conclusion

In this paper, we proposed a new latent logistic model to uncover overlapping clusters. We used both local and global variational techniques and we derived a variational EM algorithm to optimize the model parameters. We analyzed a transcriptional regulatory network of yeast and we showed that our model was able to handle sparsity. We discovered two overlapping clusters corresponding to co-regulated operons.

6. Bibliographie

[AIR 08] AIROLDI E. M., BLEI D. M., FIENBERG S. E., XING E. P., Mixed membership stochastic blockmodels, *Journal of Machine Learning Research*, , 2008.

[HAN 07] HANDCOCK M., RAFTERY A., TANTRUM J., Model-based clustering for social networks, *Journal of the Royal Statistical Society*, vol. 170, 2007, p. 1-22.

[JAA 00] JAAKKOLA T. S., JORDAN M. I., Bayesian parameter estimation via variational methods, *Statistics and Computing*, vol. 10, 2000, p. 25-37.

[LAT 08] LATOUCHE P., BIRMELÉ E., AMBROISE C., Bayesian Methods for Graph Clustering, rapport, 2008, SSB.

[LAT 09] LATOUCHE P., BIRMELÉ E., AMBROISE C., A latent logistic model to uncover overlapping clusters in networks, rapport, 2009, SSB.

[MIL 02] MILO R., SHEN-ORR S., ITZKOVITZ S., KASHTAN D., CHKLOVSKII D., ALON U., Network motifs :simple building blocks of complex networks, *Science*, vol. 298, 2002, p. 824-827.

[PAL 05] PALLA G., DERENYI I., FARKAS I., VICSEK T., Uncovering the overlapping community structure of complex networks in nature and society, *Nature*, vol. 435, 2005, p. 814-818.

Classification de variables et détermination des variables latentes sous contrainte d'ordre

Cariou Véronique

Unité de Sensométrie et Chimiométrie
ENITIAA
Rue de la Géraudière
44 322 Nantes Cedex 3
veronique.cariou@enitiaa-nantes.fr

RÉSUMÉ. L'intérêt de la classification de variables en tant que stratégie de synthèse et éventuellement de réduction d'un tableau de données a été souligné par plusieurs auteurs (voir par exemple [KET 06]). Parmi les différentes approches, la méthode de classification de variables appelée CLV [VIG 05b], vise à identifier une partition de variables de manière à maximiser l'adéquation de chaque groupe avec une variable latente résumant celui-ci. Dans le cadre de variables munies d'une structure d'ordre, présente naturellement dans des données temporelles ou spectrales, une extension de CLV [CAR 07] a été proposée. Les contraintes d'ordre sont imposées afin d'assurer que les variables de chaque classe correspondent à un même intervalle de l'échelle d'évolution. Pour la résolution de ce problème, une stratégie basée sur la programmation dynamique est adoptée. Nous présentons ici une modification algorithmique conduisant à réduire notablement les temps de calcul des variables latentes associées aux groupes. Cette démarche présente plusieurs avantages en termes d'interprétation des résultats et de temps de calcul associés. Elle est illustrée à l'aide d'une étude visant à caractériser des variétés de petits pois par spectroscopie infrarouge.

MOTS-CLÉS : Classification de variables, contrainte d'ordre, données spectrales.

1. Introduction

Dans de nombreuses applications analytiques, les données mesurées sur un ensemble d'individus se présentent sous forme de courbes (spectres infrarouge, chromatogrammes, ...). Après discrétisation, ces données peuvent être consignées dans une matrice $\boldsymbol{X}$ ayant n lignes (individus) et p colonnes (variables). Par nature, les p variables sont logiquement ordonnées (ordre temporel ou suivant les longueurs d'onde, par exemple). La classification de variables peut dans ce cadre s'avérer intéressante en tant que stratégie de synthèse et de compression des signaux.

Parmi les différentes approches, la méthode de classification de variables appelée CLV [VIG 05b], vise à identifier une partition de variables de manière à maximiser l'adéquation de chaque groupe avec une variable latente résumant celui-ci. Dans le cadre de variables munies d'une structure d'ordre, présente naturellement dans des données temporelles ou spectrales, une extension de CLV [CAR 07] a été proposée. Les contraintes d'ordre sont imposées afin d'assurer que les variables de chaque classe correspondent à un même intervalle de l'échelle d'évolution. Pour la résolution de ce problème, une stratégie basée sur la programmation dynamique est adoptée suivant le même principe que Lechevallier (voir [LEC 90]). L'avantage escompté d'une telle démarche est de permettre :

- une interprétation plus simple des résultats par les utilisateurs qui peuvent visualiser directement sur leurs données les classes de variables obtenues,
- une possibilité de compression des signaux initiaux, telle qu'elle a été également abordée par Douzal (voir [DOU 03]).

L'avantage de l'approche CLV dans le cas de variables ordonnées, réside également dans la possibilité de résumer les données de départ à l'aide des variables latentes associées aux différentes classes. Cependant, le calcul des différentes variables latentes au cours de l'algorithme de programmation dynamique s'avère important dès lors que le nombre de variables augmente. Nous présentons ici une modification algorithmique conduisant à réduire notablement les temps de calcul des variables latentes associées aux groupes. Dans une deuxième partie, nous proposons une application de notre approche à une étude de caractérisation de petits pois à l'aide de données spectrales.

2. Classification de variables sous contrainte de contiguïté

Soit $\boldsymbol{X}$ une matrice contenant p variables $x_1, ..., x_p$ mesurées sur n individus. Les variables sont supposées centrées. Nous notons P_K une partition en K classes $C_1, ..., C_k, ..., C_K$ des variables constituant le tableau $\boldsymbol{X}$. Sous contrainte d'ordre, la méthode CLV consiste à déterminer K composantes (variables latentes) $c_1, ..., c_k,..., c_K$ associées respectivement aux K classes de manière à maximiser le critère :

$$W(P_K) = \frac{1}{n} \sum_{k=1,\cdots K} c'_k X_{i_k j_k} X'_{i_k j_k} c_k \qquad \text{sous la contrainte } c_k' c_k = 1 \tag{2}$$

où $X_{i_k j_k}$ est la matrice dont les colonnes sont formées par l'ensemble des variables de C_k, c'est à dire dont l'indice est compris dans l'intervalle $[i_k,j_k]$, avec $i_k \le j_k$. De plus la contrainte d'ordre entraîne que :

$$\begin{aligned} & i_1 = 1 \\ & i_k = j_{k-1} + 1 \qquad \forall k > 1 \\ & j_K = p \end{aligned} \tag{3}$$

La valeur optimale du critère est :

$$W(P_K) = \sum_{k=1,\cdots K} \lambda_{i_k j_k}$$

où $\lambda_{i_k j_k}$ est la première valeur propre de $V_k = \frac{1}{n} X'_{i_k jk} X_{i_k j_k}$. La variable latente c_k est la première composante principale standardisée du tableau $X_{i_k j_k}$. Le problème d'optimisation ci-dessus tenant compte des contraintes d'ordre est résolu en utilisant un algorithme de programmation dynamique tel que décrit par Lechevallier [LEC 90]. Ceci consiste en une procédure itérative qui se base sur le principe que si $P_{k,l} = (\{x_1, \cdots, x_i\}, C_2, ..., C_k)$ est une partition optimale en k classes de $\{x_1, \cdots, x_p\}$ alors :

$$(C_2, ..., C_k) = P_{k-1,i+1}$$

où $P_{k-1,i+1}$ est la partition optimale en (k-1) classes de l'ensemble $\{x_{i+1}, \cdots, x_p\}$.

Lors de la prise en compte de la contrainte de contiguïté avec l'algorithme de programmation dynamique, il est nécessaire de calculer la plus grande valeur propre λ_{ij} de chaque matrice $V_{ij} = \frac{1}{n} X'_{ij} X_{ij}$ avec i (resp. j) l'indice de la première variable (resp. dernière variable) de la classe. Dans la mesure où le nombre de variables peut être élevé, la détermination des valeurs λ_{ij} au cours des itérations de l'algorithme de programmation dynamique peut s'avérer très longue. En effet, comme le souligne Van Os et Meulman [VOS 04], toute valeur propre λ_{ij} sera énumérée par toutes les classes $\{C_k \mid k=1, \ldots, K\}$ et évaluée à toutes les étapes de la récursion.

Afin de contourner cette difficulté, nous proposons donc d'effectuer l'ensemble des diagonalisations des matrices de covariances des classes contiguës au préalable. Cette opération nécessite de stocker $p(p-1)/2$ valeurs propres correspondant à l'ensemble des matrices contiguës de $\boldsymbol{X}$.

Cependant les temps de calcul associés à la détermination des différentes valeurs propres successives restent prohibitifs dès lors que le nombre de variables augmente. De plus, on observe généralement une forte multi-colinéarité entre les variables. Une seconde modification algorithmique est donc proposée ici. Elle consiste à utiliser un algorithme de puissance itérée pour la détermination de la première valeur propre tirant partie du fait que l'initialisation peut être obtenue par la solution trouvée lors de l'itération précédente de l'algorithme de programmation dynamique. Soit V_{ij} , la matrice de covariances associée à X_{ij} , λ_{ij} (resp. u_{ij}) la plus grande valeur propre de V_{ij} (resp. le premier vecteur propre). Le calcul de la valeur propre de la matrice V_{ij+1} s'effectue par l'algorithme de puissance itérée en prenant comme vecteur initial $u_{ij+1} = \begin{bmatrix} u_{ij} \\ 0 \end{bmatrix}$.

3. Application

Les données qui servent pour l'illustration de la démarche d'analyse ont été présentées et analysées par Naes et Kowalski [NAE 89] puis par Vigneau *et al.* [VIG 05a]. Il s'agit de la caractérisation de 60 variétés de petits pois surgelés par spectroscopie dans la plage spectrale allant de 1100 à 2500 nm. Afin de réduire d'éventuels artéfacts dus aux conditions instrumentales, un prétraitement des spectres par la procédure dite Standard Normal Deviate (SNV) a tout d'abord été effectué. Les spectres obtenus après correction sont présentés dans la figure 1.

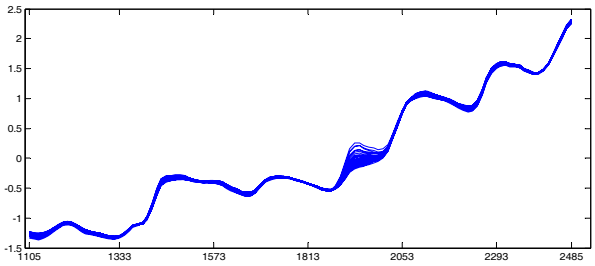


FIGURE 1 – *Spectres corrigés par SNV.*

La méthode de classification CLV sous contrainte d'ordre a été appliquée à la matrice X, centrée et standardisée, correspondant aux spectres corrigés. Une partition en 6 classes a été retenue. Le résultat de la classification est représenté dans la figure 2.

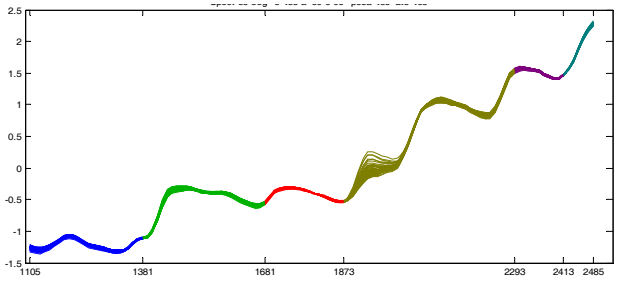


FIGURE 2 – *Spectres segmentés avec une partition en 6 classes sous contrainte d'ordre.*

4. Bibliographie

[CAR 07] CARIOU, V. ET QANNARI, E.M. "Classification de variables sous contrainte de contiguïté ", Journées de Statistique, Angers, 2007.

[DOU 03] DOUZAL, A., "Compression Technique Preserving Correlations of a Multivariate Temporal Sequence", In: M.R. Berthold, H-J Lenz, E. Bradley, R. Kruse, C. Borgelt (eds.) *Advances in Intelligent Data Analysis, V*, 566-577, Springer, 2003.

[LEC 90] LECHEVALLIER, Y., *Recherche d'une partition optimale sous contrainte d'ordre total*, Rapport de recherche INRIA, n°4247, 61 pages, 1990.

[NAE 89] NAES, T. ET KOWALSKI, B., "Predicting sensory profiles from external instrumental measures", *Food Quality And Preference*, vol. 1, 1989, p. 135-147.

[VOS 04] VAN OS, B.J. et MEULMAN, J.J., "Improving Dynamic Programming Strategies for Partitioning", *Journal of Classification*, vol. 21, 2004, p. 207-230.

[VIG 05a] VIGNEAU, E., SAHMER, K., QANNARI, E.M. ET BERTRAND, D., "Clustering of variables to analyze spectral data", *Journal of Chemometrics*, vol. 19, n° 3, 2005, p. 122-128.

[VIG 05b] VIGNEAU, E., SAHMER, K., QANNARI, E.M. ET LADIRAY, D., Classification de variables autour de composantes latentes, *Revue de Statistique Appliquée*, LIV(1), 2005, p. 27-45.

Données manquantes en ACM : l'algorithme NIPALS

Marie Chavent[1,2], **Vanessa Kuentz**[1,2], **Benoit Liquet**[3]

[1] *Université de Bordeaux, IMB, 351 cours de la Libération 33405 Talence Cedex*
{chavent,kuentz}@math.u-bordeaux1.fr
[2] *INRIA Bordeaux Sud-Ouest, Equipe CQFD*
[3] *Université Victor Segalen Bordeaux 2, INSERM, U897, 146 Rue Léo Saignat, 33076 Bordeaux Cedex*
benoit.liquet@isped.u-bordeaux2.fr

RÉSUMÉ. *Le traitement des données manquantes est une question fondamentale en analyse des données. En Analyse en Composantes Principales (ACP), les données manquantes peuvent être gérées à l'aide de l'algorithme NIPALS. Lorsque les données sont qualitatives, l'imputation de nouvelles valeurs est délicate. Dans cet article, nous proposons d'utiliser l'algorithme NIPALS pour le traitement de valeurs manquantes en Analyse des Correspondances Multiples (ACM). Nous présentons l'ACM comme une ACP appliquée aux lignes de la matrice des profils lignes ou encore aux colonnes de la matrice des profils colonnes de la matrice de fréquences. Puis nous présentons l'algorithme itératif pour le calcul de la Décomposition en Valeurs Singulières d'une matrice réelle permettant de gérer les données manquantes. Enfin cette approche est appliquée sur des exemples et comparée à d'autres approches de gestion de données manquantes.*

MOTS-CLÉS : *Analyse des Correspondances Multiples, données manquantes, algorithme NIPALS, Décomposition en Valeurs Singulières.*

1. Introduction

L'apparition de données manquantes est fréquente dans un tableau de données (appareil de mesure défectueux, individus n'ayant pas répondu à la question, etc.) De nombreux auteurs ont étudié le problème d'imputation des valeurs manquantes. Par exemple, [**?**] propose une approche de plus proches voisins pour une imputation basée sur les moindres carrés. Le cas de tables de contingence incomplètes en Analyse des Correspondances est étudié dans [**?**]. Dans [**?**], l'ACM est présentée dans un cadre d'analyse d'homogénéité et une méthode d'estimation des valeurs manquantes est proposée.

Soit $\mathbf{X}$ une matrice de données qualitatives de dimension (n, p) où n objets sont décrits sur p variables qualitatives. On se place dans le cas où certaines entrées x_{ij} sont manquantes. L'idée pour traiter ces données manquantes en Analyse des Correspondances Multiples (ACM) est d'utiliser l'algorithme NIPALS présenté dans [**?**] pour la gestion des données manquantes en Analyse en Composantes Principales (ACP). Pour cela, on définit l'ACM comme une ACP appliquée aux lignes de la matrice des profils lignes ou encore aux colonnes de la matrice des profils colonnes de la matrice de fréquences $\mathbf{F}$ construite à partir du tableau disjonctif complet $\mathbf{K}$ associé à $\mathbf{X}$. On notera $\mathbf{r} = (f_{1.}, ..., f_{i.}, ..., f_{n.})^t$, $\mathbf{c} = (f_{.1}, ..., f_{.s}, ..., f_{.q})^t$, $\mathbf{D}_n = \text{diag}(\mathbf{r})$ et $\mathbf{D}_q = \text{diag}(\mathbf{c})$ avec q le nombre total de modalités.

2. ACM et Décomposition en Valeurs Singulières

L'ACM peut être vue comme une ACP appliquée aux lignes de la matrice des profils lignes centrés $\mathbf{L} = \mathbf{D}_n^{-1}(\mathbf{F} - \mathbf{rc})$ avec les métriques $\mathbf{D}_n$ sur $\mathbb{R}^n$ et $\mathbf{D}_q^{-1}$ sur $\mathbb{R}^q$. La matrice des r premières composantes principales $\mathbf{Y}$ de dimension (n, r) s'écrit, grâce aux formules de passage, $\mathbf{Y} = \mathbf{D}_n^{-1/2}\mathbf{U}\Lambda$ où $\mathbf{U}$ et Λ sont données par la

décomposition DVS de la matrice réelle $\widetilde{\mathbf{F}} = \mathbf{D}_n^{-1/2}(\mathbf{F} - \mathbf{rc})\mathbf{D}_q^{-1/2}$. La DVS de $\widetilde{\mathbf{F}}$ de rang r s'écrit $\widetilde{\mathbf{F}} = \mathbf{U}\Lambda\mathbf{V}^t$ où Λ est la matrice diagonale des valeurs singulières, $\mathbf{U}$ est la matrice des r vecteurs propres normés de $\widetilde{\mathbf{F}}\widetilde{\mathbf{F}}^t$ et $\mathbf{V}$ est la matrice des r vecteurs propres normés de $\widetilde{\mathbf{F}}^t\widetilde{\mathbf{F}}$. On en déduit la formule de reconstruction suivante :

$$\underbrace{\mathbf{D}_n^{-1}(\mathbf{F} - \mathbf{rc})\mathbf{D}_q^{-1/2}}_{\mathbf{Z}} = \underbrace{\mathbf{D}_n^{-1/2}\mathbf{U}\Lambda}_{\mathbf{Y}}\mathbf{V}^t$$

On a donc :

$$\mathbf{Z} = \sum_{h=1}^{r} \underbrace{\mathbf{y}_h\mathbf{v}_h^t}_{\mathbf{Z}_h}$$

Cette formule de reconstruction de $\mathbf{Z}$ nous permet de calculer de manière itérative des composantes principales $\mathbf{y}_h$ en tenant compte des données manquantes.

3. Algorithme itératif pour le calcul de la DVS d'une matrice réelle

Pour le calcul de la première composante on utilise : $\mathbf{Z} = \mathbf{y}_1\mathbf{v}_1^t + \mathbf{E}$. Pour trouver $\mathbf{y}_1$ et $\mathbf{v}_1$ qui minimisent la norme de Froebonius de la matrice d'erreurs $\mathbf{E}$, on itère les deux étapes suivantes :
- Chaque colonne $\mathbf{z}_j$ de $\mathbf{Z}$ s'écrit $\mathbf{z}_j = v_{1j}\mathbf{y}_1 + \mathbf{e}_j$. Si on connaît $\mathbf{y}_1$, le coefficient v_{1j} qui minimise $\| \mathbf{e}_j \|^2$ est $v_{1j} = (\mathbf{z}_j^t\mathbf{y}_1)/(\mathbf{y}_1^t\mathbf{y}_1)$. On calcule donc v_{1j} pour $j = 1$ à q. On normalise ensuite $\mathbf{v}_1$ à 1.
- Chaque ligne $\mathbf{z}_i$ de $\mathbf{Z}$ s'écrit $\mathbf{z}_i^t = y_{i1}\mathbf{v}_1 + \mathbf{e}_i^t$. Si on connaît $\mathbf{v}_1$, le coefficient y_{i1} qui minimise $\| \mathbf{e}_i \|^2$ est $y_{i1} = (\mathbf{z}_i\mathbf{v}_1)/(\mathbf{v}_1^t\mathbf{v}_1)$. On calcule donc y_{i1} pour $i = 1$ à n.
Pour le calcul de la seconde composante on utilise : $\mathbf{Z} - \mathbf{Z}_1 = \mathbf{y}_2\mathbf{v}_2^t + \mathbf{E}_2$ et on itère les deux étapes précédentes sur $\mathbf{Z} - \mathbf{Z}_1$ jusqu'à convergence. Etc...
Les calculs itératif des composantes étant basés sur des sommes, les données manquantes ne sont pas utilisées dans ces sommes.

4. Conclusion

Cette adaptation de l'algorithme NIPALS au cas de l'ACM est appliqué sur des exemples et comparée à d'autres approches de gestion de données manquantes.

5. Bibliographie

[DELEE] De Leeuw, J., Van Der Heijden., (1988), Correspondence Analysis of incomplete contingency tables, *Psychometrika*, **53**(2), 223-233.

[SAP] Saporta, G., (2002), Data fusion and data grafting, *Computational Satistics and Data Analysis*, **38**, 465-473.

[TEN] Tenenhaus, M., (1998), La regression PLS, Editions TECHNIP.

[WAS] Wasito, I., Mirkin, B., (2006), Nearest neighbours in least-squares data imputation algorithms with different missing patterns, *Computational Satistics and Data Analysis*, **50**, 926-949.

Application des SVM à la classification des Activités de la Vie Quotidienne d'une personne à partir des capteurs d'un Habitat Intelligent pour la Santé

Anthony Fleury, Norbert Noury, Michel Vacher

[1] *Laboratoire TIMC-IMAG, équipe AFIRM, UMR CNRS/UJF 5525*
Faculté de Médecine de Grenoble bâtiment Jean Roget, 38706 La Tronche Cedex - France
fleury_anthony@hotmail.com, Norbert.Noury@imag.fr

[2] *Laboratoire LIG, équipe GETALP, UMR CNRS/UJF/INPG 5217*
220 rue de la chimie, BP 53, 38 041 Grenoble Cedex 9 - France
Michel.Vacher@imag.fr

RÉSUMÉ. *Le vieillissement accéléré de la population ces dernières années met en exergue les problèmes liés à la perte d'autonomie des personnes âgées. Les Sciences de l'Information et de la Communication permettent d'entrevoir la possibilité d'une surveillance personnalisée à domicile pour les personnes âgees. Cet article présente le choix et la disposition d'un ensemble de capteurs dans un appartement, les paramètres extraits pour la classification et l'utilisation des SVM avec la méthode un-contre-tous pour classifier un ensemble de sept activités de la vie quotidienne. Des expérimentations ont été menées sur treize personnes afin de constituer une base d'apprentissage et de validation.*

MOTS-CLÉS : *SVM, Habitat Intelligent pour la Santé, Activités de la Vie Quotidienne*

1. Introduction

L'âge moyen des populations des pays développés croît constamment ces dernières années. Les Nations-Unies prévoient que 22% des personnes seront âgées de plus de 65 ans d'ici à 2050. Pour s'adapter à cette nouvelle démographie, différents laboratoires de recherche travaillent sur l'étude des habitudes de vie de la personne afin de détecter des changements dans le comportement susceptible d'indiquer une situation à risque. Ainsi, le projet CARE [KRö 08] se base sur les Modèles de Markov Cachés avec de nombreux capteurs (localisation, température...) afin d'apprendre à reconnaître les activités « Aller aux toilettes » et « Sortir de l'appartement ». En Grande-Bretagne, [HON 08] utilise des tags RFID pour créer un modèle distinguant les activités de préparation d'une boisson chaude ou froide avec les activités d'hygiène en se basant sur la théorie de l'évidence (Dempster-Shafer). Les signatures électriques représentent une autre piste de recherche qui est actuellement explorée par les chercheurs [TSU 08, BER 08]. Nous présentons ici l'utilisation des SVM pour la classification de 7 activités de la vie quotidienne.

2. L'habitat Intelligent pour la Santé (HIS)

2.1. *Équipement de l'habitat*

L'ensemble de ces travaux se basent sur l'Habitat Intelligent pour la Santé installé à la faculté de Médecine de Grenoble. Cet habitat, présenté à la Figure 1, est un appartement de type F2 entièrement équipé et habitable, dans

lequel ont été placés un ensemble de capteurs reliés à une pièce technique (pour l'acquisition synchronisée de tous les capteurs).

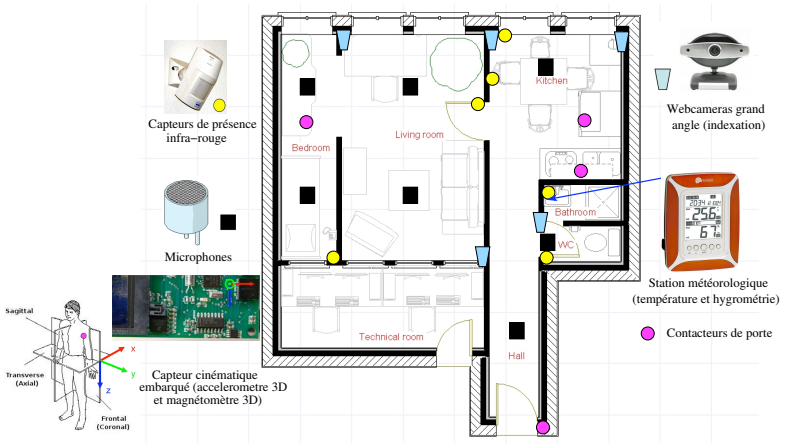


FIGURE 1. *Plan et équipement de l'Habitat Intelligent pour la Santé du laboratoire TIMC-IMAG à la faculté de médecine de Grenoble*

Parmi les capteurs présents, se trouvent tout d'abord des détecteurs infra-rouges, utilisés pour la localisation de la personne. Ils sont disposés de manière à couvrir des endroits correspondant à des tâches précises (le canapé, le lit, la table de la cuisine...). Chaque détection de mouvement dans le cône du capteur envoie un signal et l'heure de la détection ainsi que le capteur l'ayant réalisée sont conservées dans une base de donnée. Viennent ensuite les sept microphones, installés dans le plafond et répartis dans l'appartement. Ils sont connectés à un système de reconnaissance de la parole et de classification des sons développé par le Laboratoire d'Informatique de Grenoble (équipe GETALP). Ce système est responsable d'abord de la détection d'un son sur les différents microphones, en calculant un « bruit de fond » sur chacun d'entre eux et en utilisant un seuil adaptatif pour détecter le début et la fin de l'occurence d'un son. Des modèles à base de Mixtures de Gaussiennes calculent la probabilité d'appartenance aux sons de la vie courante ou à la parole. Lorsque la probabilité la plus haute est la parole, le son est transféré à une application de reconnaissance automatique de la parole dont les modèles (acoustique et de language) ont été adaptés pour la langue française. Pour un son de la vie courante, un autre ensemble de GMM va le classer dans l'une des 8 classes (bruit de pas, claquement de porte, son de serrure, son de vaisselle, cri, chute d'objet, sonnerie de téléphone et bri de verre). Un tri des détections simultanées pour prendre la décision la plus probable (en considérant le rapport signal sur bruit des différents microphones) donne la décision finale. L'habitat est également équipé d'une station météorologique donnant des indications sur la température et l'hygrométrie, placée dans la salle de bain et de contacteurs de porte sur la commode, le réfrigérateur et le placard de la cuisine pour détecter l'utilisation de ces commodités. Enfin, le dernier capteur est embarqué sur la personne. Développé par l'équipe AFIRM, il se base sur des accéléromètres et des magnétomètres pour détecter les transferts posturaux et les périodes de marche. Ces capteurs ont été testés individuellement dans différentes études du laboratoire. Enfin, des caméras sont présentes pour filmer l'appartement. Elles ne sont utilisées uniquement pour l'indexation des activités réalisées.

2.2. *Sélection des paramètres*

À partir d'analyses en composantes principales sur des données issues d'expérimentations préliminaires, nous avons déterminé les paramètres pertinents pouvant être extraits de chacune de ces mesures afin de classifier les activités de la vie quotidienne. Ces paramètres ont été choisis pour une bonne reproductibilité intra-classe et la meilleur différentiation inter-classe ainsi qu'une redondance faible. Le tableau 1 liste ces paramètres et montre des exemples d'informations apportées par celui-ci pour la classification de différentes activités.

TABLE 1. *Paramètres retenus pour les différentes modalités*

Modalité	Paramètres retenus	Exemple d'informations
Actimétrie	Pourcentage de temps des différentes postures (assis, debout, couché) et de marche	Couché pendant l'activité dormir, marche puis s'asseoir pour repos et repas...
Microphones	Nombre d'évènements par classes et par microphones	Communication (sonnerie de téléphone et parole), repas (son de vaisselle), repos (TV ou radio dans le salon)
Détecteurs infra-rouges	Pourcentage de temps par pièce et nombre d'évènements par détecteur	Repas (cuisine), hygiène (salle de bain)...
Contacteurs de porte	Pourcentage de temps « ouvert » et position la plus présente dans la fenêtre	Repas (utilisation du réfrigérateur et du placard), habillage (commode).
Environnement	Mesure différentielle (15 dernières minutes) température et hygrométrie	Hygiène (utilisation de la douche)

3. Resultats et conclusions

3.1. *Protocole experimental pour l'acquisition de données réelles*

Afin de construire une base d'apprentissage et de tester et valider les algorithmes, nous avons mené une campagne d'expérimentations permettant l'acqusition de données réelles incluant la mesure sur 13 personnes jeunes et en bonne santé (6 femmes, 7 hommes, moyenne d'âge 30 ans). La collecte des données a duré au minimum 23 minutes et au maximum 1h35 en fonction de la personne. La personne visitait l'appartement (pour s'y sentir à l'aise et connaître la place des différents éléments) puis s'y retrouvait seule pendant le temps désiré, avec comme seule consigne de réaliser au moins une fois chacune des sept activités. Elle n'avait de consigne ni sur l'ordre ni sur la manière de réaliser l'activité. Les éléments nécessaires (nourriture, TV, téléphone...) étaient à disposition.

3.2. *Mise en forme des données et validation*

Afin d'objectiver à terme des scores tels que la grille ADL, les activités ont été choisies par rapport à celles décrites dans celles-ci. L'apprentissage permettra de différentier les activités suivantes : repos, détente, prise d'un repas, hygiène, élimination, communication avec l'extérieur et habillage/déshabillage. Ces activités ont été découpées en fenêtres temporelles de 3 minutes (temps moyen de la plus courte activité) pour lesquels les paramètres sont calculés. Les SVM sont utilisés avec la méthode « un-contre-tous » associée à un vote majoritaire. Ce classifieur a été choisi tout d'abord pour sa bonne capacité de généralisation dans de nombreux problèmes[BUR 98] mais aussi pour ses bons résultats sur des ensembles de données d'apprentissage de petite taille (contrairement aux réseaux de neurones qui demandent des base de données plus importantes). Les noyaux gaussiens et polynômiaux ont été comparés (avec optimisation des hyper-paramètres). L'ensemble a été implémenté sous Matlab^{TM}. Les données ont été normalisées, en considérant la base d'apprentissage et en donnant à chaque dimension une moyenne nulle et un écart type de 1. Les données de normalisation sont ensuite utilisées pour la donnée de test. Des caméras vidéos permettaient l'indexation de l'ensemble des expérimentations. Pour des raisons évidentes de vie privée, toilettes et salle de bain n'étaient pas filmés mais la position de la porte permettait de connaître l'action en cours. Les données, synchronisées entre elles, ont été indexées afin de créer un ensemble de fenêtres temporelles de trois minutes étiquetées et disponibles pour l'apprentissage. La validation croisée s'est faite par la méthode du « leave-one-out » du fait du faible nombre de fenêtres. Le tableau 2 montre la composition de la base d'apprentissage.

3.3. *Résultats et conclusions*

Le tableau 2 montre les résultats sur les deux types de noyaux détaillés pour l'ensemble des classes. Comme nous pouvons le noter, les résultats sont encourageants surtout étant donné le faible nombre de données d'apprentissage mais il y a de grandes disparités entre les classes. Ceci s'explique par le déséquilibre entre les différentes bases d'apprentissage qui va engendrer une distorition des SVM. Globalement, les résultats atteints avec un noyau gaussien sont meilleurs qu'avec un noyau pôlynomial. Les hyper-paramètres ont à chaque fois été optimisés.

TABLE 2. *Répartition des différentes classes dans la base d'apprentissage et résultats de la classification par les séparateurs à vaste marge avec un noyau Polynômial et un noyau Gaussien*

	Base d'apprentissage		Taux de classification correcte	
Classe	Taille de la base	Pourcentage	Noyau Polynômial	Noyau Gaussien
Repos	49	19,4%	77,55%	93,87%
Détente	75	29,7%	76,71%	78,08%
Habillage	16	6,3%	56,25%	75,00%
Repas	45	17,8%	84,44%	97,78%
Élimination	16	6,3%	68,75%	93,75%
Hygiène	14	5,5%	50,00%	64,28%
Communication	17	6,7%	89,47%	89,47%
Total	252	100%	75,86%	86,21%

Ces premiers résultats sont encourageants sur la méthode utilisée et les capteurs présents. Des travaux sur la pertinence de chaque variable et des différents capteurs en considérant cette base d'activités sont en cours. De futurs travaux viendront confirmer ces résultats à l'aide de l'acquisition de nouvelles données. Notamment, il y a plusieurs points encore en suspend après ces travaux. Le premier est de savoir le comportement avec une plus grande base, plus équilibrée et plus représentative. Le second serait de travailler sur l'ajout d'une classe décrivant les transitions entre deux activités. Notre classifieur agit, pour l'instant, sur des fenêtres temporelles de 3 minutes sans tenir compte des fenêtres précédentes et suivantes. Pour implémenter ceci hors du cas de données indéxées (et donc découpées correctement), nous nous retrouverons avec des fenêtres qui ont une partie de la fin d'une activité et du début d'une autre. Le classifieur dans ce cas réagira d'une manière imprévisible. Il serait donc intéressant d'ajouter cette nouvelle classe. Nous pourrions également intégrer des connaissances *a priori*. En effet, la localisation va restreindre les activités possible et l'heure de la journée va nous donner une indication sur l'activité qui peut être réalisée. Nous avons travaillé ici sur de la classification automatique ne prenant en compte aucune donnée *a priori* mais nous pourrions améliorer les résultats en ajoutant ces connaissances.

4. Bibliographie

[BER 08] BERENGUER M., GIORDANI M., GIRAUD-BY F., NOURY N., Automatic detection of Activities of Daily Living from Detecting and Classifying Electrical Events on the Residential Power Line, *HealthCom'08 - 10th IEEE Intl. Conf. on e-Health Networking, Applications and Service*, 2008.

[BUR 98] BURGES C. J. C., A Tutorial on Support Vector Machines for Pattern Recognition, *Data Mining and Knowledge Discovery*, vol. 2, n^o 2, 1998, p. 121-167.

[HON 08] HONG X., NUGENT C., MULVENNA M., MCCLEAN S., SCOTNEY B., Evidential fusion of sensor data for activity recognition in smart homes, *Pervasive and Mobile Computing*, vol. 5, n^o 3, 2008, p. 1-17.

[KRö 08] KRÖSE B., VAN KASTEREN T., GIBSON C., VAN DEN DOOL T., CARE : Context Awareness in Residences for Elderly, *International Conference of the International Society for Gerontechnology*, Pisa, Tuscany, Italy, June 4-7 2008.

[TSU 08] TSUKAMOTO S., HOSHINO H., TAMURA T., Study on Indoor Activity Monitoring by using Electric Field Sensor, *International Conference of the International Society for Gerontechnology*, Pisa, Tuscany, Italy, June 4-7 2008.

Dissimilarity-based metric for data classification using Support Vector Classifiers

Agata Manolova[1], Anne Guerin-Dugue[2]

[1]*Technical University Sofia, 8 ave Kliment Ohridski, Sofia 1000, Bulgaria,*
amanolova@tu-sofia.bg
[2]*Department of Images and Signal, GIPSA-lab, Grenoble INP-CNRS, Grenoble, France*
anne.guerin@gipsa-lab.inpg.fr

ABSTRACT. A class is a concept of a set of objects possessing similar characteristics. This implies that the notion of "similarity/dissimilarity" is as fundamental as of "feature", since it is the similarity which groups the objects together in a class. The distance representation is most commonly used as a dissimilarity measure because is usually the simplest to calculate. Dissimilarity-based pattern recognition offers new possibilities for building classifiers on a distance representation such as kernel methods or the k nearest neighbors (kNN) rule. The goal of this work is to expand and ameliorate the advantageous and rapid adaptive approach to learn only from dissimilarity representations developed by Guerin and Celeux [GUE 01] by using the effectiveness of the Support Vector Machines algorithm for real-world classification tasks. This method can be an alternative approach to the known methods based on dissimilarity representations such as Pekalska's dissimilarity classifier [PEK 05], Haasdonk's kernel-based SVM classifier [HAA 04] and to kNN classifier and can be as effective as them in terms of accuracy for classification. Practical examples on artificial and real data show interesting behavior compared to other dissimilarity-based methods.

MOTS-CLÉS : dissimilarity representation, dissimilarity–based classification, support vector classifier, support vector machines, discriminant analysis.

1. Introduction

One of the most recent novel developments in the field of statistical pattern recognition is the concept of dissimilarity-based classifiers (DBCs). Philosophically, the motivation for DBCs is the following: if we assume that "Similar" objects can be grouped together to form a class, a "class" is nothing more than a set of these "similar" objects. Based on this idea, it is possible that the notion of proximity (similarity or dissimilarity) is actually more fundamental than that of a feature. Thus, DBCs are a way of defining classifiers between the classes, which are not based on the feature measurements of the individual patterns, but rather on a suitable dissimilarity measure between them. The advantage of this methodology is that since it does not operate on the class-conditional distributions, the accuracy can exceed theoretically the Bayes' error bound. Another salient advantage of such a paradigm is that it does not have to confront the problems associated with feature spaces such as the "curse of dimensionality", and the issue of estimating a large number of parameters.
The distance representation is most commonly used as dissimilarity because is usually the simplest measure. A dissimilarity value expresses a magnitude of difference between two objects and becomes zero only when they are identical. A straightforward way of dealing with dissimilarity representation is based on a distance-relations between objects, which leads to the rank-based methods, e.g. to the k nearest neighbors rule. The kNN rule works well, but suffers from local decisions, needs a large learning set to be effective - noise sensitive and not suited for distances which do not respect the triangular inequality. There are few other methods dealing with dissimilarities such as (i) the prototype dissimilarity classifier developed by Duin and Pekalska [DUI 02] which is a global technique depending on the choice of representative prototypes for each class and (ii) the distance on substitution kernels with SVM developed by Haasdonk [HAA 04].
This paper focuses on the incorporation of SVM in to the dissimilarity-based classifier "Shape Coefficient" described in [GUE 01], [MAN 08]. The Shape Coefficient (Cs) is defined from simple statistics (mean and

variance) on the dissimilarity data. The proposed decision rules are based on this Shape Coefficient description and on optimal separating hyperplane with SVC, using the CS coefficient as dissimilarity on the input space. This provides a decision rule with a limited number of parameters per class.
The paper is organized as follows: in Section 2 we describe the theoretical basis of this approach; in Section 3 we provide experimental results on artificial and real-life data sets. Finally, Section 4 concludes the paper.

2. Description of the « Shape Coefficient »

Let us consider a two-class classification problem where ω_1 is the first class and ω_2 the second class. Let N be a set of objects o_i to be classified, D is the dissimilarity ($N \times N$) table between each object such as: $D = \left[d(o_i, o_j) : 1 \le i, j \le N \right]$. Following [GUE 01] and [MAN 08], the Shape Coefficient describes the proximity of an object to a given class (for example for ω_1, eq. 1):

$$Cs(o_i, \omega_1) = \frac{\gamma_1 [\overline{d^2(o_i, \omega_1)} - I(\omega_1)]^2}{[\mathrm{var}(d^2(o_i, \omega_1))]^{\delta_1}}, \quad (1)$$

where $\overline{d^2(o_i, \omega_1)}$ is the empirical average of the dissimilarity between object o_i and all the observations in class ω_1, $var(d^2(o_i, \omega_1))$ is the empirical variance, and $I(\omega_1)$ is the class inertia computed as the empirical mean of all the squared dissimilarities between objects in class ω_1. The numerator deals with the "position" of the observation o_i relatively the class center. The denominator interpretation is more complex, taking into account the "structure" (orientation, shape, intrinsic dimension…) of the observations distribution in the class. Then the parameters γ_1 and δ_1 are learning parameters to best fit this data structure. The equation for $Cs(o_i, \omega_2)$ with the class ω_2 is equivalent to (1) and has two fitting parameters γ_2 and δ_2. The decision rule for a two-class classification problem for an object o_i is given then by the following relation:

$$Cs(o_i, \omega_1) \underset{2}{\overset{1}{\lessgtr}} Cs(o_1, \omega_2) \quad (2)$$

This relation is defined by four learning parameters for a two-class problem. From (2), we have defined in [MAN 08] learning procedures which overcome or have similar performance compared the kNN rule. Here we proposed a more straightforward learning procedure based on SVM optimization, based on the three independent parameters: δ_1, δ_2, γ_1/γ_2.

3. Decision rule using SVC optimization

The idea is to propose a new representation of the observations which must be compatible with a linear decision rule in this new features space. The quantities $Cs(o_i, \omega_1)$ and $Cs(o_i, \omega_2)$ being positive, we can transform (2) using the logarithmic function as follows:

$$\log\left(\frac{\gamma_1}{\gamma_2}\right) + 2\log(\overline{d}^2(o_i, \omega_1) - I(\omega_1)) - 2\log(\overline{d}^2(o_i, \omega_2) - I(\omega_2)) - \delta_1 \log(\mathrm{var}(d^2(o_i, \omega_1))) + \delta_2 \log(\mathrm{var}(d^2(o_i, \omega_2))) \underset{2}{\overset{1}{\lessgtr}} 0 \quad (3)$$

This is in fact, a linear decision rule in a 4-dimensional input space. Following (3), we can represent each object o_i using a four dimensional vector x_i ($x_i = [x_{i1}\ x_{i2}\ x_{i3}\ x_{i4}]^T$):

$$\begin{aligned} x_{i1} &= 2\log(\overline{d^2(o_i, \omega_1)} - I(\omega_1)) & x_{i3} &= -\log(\mathrm{var}(d^2(o_i, \omega_1))) \\ x_{i2} &= -2\log(\overline{d^2(o_i, \omega_2)} - I(\omega_2)) & x_{i4} &= \log(\mathrm{var}(d^2(o_i, \omega_2))) \end{aligned} \quad (4)$$

So now, the decision rule (3) becomes: $\beta^T x_i + \beta_0 \underset{2}{\overset{1}{\lessgtr}} 0$, (5)

where $\beta = [1\ 1\ \delta_1\ \delta_2]^T$ is the normal to the optimal separating hyperplane and $\beta_0=\log\left(\pi/\gamma_2\right)$ is the bias from the hyperplane to the origin. Labeling the objects with the auxiliary variables per class, such as $y_i=-1$ *for* $o_i\in\omega_1$ and $y_i=1$ *for* $o_i\in\omega_2$, we have the following classical linear decision rule : $y_i=sign(\beta^T x_i+\beta_0)$. (6)

This is the standard decision rule for SVC. Here, the difference is the vector β normal to the optimal hyperplane: *it is constraint to have the same two first components:* $\beta_1=\beta_2$.

Thus finding the optimal hyperplane when the 2 classes are non separable consists of this optimization problem solved by using the Lagrange multipliers [BUR98]:

$$\min_{\beta,\beta_0} \frac{1}{2}\|\beta\|^2 + C\sum_{i-1}^{N}\zeta_i \quad , \qquad (8)$$
$$subject\ to\ y_i(\beta^T x_i+\beta_0)\geq 1-\zeta_i, \zeta_i\geq 0, i=0,...,N$$

where ζ_i are the slack variables, associated with all the objects. If the object o_i is classified in the wrong class then $\zeta_i>1$. The parameter C corresponds to the penalty for errors and it is chosen by the user. In order to introduce the constraints on the β vector, we consider the observations x_i into two orthogonal subspaces such as:

$x_i=[x_i'\ x_i'']^T, x_i'=[x_{i1}\ x_{i2}]^T, x_i''=[x_{i3}\ x_{i4}]^T$ and also $\beta=[\beta'\ \beta'']^T, \beta'=\|\beta'\|[1\,1]^T/\sqrt{2}, \beta''=[\beta_3\ \beta_4]^T$. (9)

After introduction of the scalar product u_i', the optimization problem is then transformed such as:

$$\min_{\|\beta'\|,\beta'',\beta_0} \frac{1}{2}\|\beta'\|^2+\frac{1}{2}\|\beta''\|^2+C\sum_{i-1}^{N}\zeta_i \quad . \qquad (10)$$
$$subject\ to\ y_i(\|\beta'\|u_i'+\beta''^T x_i''+\beta_0)\geq 1-\zeta_i, \zeta_i\geq 0, i=0,...,N \text{ with } u_i'=\left\langle [1\quad 1]^T, x_i'\right\rangle/\sqrt{2}$$

Very interesting is the case of Least Square Support Vector Machines (LSSVM) where the optimization problem consists of:

$$\min_{\beta,\beta_0} \|\beta\|+\frac{C}{2}\sum_{i-1}^{N}\zeta_i^2 \quad . \qquad (11)$$
$$subject\ to\ y_i(\beta^T x_i+\beta_0)=1-\zeta_i,\ i=0,...,N$$

The LSSVM leads to a solution of a linear system instead of a quadratic programming problem [SUY 99]. A similar straightforward modification can be done to the constraints of the β vector.

4. Experimental results

All the experiments are done using SVM^{Light}, an implementation of Support Vector Machines in C by Thorsten Joachims (http://svmlight.joachims.org/) and the "LS-SVMlab" in Matlab, developped in K.U. Leuven University, Belgium (http://www.esat.kuleuven.ac.be/sista/lssvmlab/). We have made source modifications in order to implement the supplementary constraints on the β vector.

In order to get insights into the proposed method, we first perform 2D experiments on two artificial data sets: two overlapped Gaussian distributions named Gauss2 and Gauss3 in table 1. Then we target at a real-world problem. By cross validation, the regularization parameter is set to C = 10. The learning set is composed on 100 observations per class, ten learning sets are generated. The classification errors are determined on one independent test set of 200 examples per class. After each learning phase, a test-error is computed. Table 1 presents the mean test-error and the variance after the ten learning phases. We include two reference classifiers to compare the overall classification performance: the kNN and the Shape Coefficient from [MAN 08].

The real-world experiment is done with the database of 3-8 handwritten digits. The input dissimilarity matrix (200 images per class) is made up of the Hausdorff distances between each pair of binary images. This dataset is a two-class database of 400 handwritten binary numbers, for the digits "3" and "8". This dataset comes from the NIST database [GAR 1997]. The observations are then compared using the Hausdorff distance. This database has been kindly sent by E. Pekalska [PEK 2005]. The results are displayed in Table 2.

***TABLE 1** – Average classification errors [in %], numbers in parenthesis denote standard deviations.*

	kNN	Cs	Cs-SVM	Cs-LSSVM
Gauss2	4.03 (0.4)	4.15 (0.9)	3.97 (0.3)	7.57 (4.10)
Gauss3	11.83 (1.17)	10.67 (0.63)	10.82 (0.54)	13.05 (2.86)

***TABLE 2** – Average classification errors [in %].*

	kNN	Cs	Cs-SVM	Cs-LSSVM
3-8 digits	7.75	3.75	3	3

5. Conclusions, Perspectives

We have proposed a new way of optimizing the parameters of the proximity index "Shape Coefficient". It used the SVM decision rules which allow us to find the optimal solution for our classification problem. With only two parameters per class, the model for class description is compact and parsimonious. The model is flexible, effective and fast in different classification tasks. The result of the comparison with the kNN and the standard Cs shows better results for the classification error. The LSSVM are simpler to implement because they use linear system for the optimization but the results show that they are more volatile and depend much more on the data to classify then the standard SVM and they are more sensitive to noisy objects. It is proven that the LSSVM perform better with kernels different then the linear kernel. The Cs-SVM is a global method with adjustable parameters according to the properties of the class so it performs better then the kNN rule in case of (1-NN or 3-NN). For the Gauss2 distributions the kNN performs a little bit better than the CS-SVM but these results are due to the higher number of neighbors needed to better classify the objects (more then 9 neighbors). In the [MAN 08] we have proven that the range of parameter δ is limited. This is also true in the case of Cs-SVM and Cs-LSSVM. The results with the real-word experiments encourage us to propose this metric as a good alternative to other dissimilarity-based classifiers. Because the metric uses only 2 parameters per class and a linear kernel, data classification is very fast (0.07 seconds in SVM^{Light} for 200 points). This work goes on with the applications of the proposed method for multi-class real-world problems and comparisons with the Pekalska and Haasdonk's classifiers.

6. Bibliography

[MAN 08] MANOLOVA A, GUERIN-DUGUE A., C*assification of dissimilarity data with a new flexible Mahalanobis-like metric*. Pattern Anal. Appl. 11(3-4): 337-351 (2008), Springer Link

[PEK 05] PEKALSKA E., *Dissimilarity representations in pattern recognition: Concept, theory and applications*, Phd Thesis, ISBN 90-9019021-X, 2005.

[HAA 04] HAASDONK B., BAHLMANN C., "Learning with Distance Substitution Kernels", *Pattern Recognition - Proc*. of the 26th DAGM Symposium, Tubingen, Germany, August/September 2004.

[GUE 01] GUERIN-DUGUE A., CELEUX G., "Discriminant analysis on dissimilarity data : A new fast gaussian-like algorithm ". AISTATS'20001, Florida, USA, pages 202-207, january 2001

[DUI 02] DUIN R., PEKALSKA E., "Dissimilarity representation for building good classifiers", *Pattern Recognition Letters*, vol. 23, no. 8, 2002, 943-956.

[BUR 98] BURGES J. C., *A tutorial on Support Vector Machines for pattern recognition, Data Mining and Knowledge discovery*, 2, 121-167 (1998), Kluwer Academic Publishers

[GAR 97] GARRIS M, BLUE J., CANDELA G., GROTHER P., JANET S., and WILSON C., *NIST Form-Based Handprint Recognition System (Release 2.0)*, Internal report, January 1997.

[SUY 99] SUYKENS J.A.K., LUCAS L., VAN DOOREN P., DE MOOR B., VANDERWALLE J., "Least squares support vector machine classifiers : a large scale algorithm", in *Proc. of the European Conference on Circuit Theory and Design* (ECCTD'99), Stresa, Italy, Sep. 1999, pp. 839-842.

Analyse Discriminante Dissymétrique

Rafik Abdesselam

CREM UMR CNRS 6211,
Université de Caen Basse-Normandie,
Campus 4 - Claude Bloch, 14032 Caen, France
rafik.abdesselam@unicaen.fr

RÉSUMÉ. *L'approche factorielle proposée peut être considérée comme un complément utile pour analyser l'association symétrique et/ou dissymétrique entre un ensemble de variables quantitatives explicatives et une variable qualitative cible à expliquer (détermination des moments principaux et représentations graphiques simultanées et barycentriques), dans un but de discrimination et de classement. Nous utilisons une famille de générateurs de produits scalaires dans l'espace des individus à expliquer afin de déterminer celui qui maximise au mieux les critères usuels d'inertie expliquée et du pourcentage de bien classés. Un exemple sur données réelles illustre cette approche dissymétrique dont les résultats sont comparés à ceux de l'analyse discriminante classique (symétrique), de la régression logistique et de la régression logistique PLS.*

MOTS-CLÉS : *Coefficients d'association symétrique et dissymétrique, analyse en composantes principales, analyse de la variance multivariée, analyse factorielle discriminante.*

1. Introduction

L'approche proposée consiste à utiliser des générateurs de produits scalaires dans un but de discrimination et de classement. L'analyse discriminante (AD), travaux de Fisher [FIS 38], utilise les distances classiques de Mahalanobis et du khi-deux. L'analyse factorielle discriminante dissymétrique (ADD) proposée consiste à générer des produits scalaires dans l'espace des individus à expliquer afin de rechercher celui qui maximise au mieux les critères de classement, à savoir, le coefficient d'association et le pourcentage d'individus bien classés. L'approche proposée est évaluée puis comparée, sur la base d'une application sur données réelles, à l'analyse discriminante classique ainsi qu'aux modèles logistique et logistique PLS.

2. Analyses de l'association dissymétrique

On utilisera les notations suivantes pour exposer l'approche discriminante dissymétrique proposée.
$X_{(n,p)}$ est le tableau de données quantitatives associé à l'ensemble $\{x^j\ ;\ j = 1, p\}$, des variables discriminantes centrées à n lignes-individus et p colonnes-variables,
$Y_{(n,q)}$ est la matrice des variables indicatrices $\{y^k\ ;\ k = 1, q\}$ associées aux q modalités de la variable cible y,
$E_x = \mathbf{R^p}$ [resp. $E_y = \mathbf{R^q}$] est le sous-espace explicatif [resp. à expliquer] des individus associés au tableau de données $X_{(n,p)}$ [resp. $Y_{(n,q)}$],
$N_x = \{x_i \in E_x\ ;\ i = 1, n\}$ [resp. $N_y = \{y_i \in E_y\ ;\ i = 1, n\}$] est le nuage des individus associé à X [resp. Y],
M_x [resp. M_y] est la matrice associée au produit scalaire dans le sous-espace des individus E_x [resp. E_y],
$D_n = \frac{1}{n} I_n$ est la matrice diagonale des poids des individus, où, I_n désigne la matrice unité d'ordre n,
D_q est la matrice diagonale des poids des q modalités de la variable cible, définie par $[D_q]_{kk} = \frac{n_k}{n}\ \ \forall k = 1, q$,
$\chi^2_y = D_q^{-1}$ est la matrice de la distance du khi-deux,
V_x^- est la matrice de la distance de Mahalanobis, inverse généralisée de Moore-Penrose de la matrice de variances et covariances $V_x = {}^t X\, D_n\, X$,

$[(V_yM_y)^{1/2}]^+$ est l'inverse généralisée de Moore-Penrose, relativement à M_y, de $(V_yM_y)^{1/2}$,
$N_y^x = \{P_{E_x}(y_i)\,;\, i=1,n\}$ [resp. $N_x^y = \{P_{E_y}(x_i)\,;\, i=1,n\}$] est le nuage des individus associé au tableau de données $\widehat{Y}^x$ [resp. $\widehat{X}^y$] ; projection orthogonale de N_y [resp. N_x] sur E_x muni de la distance de Mahalanobis [resp. sur E_y muni de la distance du khi-deux].

Définition :

L'ADD ou l'analyse factorielle de l'association dissymétrique entre un ensemble de variables quantitatives explicatives $\{x^j\,;\, j=1,p\}$ et une variable qualitative à expliquer y, consiste à effectuer les deux analyses en composantes principales (ACP) suivantes :

$$\text{ACP}\,\{\widehat{Y}^x = Y\,M_y[(V_yM_y)^{1/2}]^+\,V_{yx}\ ;\ V_x^-\ ;\ D_n\}\ \text{ ou }\ \text{ACP}\,\{G=\chi^2_y\,(V_yM_y)^{1/2}\,V_{yx}\ ;\ V_x^-\ ;\ D_q\}\quad(1)$$

$$\text{ACP}\,\{\widehat{X}^y = XV_x^-V_{xy}\ ;\ M_y\ ;\ D_n\}\quad(2)$$

La construction statistique et géométrique du nuage $N_y^x \subset E_x$ analysé dans l'ACP(1), joue un rôle fondamental dans notre approche. On montre que le produit scalaire choisi dans le sous-espace explicatif E_x pourrait être quelconque. En effet, pour l'ACP(1), l'inertie du nuage N_y^x (par rapport à son centre de gravité) ne dépend pas de M_x. On choisit alors $M_x = V_x^-$ (distance de Mahalanobis) dans E_x afin de simplifier les calculs et de retrouver les résultats fondamentaux de l'AD et du MANOVA qui correspondent respectivement à ceux de l'ACP(1) et de l'ACP(2), dans le cas unique où $M_y = \chi^2_y$.

Quant au choix de M_y dans le sous-espace à expliquer E_y, on utilise des générateurs de produits scalaires $M_y(\alpha)$. Dans le contexte de notre approche, nous suggérons les expressions simples suivantes :

$$^1M_y(\alpha) = \alpha\,I_q + (1-\alpha)\chi^2_y \qquad \text{et} \qquad ^2M_y(\alpha) = \alpha\,D_q + (1-\alpha)I_q \qquad \text{avec} \qquad \alpha\in[0,1]$$

ces générateurs vont évoluer de la position symétrique $^1M_y(0) = \chi^2_y$ (distance du khi-deux) vers la position dissymétrique $^2M_y(1) = D_q$ en passant par la position dissymétrique $^1M_y(1) = I_q = {}^2M_y(0)$, où I_q désigne la matrice unité d'ordre q.

Soient,

- $\{(\lambda_j(x)\,;\,u_j)\,/j=1,r\}$ [resp. $\{(\lambda_j(y)\,;\,v_j)\,/j=1,r\}$] les moments principaux non nuls et les vecteurs axiaux principaux V_x^--normés de l'ACP(1) [resp. M_y-normés de l'ACP(2)],
- $\{U^j/j=1,r\}$ [resp. $\{V^j/j=1,r\}$] les composantes principales correspondant D_q-normées [resp. D_n-normées].

La Figure 1 présente les schémas de dualité ainsi que les liens de passage entre les deux analyses factorielles. Où, $P = {}^t[(V_yM_y)^{1/2}]\,D_{[1/\sqrt{\lambda_j}]}$ désigne la matrice de passage et $D_{[1/\sqrt{\lambda_j}]}$ la matrice diagonale dont les éléments sont les inverses des racines carrées des moments principaux.

$$\begin{array}{ccccccc}
(\lambda_j\,;\,u_j)_{j=1,r} & {}^tG & & {}^tP & (\lambda_j\,;\,v_j)_{j=1,r} & {}^t\widehat{X}^y & \\
E_x=R^p & \longleftarrow & F_x^\star=R^q & \longleftarrow & E_y=R^q & \longleftarrow & F_y^\star=R^n \\
V_x^-\downarrow & \textbf{\textit{ACP(1)}} & \uparrow D_q & & M_y\downarrow & \textbf{\textit{ACP(2)}} & \uparrow D_n \\
E_x^\star=R^p & \longrightarrow & F_x=R^q & \longrightarrow & \mathrm{E}_y^\star=R^q & \longrightarrow & F_y=R^n \\
 & G & (U^j)_{j=1,r} & P & & \widehat{X}^y & (V^j)_{j=1,r}
\end{array}$$

FIG. 1. *Schémas de dualité.*

Propriété :

M_y étant un produit scalaire quelconque dans E_y, $\forall j = 1, r$ on a les formules de transition suivantes :

1) $\lambda_j(x) = \lambda_j(y) = \lambda_j$

2) $u_j = \frac{1}{\sqrt{\lambda_j}} V_{xy}\, M_y\, v_j$; $v_j = \frac{1}{\sqrt{\lambda_j}} V_{yx} V_x^-\, u_j$

3) $U^j = \frac{1}{\sqrt{\lambda_j}}\, \chi_y^2\, (V_y M_y)^{1/2}\, {}^t Y D_n V^j$; $\mathrm{V}^j = \frac{1}{\sqrt{\lambda_j}}\, X V_x^- V_{xy}\, {}^t[(V_y M_y)^{1/2}]\, U^j$

4) $\widetilde{V}^j = \frac{1}{\lambda_j}\, X V_x^- V_{xy}\, {}^t[(V_y M_y)^{1/2}]\, U^j = \frac{1}{\sqrt{\lambda_j}}\, V^j$; $U^j = \chi_y^2\, (V_y M_y)^{1/2}\, {}^t Y D_n \widetilde{V}^j$.

On montre que les deux analyses ont les mêmes valeurs propres non nulles $\{\lambda_j\ /j = 1, r\}$; égalité des inerties : $I(N_y^x) = I(N_g) = trace(\, V_{xy}\, M_y\, V_{yx}\, V_x^-\,) = trace(\, V_{yx}\, V_x^-\, V_{xy}\, M_y\,) = I(N_x^y) = \sum_{j=1}^{r} \lambda_j$.
En particulier, pour $M_y = \chi_y^2$ (cas symétrique), l'inertie expliquée $I(N_y^x)$ [resp. $I(N_x^y)$] est égale au rapport de corrélation généralisé (AD classique) [resp. à la trace de Pillai (MANOVA)]. On déduit aisément les formules de passage des assertions 2) et 3), compte tenu de ce que u_j [resp. v_j] est normé pour V_x^- [resp. M_y].
L'assertion 4) concerne les représentations simultanées barycentriques dans les plans principaux de l'ACP(1).

Le coefficient d'association dissymétrique entre les variables quantitatives explicatives $\{x^j\ ;\ j = 1, p\}$ et la variable de groupe y, relativement à $M_y(\alpha)$, est mesuré par la quantité : $Q_{(x \mapsto y)} = \frac{I(N_y^x)}{I(N_y)} = \frac{trace(\, V_{xy}\, M_y(\alpha)\, V_{yx}\, V_x^-\,)}{trace(\, V_y\, M_y(\alpha)\,)}$.

Il correspond au critère du rapport d'inertie expliquée et englobe le cas symétrique, dans le cas particulier où $M_y(\alpha) = \chi_y^2$, il est alors égal à $trace(\, V_{xy}\, \chi_y^2\, V_{yx}\, V_x^-)/(q-1)$.

Cette approche consiste donc à rechercher le "meilleur" produit scalaire $M_y(\alpha)$, noté $M_y^\star$, qui maximise au mieux les critères usuels du rapport d'inertie expliquée et du pourcentage d'individus bien classés. On propose d'appeler l'ACP(1) "Analyse Discriminante Dissymétrique" (ADD), relativement à $M_y^\star$.

3. Exemple d'application - Vins de Bordeaux

Pour évaluer l'intérêt de l'approche proposée, nous reprenons l'exemple analysé par Tenenhaus dans [TEN 05], concernant les données météo. Quatre variables : Température (somme des températures moyennes journalières en degrès Celsus), Soleil (durée d'insolation en heures), Chaleur (nombre de jours de grande chaleur) et Pluie (hauteur des pluies en millimètres), ainsi qu'une variable de groupe : qualité du vin (1-bonne, 2-moyenne, 3-médiocre), ont été mesurées sur 34 années (1924-1957). Toutes les analyses ont été réalisées sur les variables météo centrées-réduites. l'objectif est de mettre en évidence les caractéristiques météo qui différencient et séparent au mieux les trois qualités. L'échantillon de base de cet exemple étant petit, nous n'avons donc pas choisi d'échantillon-test.

TAB. 1. *Choix du produit scalaire M_y.*

	AD	ADD	→				
α	0	1/3	2/3	1 0	1/3	2/3	1
$M_y(\alpha)$	χ_y^2	${}^1M_y(\frac{1}{3})$	${}^1M_y(\frac{2}{3})$	I_q	${}^2M_y(\frac{1}{3})$	${}^2M_y(\frac{2}{3})$	$\mathbf{D}_q$
$Q_{(x \mapsto y)}$ (%)	**44.40**	44.49	44.61	44.93	45.01	45.15	**45.48**
Bien classés (%)	**70.59**	73.53	79.41	79.41	82.35	82.35	**85.29**

Le tableau 1 donne les résultats des critères d'inertie expliquée et de classement, de l'AD et de six ADD selon le choix du produit scalaire $M_y(\alpha)$ généré. Pour ces données, on obtient de meilleurs pourcentages des critères avec ${}^2M_y(1) = M_y^\star = D_q$.

Le tableau 2 présente les moments d'inerties des facteurs discriminants de l'AD et de la "meilleure" ADD obtenue avec $M_y = D_q$.

TAB. 2. *Inerties et moments d'inerties.*

Facteur	F^1	F^2	**AD**	Inerties	**ADD**	Facteur	F^1	F^2
Valeur propre	0.766	0.122	0.888	B-Inter classes	0.101	Valeur propre	0.088	0.013
% Inertie	86.29	13.71	1.112	W-Intra classe	0.121	% Inertie	87.10	12.90
% Cumulé	86.29	100.00	2.000	T-Totale	0.222	% Cumulé	87.10	100.00

TAB. 3. *Tableaux de classement croisant qualité observée et prédite.*

AD : 70.59%	1	2	3	Total	**Logistique : 79.41%**	1	2	3	Total
1 : Bonne	**8**	3	0	11	1 : Bonne	**8**	3	0	11
2 : Moyenne	3	**7**	1	11	2 : Moyenne	2	**8**	1	11
3 : Médiocre	0	3	**9**	12	3 : Médiocre	0	1	**11**	12
Total	10	12	12	34	Total	10	12	12	34
ADD : 85.29%	1	2	3	Total	**Logistique-PLS : 82.35%**	1	2	3	Total
1 : Bonne	**11**	0	0	11	1 : Bonne	**9**	2	0	11
2 : Moyenne	3	**7**	1	11	2 : Moyenne	1	**8**	2	11
3 : Médiocre	0	1	**11**	12	3 : Médiocre	0	1	**11**	12
Total	10	12	12	34	Total	10	12	12	34

A titre de comparaison avec d'autres modèles de prédiction, le tableau 3 présente les matrices de confusion de l'AD et de l'ADD, ainsi que celles des régressions logistique et logistique-PLS données et commentées dans [TEN 05]. On constate que l'ADD donne un meilleur pourcentage de bien classés.

4. Conclusion

L'approche dissymétrique proposée de l'analyse discriminante peut être utile pour enrichir les résultats de l'analyse discriminante ACP(1) par ceux fournis par l'analyse factorielle de l'ACP(2). L'analyse dissymétrique est parfois plus appropriée et s'adapte à la structure des données observées (multicolinéarité des prédicteurs dans l'exemple des vins de Bordeaux). Il serait intéressant de la comparer à l'AD-PLS [NOC 05]. Enfin, cette approche peut être utilisée lorsque les prédicteurs sont qualitatifs ([SAP 77]) ou mixtes ([ABD 09]).

5. Bibliographie

[ABD 09] ABDESSELAM R., Discriminant Analysis on Mixed Predictors, *Data Analysis and Classification : from the exploratory to the confirmatory approach*, In Book Studies in Classification, Data Analysis, Springer, 2009.

[BAK 03] BAKER M., RAYENS W., PLS for discrimination, *Journal of Chemometrics*, vol. 17, 3, 2003, p. 153–197.

[FIS 38] FISHER R., The Statistical Utilization of Multiple Measurements, *Annals of Eugenics*, vol. VIII, 1938, p. 376–386.

[GEO 04] GEOFFREY J. M., Discriminant Analysis and Statistical Pattern Recognition, rapport n^{o}526, 2004, Wiley Series in Probability and Statistics, New Ed.

[NOC 05] NOCAIRI H., QANNARI E.M., VIGNEAU E., BERTRAND D., Discrimination on latent components with respect to patterns. Application to multicollinear data, *Computational Statistics Data Analysis*, vol. 48, 2005, p. 139–147.

[PAL 99] PALM R., L'analyse de la variance multivariée et l'analyse canonique discriminante : principes et applications, rapport n^{o}40, 1999, Gembloux, Presses Agronomiques.

[SAP 77] SAPORTA G., Une méthode et un programme d'analyse discriminante sur variables qualitatives, *Analyse des Données et Informatique*, INRIA, France, 1977.

[TEN 05] TENENHAUS M., La régression logistique PLS, *Modèles statistiques pour données qualitatives*, Editions Technip, J.J. Droesbeke, M. Lejeune, G. Saporta, Editeurs., 2005, p. 263–276.

Classification supervisée avec second étage optionnel pour variables de covariance conditionnelle hétérogène

Thomas Burger, Thierry Dhorne

Université Européenne de Bretagne, Université de Bretagne-Sud, CNRS, Lab-STICC, Centre de Recherche Yves Coppens, BP 573, F-56017 Vannes cedex, FRANCE
{prenom}.{nom}@univ-ubs.fr – http ://www-labsticc.univ-ubs.fr/~burger/

RÉSUMÉ. *En informatique, il est fréquent qu'un problème de classification fasse intervenir des variables de nature hétérogène, dont les différentes covariances (conditionnellement à la classe) n'ont pas le même ordre de grandeur. Si la procédure d'apprentissage supervisé est basée sur des modèles génératifs, certaines variables, de faible covariance, mais discriminantes peuvent ne pas être suffisamment considérées. Nous présentons un schéma de classification à deux étages permettant de ne pas perdre l'information discriminante issue de ces variables sous-représentées durant l'apprentissage. Le second étage est optionnel : il n'est utilisé que lorsqu'une information pertinente est manquante pour la classification, et que celle-ci peut provenir de variables sous-représentées. La difficulté du problème est de déterminer automatiquement dans quels cas le second étage doit être utilisé. Pour résoudre ce problème, nous proposons de travailler dans le formalisme des fonctions de croyance (FC), et d'utiliser une procédure de décision basée sur une généralisation de la transformée pignistique.*

MOTS-CLÉS : *Classification de données multimodales/multimédia, apprentissage automatique, modèles génératifs, fonctions de croyance, transformée pignistique*

1. Introduction

En informatique, les problèmes de classification supervisée ont des spécificités [DUD 01] qui rendent parfois l'utilisation des méthodes statistiques classiques inadaptées. Ainsi, en reconnaissance de forme (vision par ordinateur, compression MPG7), en reconnaissance de la parole [RAB 89], en indexation automatique de contenus multimédia [MUL 07], nous sommes confrontés aux problèmes suivants :

- bruit des données (dû à la compression de l'image, à la qualité des capteurs sonores/vidéos, etc).
- besoin de généricité par rapport au nombre de classes.
- contraintes machines : sensibilité des conditions d'acquisitions, calcul temps-réel, bufferisation, etc.
- nature hétérogène des variables de classification (il est en effet fréquent de mélanger des descripteurs de couleur, de texture, de forme, etc.), dont la prise en compte de manière conjointe est difficile.

En raison de tout cela, il est classique d'utiliser des méthodes génératives plutôt que discriminantes [ARA 06] : HMM avec observation gaussiennes, ACP et modèles des classes basés sur des mélanges gaussiens, etc..

Cependant, l'intérêt des méthodes génératives pour la réduction du bruit peut aussi devenir un inconvénient. Les modèles génératifs ne considèrent que les grandes tendances des échantillons d'apprentissage, et "lissent" le reste en perdant une partie de l'information. Or, dans certains cas, c'est cette information perdue qui permet la séparation entre les classes. En pratique, cela peut en particulier arriver si, conditionnellement à chaque classe, les matrices de covariance sont différentes. Et cela arrive de manière fréquente lorsque les variables sont issus de différents jeux de descripteurs de nature hétérogène.

Dans [ARA 07, ARA 08, ARA 09], nous nous sommes intéressés à la reconnaissance automatique de gestes multimodaux de la Langue des Signes Américaine (ASL). Tous les problèmes énumérés ci-dessus ont été rencontrés. Dans ces articles, nous avons apporté plusieurs éléments de solutions, et plusieurs variantes. Néanmoins :

– La méthode de classification proprement dite n'est ni isolée de son application, ni des autres traitements informatiques n'ayant pas de rapport avec la classification. Elle n'est pas isolée non plus des prétraitements que nous avons dû utiliser pour standardiser le problème vis-à-vis des autres méthodes de l'état de l'art, et ainsi proposer un protocole de comparaison rigoureux.
– Les deux transformations mathématiques appliquées aux structures d'information mise en jeu dans l'algorithme de classification ne sont pas justifiées d'un point de vue théorique. Depuis, la première a même été améliorée, et cette amélioration correspond au travail publié en parallèle par une autre équipe de recherche [DUB 08]. La seconde a été formalisée, justifiée et publiée depuis [BUR 09].

Cet article a donc pour objectif de détailler les aspects méthodologiques qui manquent à nos travaux sur l'ASL, en s'appuyant sur les résultats expérimentaux de [ARA 07, ARA 08, ARA 09], les résultats théoriques de [DUB 08, BUR 09], et de replacer le tout dans un contexte statistique appliquée à l'informatique. Dans la deuxième partie, nous fournissons les éléments théoriques de manipulation des fonctions de croyance nécessaires à la méthode. Dans la troisième partie, nous décrivons la méthode de classification et nous résumons les résultats expérimentaux.

2. conversions entre probabilités et fonctions de croyance

Nous supposons le lecteur familier de la théorie de Dempster-Shafer et des fonctions de croyance (FC). Ainsi, nous ne reprenons pas les définitions de base, et celles-ci peuvent être trouvées dans [DEM 68, SHA 76, SME 94].

Dans de nombreuses situations, il est nécessaire de passer du formalisme probabiliste au formalisme des FC, et vice-versa. En fonction de ce que les probabilités et les FC modélisent, il y a plusieurs manières d'effectuer ces conversions [COB 03], et le choix de certaines plutôt que d'autres fait débat. Dans notre cas, nous retiendrons la conversion d'une probabilité en une fonction de croyance consonante décrite dans [DUB 08]. Soient p une fonction de probabilité sur Ω, $\{h_1, \ldots, h_N\}$ l'ensemble des éventualités de Ω ordonnés par valeurs de probabilité décroissante, et m la FC correspondant au résultat de la conversion. On a m nulle partout, sauf pour les éléments focaux du type $\{h_1, \ldots, h_k\}$, pour lesquels, on a : $m(\{h_1, \ldots, h_k\}) = k \times [p(h_k) - p(h_{k+1})]$.

A l'inverse, nous considérons la conversion d'une FC en une probabilité grâce à la transformée pignistique [SME 94]. Dans [BUR 09], nous proposons une généralisation de la transformée pignistique, dont le résultat n'est pas nécessairement une probabilité. Elle dépend d'un paramètre γ, mais dans le cas où $\gamma = 1$ nous retombons sur la transformée pignistique originale. Nous avons :

$$\mathbb{B}_\gamma(B) = m(B) + \sum_{\substack{B \subset A \subseteq \Omega \\ A \notin \Delta_\gamma}} \frac{m(A) \cdot |B|}{N(|A|, \gamma)} \quad \forall B \in \Delta_\gamma \qquad \text{avec} \qquad N(|A|, \gamma) = \sum_{k=1}^{\gamma} \frac{|A|!}{k!(|A| - k)!} \cdot k \tag{1}$$

avec $\mathbb{B}_\gamma$ désignant le résultat de la transformée de m, $|A|$ désignant le cardinal de A, et Δ_γ désignant l'ensemble des éléments de 2^Ω de cardinal inférieur ou égale à γ. L'intérêt de la transformée pignistique est classiquement de convertir une FC en une probabilité juste avant la prise de décision, de manière à prendre une décision conforme à la notion classique de pari en théorie des jeux. Dans le cas où $\gamma \neq 1$, la généralisation proposée entraîne une décision qui ne sera pas forcement focalisée sur une seule éventualité, mais sur plusieurs (au maximum γ), parmi lesquelles l'hésitation est justifiée. Par contre, le nombre d'éventualités ainsi retenu ne sera pas forcément de γ si le problème ne le justifie pas. Ainsi, on sort du cadre de la théorie des jeux, comme cela est justifié dans [BUR 09].

3. Schéma de classification à second étage optionnel

Considérons d'abord le schéma classique d'une classification basée sur des modèles génératifs. Notons $C_1, \ldots, C_K$ les K classes du problème. Pour chaque classe C_q, un modèle génératif, noté G_q, a été appris sur un échantillon représentatif. Ensuite, chaque individu I_i à classer est considéré : on calcule sa vraisemblance pour chacun des modèles G_q. Cela permet d'obtenir un ensemble de K vraisemblances $\mathcal{L}^i(G_q), \forall q \leq K$. De manière clas-

sique, I_i est classée selon la méthode du maximum a posteriori (MAP), c'est à dire qu'il est associé à la classe $C_* = \text{argmax}_{(C_q)}(\mathcal{L}^i(G_q))$.

Nous proposons de raffiner cette méthode : dans un premier temps, une étude sommaire des corpus de données doit permettre d'évaluer quelles sont les variables susceptibles d'être perdues durant la génération des modèles. Par exemple, il est classique d'avoir un premier jeu de descripteurs d'une nature donnée (de formes, par exemple) dont la prise en compte dans les modèles génératifs est minimisée par un autre jeu de descripteurs (de couleurs par exemple). Néanmoins, il est à première vue difficile de déterminer dans quelles proportions les informations issues du premier jeu de descripteurs vont être perdues. Par souci de notations simples, notons V_p l'ensemble des variables qui sont potentiellement **p**répondérantes, et V_l, l'ensemble de celles dont l'influence risque d'être **l**issée dans les modèles. Ensuite, pour chaque classe C_q, non plus un, mais deux modèles génératifs sont appris : le premier, $G_q^{p,l}$ sur l'ensemble $\{V_p, V_l\}$, et le second G_q^l, sur V_l uniquement. Le premier étage de classification consiste à calculer pour tout individu I_i les vraisemblances $\mathcal{L}^i(G_q^{p,l})$ des modèles $G_q^{p,l}$. Deux situations sont possibles :

- L'une des vraisemblances est clairement plus élevée que les autres. La décision est donc facile à prendre. Cela signifie que l'information était suffisante pour mener à bien la tâche de classification.
- Plusieurs vraisemblances ont le même ordre de grandeur, et le choix de l'une par rapport à l'autre peut sembler arbitraire. Il est donc nécessaire de vérifier si une information provenant de V_l, déterminante pour la classification, n'a pas été perdue durant la génération des modèles $G_q^{p,l}$. Dans un tel cas, il est donc judicieux de ne considérer que les quelques classes dont les vraisemblances sont suffisamment élevées, et de procéder à un second tour parmi celles-ci, en n'utilisant cette fois-ci, que V_l.

Cette stratégie, relativement simple sur le principe a le mérite de permettre de récupérer l'information manquante pour finaliser la classification, comme dans les méthodes de boosting. Par contre, à l'inverse du boosting, elle est générique par rapport au nombre de classes : Ainsi, quand une nouvelle classe (la $K+1$ième) doit être ajoutée au problème (un mot supplémentaire en reconnaissance de la parole, un nouveau visage dans un système d'identification, etc.), il suffit de fournir les deux modèles $G_{K+1}^{p,l}$, G_{K+1}^l, et cela ne remet pas en cause les modèles des autres classes. Cependant, quelle stratégie de décision adopter pour pouvoir automatiquement arbitrer entre les deux situations décrites plus haut ? En effet, il est difficile de décider automatiquement dans quels cas on peut considérer qu'une unique classe ressort des comparaisons de vraisemblance, et dans quels cas, un second étage est nécessaire.

A la fin du premier étage, plutôt qu'une stratégie du MAP, nous proposons la stratégie suivante : Nous normalisons les K vraisemblances en les divisant par la somme des vraisemblances, afin que leur somme valle 1. Ainsi, après cette normalisation, nous avons une distribution de probabilité subjective modélisant l'appartenance à chacune des classes. Ensuite, cette probabilité est convertie en une FC par la méthode indiquée dans la partie 2 [DUB 08]. Enfin, la généralisation de la transformée pignistique est utilisée [BUR 09]. Si la quantité d'information est suffisante pour qu'une seule classe ressorte, alors la décision prise à l'issue du premier étage sera la même qu'avec le MAP, et la procédure se terminera ainsi. En revanche, pour les cas limites, un sous-ensemble restreint de K' classes (avec $K' \leq \gamma$), parmi lesquelles il est difficile d'opérer une discrimination, est retenu. Dès lors, le second étage de classification est utilisé : Comparaison des K' modèles G_q^l, et prise de décision avec la stratégie du MAP.

Le dernier point à discuter est le choix de la valeur de γ. Il y a plusieurs stratégies possibles : Celle qui consiste à bien étudier le jeu de données, afin d'en connaître les spécificités, et déterminer le nombre maximal de classes parmi lesquelles une hésitation est possible (nous appelons un tel regroupement de classe, un cluster). Cette stratégie exige une connaissance précise de la topologie des données qu'il est en pratique difficile d'avoir dans certains problèmes en informatique. C'est pourquoi, une méthode plus automatique peut être préférée. En théorie, il est envisageable d'effectuer une validation croisée sur toutes les valeurs possibles de γ, et de choisir la valeur donnant le meilleur résultat. En pratique, cela n'est ni très élégant, ni robuste à l'augmentation du nombre de classes dans le problème. Dès lors, une solution intermédiaire consiste à définir de manière automatique des clusters au sein desquels une hésitation a un sens. Dès qu'une décision incomplète surgie au premier étage de classification, le second étage doit discriminer entre toutes les classes du cluster contenant l'hésitation. Ainsi, la valeur de γ n'a en pratique plus d'importance, et la valeur $\gamma = 2$ fait parfaitement l'affaire, quelque soit la taille du cluster.

C'est cette dernière solution que nous préconisons (si le problème le permet) et que nous avons utilisée pour la reconnaissance de l'ASL. Dans ce problème, la reconnaissance de chaque signe de l'ASL est effectuée au moyen d'un HMM, V_p représente les descripteurs des mouvements des mains, et V_l représente les descripteurs des mouvements des autres parties du corps (mouvements de têtes, expression faciale, etc.). Cette méthode nous a permis d'obtenir des taux de classifications meilleurs que les méthodes classiques :

- 23.7% des erreurs sont évitées par rapport au schéma de base indiqué plus haut, correspondant à la classification systématique par MAP à l'issu du premier étage.
- 31.1% des erreurs sont évitées par rapport à l'utilisation conjointe des deux modèles $G^{p,l}$ et G^l et de leur mise en cascade systématique (second étage utilisé de manière systématique). En effet, dans un tel cas, certaines décisions correctes à l'issu du premier étage sont remises en cause au niveau du second.
- 37.4% des erreurs sont évitées par rapport à l'utilisation conjointe de deux modèles G^p et G^l en parallèle (et non en cascade) et de la fusion systèmatique des vraisemblances.

4. Conclusion

Nous proposons dans cet article une méthode de classification ayant un second étage optionnel, permettant de s'affranchir de nombreux défauts de la classification supervisée par modèles génératifs. Cette méthode est basée sur une méthode de décision originale nécessitant de transformer l'information probabiliste issue des modèles de classes en une fonction de croyance. La justification de la méthode est basée sur la validation théorique de transformations mises en jeu et de son utilisation dans le cas d'un problème réel, la reconnaissance de gestes de la Langue des Signes. La suite de ce travail consiste en application de cette méthode à de nouveaux problèmes.

5. Bibliographie

[ARA 06] ARAN A., AKARUN L., Recognizing Two Handed Gestures with Generative, Discriminative and Ensemble Methods via Fisher Kernels, *Lecture Notes in Computer Science : Multimedia Content Representation, Classification and Security International Workshop, MRCS 2006*, , 2006, p. 159–166.

[ARA 07] ARAN O., BURGER T., CAPLIER A., AKARUN L., Sequential Belief-Based Fusion of Manual and Non-Manual Signs, *Gesture Workshop (GW'07)*, 2007.

[ARA 08] ARAN O., BURGER T., CAPLIER A., AKARUN L., A Belief-Based Sequential Fusion Approach for Fusing Manual and Non-Manual Signs, *Pattern Recognition*, vol. 42(5), 2008, p. 812–822.

[ARA 09] ARAN O., BURGER T., CAPLIER A., AKARUN L., Sequential Belief-Based Fusion of Manual and Non-Manual Information for Recognizing Isolated Signs, *Gesture-Based Human-Computer Interaction and Simulation, Sales Dias, M. ; Gibet, S. ; Wanderley, M.M. ; Bastos, R. (Eds.), LNCS/LNAI*, vol. 5085, 2009, Springer.

[BUR 09] BURGER T., CAPLIER A., A Generalization of the Pignistic Transform for Partial Bet, *accepted to ECSQARU'09*, July 2009.

[COB 03] COBB B. R., SHENOY P., A Comparison of Methods for Transforming Belief Functions Models to Probability Models, *Lecture Notes in Artificial Intelligence*, vol. 2711, 2003, p. 255–266.

[DEM 68] DEMPSTER A., A generalization of Bayesian inference, *Journal of the Royal Statistical Society, Series B*, vol. 30(2), 1968, p. 205–247.

[DUB 08] DUBOIS D., PRADE H., SMETS P., A definition of subjective possibility, *International Journal of Approximate Reasoning*, vol. 48(2), 2008, p. 352–364.

[DUD 01] DUDA R., HART P., STORK D., *Pattern Classification*, Wiley, 2001.

[MUL 07] MULHEM P., QUÉNOT G., BERRUT C., Recherche d'information multimédia, *Patrick Gros, éditeur, L'indexation multimédia - Descriptions et recherche automatiques*, , 2007, p. 25–49.

[RAB 89] RABINER L., A tutorial on hidden Markov models and selected applications in speech recognition, *proceedings of the IEEE*, vol. 7(2), 1989, p. 257–286.

[SHA 76] SHAFER G., *A Mathematical Theory of Evidence*, Princeton University Press, 1976.

[SME 94] SMETS P. K. R., The transferable belief model, *Artificial Intelligence*, vol. 66(2), 1994, p. 191–234.

Discrimination sur des données arborescentes

David Grosser, Henry Ralambondrainy, Noël Conruyt

IREMIA, Université de la Réunion
15 av Cassin, 97715 Saint-Denis Message Cedex 9, Réunion
(conruyt,grosser,ralambon)@univ-reunion.fr

RÉSUMÉ. *La formalisation de connaissances descriptives issues des sciences du vivant, par les experts en biologie et en Systématique, produit des représentations arborescentes dont les noeuds peuvent être présents, absents, ou inconnus. Afin d'améliorer la robustesse du processus de classification de ces objets complexes, souvent partiellement décrits, nous proposons une méthode de discrimination itérative, interactive et semi-dirigée. Elle combine des techniques inductives pour le choix de variables discriminantes et la recherche de plus proches voisins pour la reconnaissance des classes. La méthode s'appuie sur nos précédents travaux, où nous avons défini différents indices de similarité qui prennent en compte à la fois la structure et les valeurs des objets pour la recherche de voisins.*

MOTS-CLÉS : *Classification, discrimination, similarité, données structurées*

1. Introduction

Outre l'étude des liens de parenté entre espèces (phylogénétique), la description de *spécimens* biologiques à des fins d'identification et de classification est une part essentielle du travail des systématiciens (phénétique). L'automatisation de ce processus par des outils informatiques dans le but de construire des systèmes classificatoires pose d'intéressants problèmes de représentation et de traitement de la connaissance. Ceci est particulièrement vrai en biologie marine et pour certains taxons comme les coraux, où la nature polymorphe des individus, les colonies coralliennes, rend difficile leur description par des représentations classiques de type attributs-valeurs, du fait notamment des relations de dépendances entre caractères induits par une forte variabilité morphologique. Le modèle de représentation par descriptions arborescentes que nous avons développé offre un cadre mieux adapté pour représenter ces objets biologiques complexes, permettant d'exprimer les relations de présence-absence entre composants d'une observation. Ce modèle permet ainsi de représenter des descriptions structurées dont les caractères sont exprimés par des valeurs complexes, de type ordinal ou continu, ensemble ou intervalle, qui peuvent être manquantes ou inconnues.

Ainsi, les méthodes classiques de discrimination développées dans le cadre de l'analyse discriminante ou de l'apprentissage, telles que les arbres de discrimination ou de décision, sont rarement satisfaisantes dans ce cadre, car elles ne prennent en compte ni les relations entre attributs, ni les données manquantes. Elles sont également très peu tolérantes aux erreurs. Le problème que nous considérons est celui de la détermination de la classe d'appartenance d'une description arborescente partiellement renseignée et possédant éventuellement des erreurs, à partir d'une base de cas de références partitionnées en k-classes *a priori*, qualifiées par des experts. La méthode de discrimination proposée procède par inférence de voisinages successifs. Elle est inductive, interactive, itérative et semi-dirigée. Elle combine des techniques d'induction de variables discriminantes et de recherche de voisins à l'aide d'un indice de similarité qui prend en compte la structure et le contenu des descriptions [GRO 08]. Elle s'appuie sur l'espace des descriptions, structuré en inf demi-treillis pour garantir la cohérence des descriptions et le calcul de descriptions généralisantes. La méthode est en cours de développement au sein du Système de Gestion de Base de Connaissances IKBS.

2. Représentation des descriptions

Au sein d'une base de connaissances, les observations sont décrites à l'aide de *modèles descriptifs*. Un modèle descriptif représente une connaissance d'ordre ontologique, portant sur la structure et l'organisation des descripteurs associé à une classe d'individus. Il est constitué principalement d'une arborescence $\mathcal{M} = (\mathcal{A}, \mathcal{U})$, où $\mathcal{A}$ est un ensemble d'attributs formant les sommets de l'arbre. Les noeuds terminaux sont des attributs classiques, de type qualitatif ou numérique. Les noeuds intermédiaires sont des "attributs structurés" (ou *composants*) notés : $A_j :< A_1, \ldots, A_p >$ où A_j est la racine du sous-arbre dont les fils sont $A_1, \ldots, A_p$, des attributs structurés ou simples. Une arête $(A_j,\ A_l) \in \mathcal{U}$ exprime une relation de dépendance entre A_j et A_l. Un modèle descriptif peut également être utilisé comme support à la représentation des classes d'individus (ou taxons) et mis en relation à des Ontologies plus vastes du domaine.

2.1. *Structure d'une description*

Une observation est une arborescence dérivée de celle de A (appelée *squelette*) pour laquelle les attributs simples ont été valués par un élément du domaine. Un *squelette* décrit la structure morphologique d'une observation, il renseigne le statut de chaque composant, dont l'état peut être présent (+), absence (-) ou inconnu ($*$). Notons $S = \{+, -, *\}$, Une application $\sigma : \mathcal{A} \rightarrow S$ définit le squelette H_σ par l'arborescence annotée par S : $H_\sigma = (\mathcal{A}_\sigma, \mathcal{U})$ avec $\mathcal{A}_\sigma = \{(A_j, \sigma(A_j))_{j \in J}\}$, avec $A_j \in A$. Sur l'ensemble des squelettes les règles de cohérence suivantes sont imposées. Pour tout attribut structuré $B :< B_l >_{l \in L}$, on a :

1. "Les fils d'un noeud absent sont absents" : $\sigma(B) = -$ alors $\sigma(B_l) = -$ pour $l \in L$,
2. "Les fils d'un noeud inconnu sont inconnus ou absents" : $\sigma(B) = *$ alors $\sigma(B_l) = *|-$ pour $l \in L$.
3. "Les fils d'un noeud présent peuvent être présents, absents ou inconnus" : $\sigma(B) = +$ alors $\sigma(B_l) = +|-|*$ pour $l \in L$.

On note $\mathcal{H}$ l'ensemble des squelettes vérifiant les règles de cohérence précédentes. Les valeurs d'une observation sont relatives aux attributs simples (les feuilles). Un attribut simple A_q est valué dans son domaine D_q s'il est présent dans une observation. Lorsqu'il est absent, on lui attribuera la valeur $\perp$ (valeur inapplicable) et $*$ lorsqu'il n'est pas renseigné. On note $\Gamma_q = D_q \cup \{\perp\} \cup \{*\}$, et $\Gamma = \sqcap_{q \in Q} \Gamma_q$. L'ensemble des valeurs d'une observation o est noté $v_o = (v_q)_{q \in Q} \in \Gamma$ avec $v_q \in \Gamma_q$. Une observation est décrite par son squelette et ses valeurs : $o = (H_{\sigma_o}, v_o) \in \mathcal{E} = \mathcal{H} \times \Gamma$.

2.2. *Inf demi-treillis sur les descriptions*

On munit S de l'ordre suivant $* < +, * < -$, qui s'interprète : une valeur présente ou absente est plus précise qu'inconnue. On a $* = + \wedge -$, S est un inf demi treillis $(S; <, \wedge)$. Il s'ensuit que l'ensemble des fonctions $S^A = \{\sigma : \mathcal{A} \rightarrow S\}$ est aussi un inf demi treillis. L'ensemble des squelettes $(\mathcal{H}; <, \wedge)$ est donc un inf demi treillis tel que pour $H_{\sigma_1}, H_{\sigma_2} \in \mathcal{H}$, on ait : $H_{\sigma_1} < H_{\sigma_2} \Longleftrightarrow \sigma_1 < \sigma_2$, et $H_{\sigma_1 \wedge \sigma_2} = H_{\sigma_1} \wedge H_{\sigma_2}$. Il est facile de vérifier que $\mathcal{H}$ est stable pour l'opérateur $\wedge$ cad que l'inf de deux squelettes cohérentes reste cohérent.

Chaque domaine $(D_q; <, \wedge, *)$ est supposé muni d'une structure d'inf demi treillis qui dépend du type de l'attribut simple concerné. On considère $(\Gamma_q; <, \wedge)$ comme l'inf demi treillis dont le plus petit élément est $*$ et l'élément $\perp$ est incomparable avec les $v_q \in D_q$, cad $v_q \wedge \perp = *$. L'ensemble $(\Gamma = \sqcap_{q \in Q} \Gamma_q; <, \wedge)$ est alors un inf demi treillis comme produit d'inf demi treillis. L'espace de description des observations est l'inf demi treillis produit : $(\mathcal{E} = \mathcal{H} \times \Gamma; <, \wedge)$. L'inégalité entre deux descriptions $e < f$ s'interprète comme suit : la description e est plus générale que f sur le plan structurel et concernant les valeurs.

Les applications extension $ext : \mathcal{E} \longrightarrow \mathcal{P}(O)$ et intension $int : \mathcal{P}(O) \longrightarrow \mathcal{E}$ sont définis usuellement dans un inf demi treillis : $ext(e) = \{o \in O | e \leq o\}$ l'extension de la description e est l'ensemble des observations qu'elle reconnait. L'intension de $L \subset O$ est la description qui généralise les observations de L : $int(L) = \bigwedge_{o \in L} o = (\wedge_{o \in L} H_{\sigma_o}, \wedge_{o \in L} v_o)$.

2.3. *Indices de similarité sur les observations*

Nous avons proposé dans [GRO 08] différents indices de similarité qui prennent en compte à la fois les aspects structurels et valeurs d'observations arborescentes. La similitude structurelle $\zeta_S \in [0,1]^{\mathcal{H}\times\mathcal{H}}$ de deux squelettes est évaluée comme la moyenne des valeurs de similitude de leurs noeuds. Elle est calculée de manière récursive à partir d'une fonction de comparaison sur le statut des noeuds. Une mesure de similarité étant donnée sur chaque domaine d'un attribut simple, un indice de similarité dite "locale" $\zeta_L \in [0,1]^{\Gamma\times\Gamma}$ entre deux observations est calculable sur les attributs simples présents en commun. On définit alors un indice de similarité globale $\zeta_G \in [0,1]^{\mathcal{E}\times\mathcal{E}}$ qui tient compte à la fois de la structure et du contenu. Il est calculé de manière récursive ou comme la moyenne des deux indices précédents. On note l'indice de distance $d = 1 - \zeta_G$ déduite de l'indice de similarité globale.

3. Discrimination par voisinages successifs

La méthode de discrimination proposée permet de déterminer l'appartenance à une classe d'un individu partiellement décrit par un utilisateur, et pouvant comporter des erreurs, en tenant compte des relations de dépendances entre attributs. Le principe consiste à sélectionner un voisinage V, i.e. un ensemble de descriptions (spécimens ou classes) proches de la description courante, à l'aide de l'indice de distance d. Un ensemble de classes candidates est calculée à partir de l'ensemble des voisins. La méthode cherche ensuite à compléter l'information manquante, d'une part par l'application de règles de cohérences et d'autre part en proposant à l'utilisateur un ensemble d'attributs discriminants qu'il peut renseigner. Un nouveau voisinage est recalculé sur la base de la nouvelle description partielle. Le processus est réitéré jusqu'à obtention d'un ensemble quasi-homogène de descriptions.

3.1. *Cohérence des descriptions*

Les descriptions sont exprimées dans l'espace $(\mathcal{E} = \mathcal{H} \times \Gamma; <, \wedge, d, ext, int)$. On note $e \in \mathcal{E}$ la description partielle qui vérifie les règles de cohérences suivantes :

1. Les descendants d'un noeud non renseigné sont notés "$*$",
2. Les descendants d'un noeud absent sont notés "$-$",
3. Les ancêtres d'un noeud présent ou d'une feuille valuée sont notés "$+$".

Si d'autres règles d'inférence liées aux domaines sont disponibles, elles sont appliquées à ce stade.

3.2. *Algorithme de discrimination*

L'algorithme de détermination de la classe d'appartenance est itératif et comporte trois étapes, à l'itération m, on considère une description e_m.

3.2.1. ***Etape 1*** *: Calcul du voisinage*

L'ensemble des voisins de e_m à l'itération m est obtenu par $V_{(m)} = \{o \in O \mid d(e_m, o) < \Delta_m\}$, où Δ_m est un seuil recalculé à chaque itération et d la mesure proposée au 2.3.

On note $D(e, A)$ une mesure de proximité de e à un ensemble A qui sera dans notre cas :

$$D_{max}(e, A) = max_{a\in A} d(e, a)$$

où $D_{moy}(e, A) = \frac{1}{|A|}\Sigma_{o\in A} d(e, o)$, soit la distance maximum ou moyenne des éléments de A à e, ou encore l'inertie de A par rapport à e : $D_{disp}(e, A) = \frac{1}{|A|}\Sigma_{o\in A} d^2(e, o)$ qui mesure la dispersion de A autour de$e$. Ces mesures sont ordonnées comme suit : $D_{disp} < D_{moy} < D_{max}$. On définit pour un choix de D donné :

$$\Delta_m = D(e_{m-1}, V_{(m-1)})$$

3.2.2. ***Etape 2*** *: Classes candidates*

Notons $\{C_l\}_{l\in K}$ l'ensemble des classes à priori et $O^l_{(m)} = C_l \cap V_{(m)}$. Soit $f^l_{(m)} = \frac{|O^l_{(m)}|}{|V_{(m)}|}$ la fréquence relative de classe C_l dans l'ensemble de voisins $V_{(m)}$. Les classes candidates à l'affectation du spécimen proposées à l'utilisateur sont les classes $O^l_{(m)}$ telles que leurs fréquences relatives $f^l_{(m)}$ sont significativement différentes de leurs fréquences absolues dans la population qui sont les $f^l_{(0)}$. Les classes sont présentées dans l'ordre de leur degré de proximité $D(e_m, O^l_{(m)})$.

3.2.3. ***Etape 3*** *: Sélection d'attributs discriminants*

Une liste ordonnée de variables informatives est calculée à chaque étape du processus d'identification. Le premier élément est exposé sous forme de question à l'utilisateur qui peut choisir une variable alternative de la liste ou apporter une réponse inconnue. La liste est construite en fonction d'une combinaison de plusieurs critères :

1. Position des attributs dans l'arborescence. La méthode considère les attributs qui peuvent être renseignés, i.e. ceux pour lesquels il existe une chaîne de composants (noeuds) dont la présence est avérée. Les questions relatives à la présence ou l'absence de composants sont également considérées.

2. Pouvoir discriminant. Choix de critère classique de calcul du gain d'information utilisé en apprentissage, tel que l'*entropie de Shannon* ou le *Gini Index*. Ce type de critère permet de minimiser le nombre de questions.

3. Pondération des attributs dans le modèle descriptif. En biologie, certains caractères sont particulièrement difficiles à observer (nécessite du matériel spécifique), à décrire ou sujet à interprétation. Le poids relatif de chaque attribut peut être pondéré par l'expert sur une échelle de 0 à 1. Cette pondération exprime une certaine connaissance, d'ordre stratégique, qui permet d'influer directement sur l'ordre de sélection des attributs.

3.2.4. *Condition d'arrêt*

Le nombre d'attributs susceptible d'être valué étant fini, le processus ne peut se poursuivre indéfiniment. D'autre part, il est facile de montrer que pour les mesures de proximité $D_{max}, D_{moy}, D_{disp}$, la suite de nombres positifs Δ_m est décroissante. Il est alors possible de prendre comme critère d'arrêt, une taille minimale de l'effectif de l'ensemble des voisins, comme dans les méthodes d'arbre de décision.

4. Conclusion

Pour identifier un objet biologique et lui associer un taxon, les systématiciens procèdent la plupart du temps en deux phases. La phase *synthétique*, par observation globale des caractères les plus visibles permet de réduire le champs d'investigation. La phase *analytique*, par observation fine de caractères discriminants permet d'affiner la recherche jusqu'à obtention du résultat. La méthode de discrimination que nous proposons présente l'avantage de correspondre au raisonnement mis en oeuvre par les biologistes. A partir d'une description partielle contenant généralement les caractères les plus visibles ou faciles à décrire, la méthode suggère à l'utilisateur une information pertinente qu'il faut compléter pour déterminer la classe la plus probable. Elle est de plus tolérante aux erreurs dans la mesure ou une information erronée peut quand même aboutir à un résultat satisfaisant du fait que l'on n'effectue pas un filtrage strict sur les caractères renseignés. La méthode est générique et applicable à tout domaine où l'on considère des données structurées (type XML). Il suffit de disposer d'un opérateur de généralisation et d'un indice de proximité adaptés aux données considérées. Cette méthode est en cours d'évaluation sur une base "Coraux des Mascareignes" qui compte environ 150 taxons et 800 descriptions.

5. Bibliographie

[GRO 08] GROSSER D., RALAMBONDRAINY H., Indices de similarité sur des données arborescentes, *Proceedings of First joint meeting of the société francophone de classification and the classification and data analysis group of the italian statistical society, SFC-CLADAG*, 2008, p. 317–320.

Reliability of error estimators in small-sample high-dimensional classification problems

Blaise Hanczar

CRIP5 - Université Paris Descartes
45, rue des Saint-Pères,
75006 PARIS
blaise.hanczar@parisdescartes.fr

RÉSUMÉ. *The validity of a classifier model, consisting of a trained classifier and it estimated error, depends upon the relationship between the estimated and true errors of the classifier. Absent a good error estimation rule, the classifier-error model lacks scientific meaning. This paper demonstrates that in high-dimensionality feature selection settings in the context of small samples there can be virtually no correlation between the true and estimated errors. This conclusion has serious ramifications in the domain of high-throughput genomic classification, such as gene-expression classification, where the number of potential features (gene expressions) is usually in the tens of thousands and the number of sample points (microarrays) is often under one hundred.*

MOTS-CLÉS : *classification, error estimation, small-sample, high dimension.*

1. Introduction

The validity of a classifier model, the designed classifier and the estimated error, depends upon the relationship between the estimated and true errors of the classifier. Model validity is different than classifier goodness A good classifier is one with small error, but this error is unknown when a classifier is designed and its error is estimated from sample data. In this case, its performance must be judged from the estimated error. Since the error estimate characterizes our understanding of classifier performance on future observations and since we do not know the true error, model validity relates to the design process as a whole : What is the relationship between the estimated and true errors resulting from applying the classification and error estimation rules to the feature-label distribution when using samples of a given size ?

Since in the context of sampling we are concerned with the estimation of one random variable, the true error, by another, the estimated error, we naturally would like the true and estimated errors to be strongly correlated. In this paper, we demonstrate that when there is high-dimensionality, meaning a large number of potential features, and a small sample, one should not expect significant correlation between the true and estimated errors. This conclusion has serious ramifications in the domain of high-throughput genomic classification, such as gene-expression classification. For instance, with gene-expression microarrays, the number of potential features (gene expressions) is usually in the tens of thousands and the number of sample points (microarrays) is often under one hundred.

Concerns regarding the validity of microarray-based classification go back to practically the outset of its use [DOU 01], with the accuracy of small-sample error estimation being of particular concern. In particular, attention has been directed at the deleterious effect of cross-validation variance on error estimation accuracy [BRA 04, ISA 08] and the even worse performance of cross-validation in the presence of feature selection [MOL 05]. Whereas the preceding studies have focused on the increased variance of the deviation distribution between the estimated and true errors, here we show that it is the decorrelation of the estimated and true errors in the case of feature selection that lies at the root of the problem.

2. Precision of the error estimation

The precision of an error estimator relates to the difference between the true and estimated errors. We require a probabilistic measure of this difference. We use the root-mean-square error (square root of the expectation of the squared difference), $RMS = \sqrt{E[|\varepsilon_{est} - \varepsilon_{tru}|^2]}$. It is helpful to understand the RMS in terms of the deviation distribution, $\varepsilon_{est} - \varepsilon_{tru}$. The RMS can be decomposed into the bias, $Bias[\varepsilon_{est}] = E[\varepsilon_{est} - \varepsilon_{tru}]$, of the error estimator relative to the true error, and the deviation variance, $Var_{dev}[\varepsilon_{est}] = Var[\varepsilon_{est} - \varepsilon_{tru}]$, namely, $RMS = \sqrt{Var_{dev}[\varepsilon_{est}] + Bias[\varepsilon_{est}]^2}$.

A straightforward expectation computation shows that the deviation variance can be further decomposed into $Var_{dev}[\varepsilon_{est}] = \sigma^2_{est} + \sigma^2_{tru} - 2\rho\sigma_{est}\sigma_{tru}$. For large samples, the designed classifier tends to be stable across samples, so that $\sigma^2_{tru} \approx 0$. If the classification rule is consistent, then the expected difference between the error of the designed classifier and the Bayes error tends to 0. Moreover, popular error estimates tend to be precise for large samples and therefore $\sigma^2_{est} \approx \sigma^2_{tru} \approx 0$. Hence, the right-hand side of the preceding decomposition is close to 0, so that, $Var_{dev}[\varepsilon_{est}] \approx 0$.

The situation is starkly different for small samples. For these, σ^2_{tru} is not small, and σ^2_{est} can be substantially larger, depending on the error estimator. Now correlation plays an important role. For instance, if $\rho \approx 1$, then $Var_{dev}[\varepsilon_{est}] \approx (\sigma_{est} - \sigma_{tru})^2$ but if $\rho \approx 0$, then $Var_{dev}[\varepsilon_{est}] \approx \sigma^2_{est} + \sigma^2_{tru}$. This is a substantial difference when σ^2_{tru} and σ^2_{est} are not small. As we will see, feature selection with small samples will drive down the correlation.

3. Simulation study

3.1. *Experimental design*

Our simulation study uses the following protocol when using feature selection :

1. A training set S_{tr} and test set S_{ts} are generated. For the synthetic data, N examples are created for the training set and 10000 examples for the test set. For the microarray data, the examples are separated into training and test sets with 50 examples for the training set and the remaining for the test set.

2. A feature-selection method is applied on the training set to find a feature subset $\Omega_d(S_{tr})$, where d is the number of selected features chosen from the original D features.

3. A classification rule is used on the training set to build a classifier.

4. The true classification error rate is computed using the test set.

5. Three estimates of the error rate are computed from S_{tr} using the three estimators : leave-one-out, cross-validation, and .632 bootstrap.

This procedure is repeated 10000 times. In the full study, we consider three feature-selection methods, t-test, relief and mutual information, and we consider five classification rules, 3-nearest-neighbor (3NN), linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), linear support vector machine (SVM), and decision trees (CART) [HAN 07]. For cross-validation we use 5 runs of 5-fold cross-validation and for .632 bootstrap we do 100 replications. In the case of no feature selection, step 2 is omitted because we begin with the d features to be used. In the full study, we consider synthetic data and data from two micorarray studies. In these proceedings we focus on the linear SVM classifier and t-test feature selection using synthetic data. In the full study, similar results are seen across the full range of experiments, for all feature selection methods, classification rules, and data types.

Our simulation are done using artificial data. These data are generated from 2-classes Gaussian model. We use different value for the parameters of Gaussian (mean, covariance matrix) in order to generate linear and non linear classification problem with correlated and non correlated features.

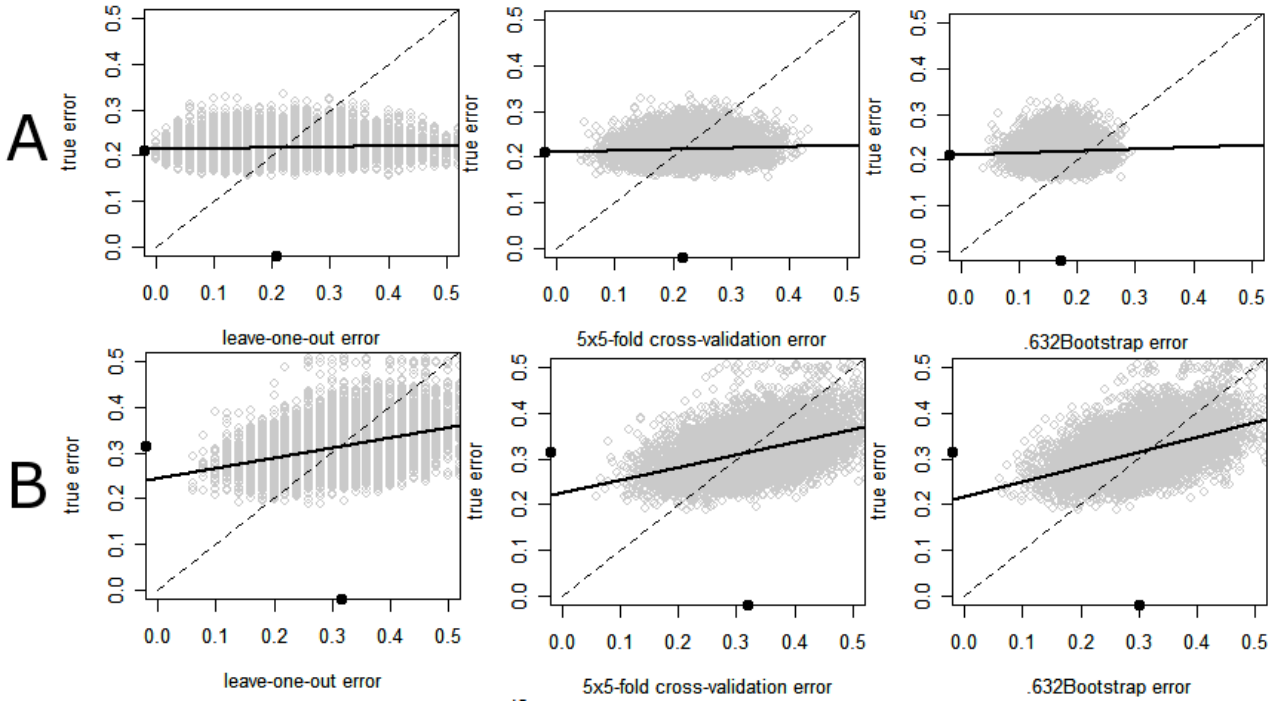


FIG. 1. *Comparison of the true and estimated errors. (A) experiment 6 with a linear model,N=50, D=200, d=5, t-test selection and SVM. (B) experiment 5 with a linear model, N=50, D=5, no feature selection and SVM.*

3.2. *Results*

We report in detail on the results obtained from the linear model with uncorrelated features, $n = 50$, $D = 200$, $d = 5$, feature selection by the t-test, and classification by a linear SVM. Figure 1A shows the estimated- and true-error pairs. The horizontal and vertical axes represent ε_{est} and ε_{tru}, respectively. The dotted 45-degree line corresponds to $\varepsilon_{est} = \varepsilon_{tru}$. The black line is the regression line. The means of the estimated and true errors are marked by dots on the horizontal and vertical axes, respectively. The three plots in Fig. 1A represent the comparison of the true error with the leave-one-out error, the 5×5-fold cross-validation error, and the .632 bootstrap error. The difference between the means of the true and estimated errors give the biases of the estimators : $E[\varepsilon_{tru}] = 0.21$, whereas $E[\varepsilon_{loo}] = 0.21$, $E[\varepsilon_{cv}] = 0.22$, and $E[\varepsilon_{b632}] = 0.17$. The leave-one-out and cross-validation estimators are virtually unbiased and the bootstrap is slightly biased. Estimator variance is represented by the width of the scatter plot.

Our focus is on the correlation and regression for the estimated and true errors. To distinguish feature selection from no feature selection, we denote these by $\hat{\rho}_{fs}$ and $\hat{\rho}_0$, respectively. To emphasize the error estimator, for instance,leave-one-out, we will write $\hat{\rho}_{fs}^{loo}$ or $\hat{\rho}_0^{loo}$. In Fig. 1A, the regression lines are almost parallel to the x-axis. The correlation is close to 0 for each estimation rule : $\hat{\rho}_{fs}^{loo} = 0.07$, $\hat{\rho}_{fs}^{cv} = 0.08$, and $\hat{\rho}_{fs}^{b632} = 0.06$. Ignoring the bias, which is small in all cases, the impact of the lack of correlation is not negligible since the variances of errors σ_{est}^2 and σ_{tru}^2 are not small.

To see the effect of feature selection, we compare the preceding results with those obtained for the linear model, 5 uncorrelated features, $n = 50$, t-test, and SVM classification, except without feature selection, the model being directly generated with $d = 5$ features. Figure 1B shows the data plots and regression lines for this case. There is significant regression in all three cases with $\hat{\rho}_0^{loo} = 0.43$, $\hat{\rho}_0^{cv} = 0.48$, and $\hat{\rho}_0^{b632} = 0.49$. Feature selection has caused drastic losses of correlation and regression.

In the full study consisting of 84 experiments, 60 with synthetic data and 24 with microarray data, in all cases, except for a single model, $\hat{\rho}_{fs} < \hat{\rho}_0$, and often $\hat{\rho}_{fs}$ is very small. The correlation increases with increasing sample size. Note that this increase has little practical impact because small error variances imply a small deviation variance, irrespective of the correlation. It is with small samples that lack of correlation and regression render error estimation virtually useless.

4. Conclusion

Our study shows that feature selection has a strong decorrelating effect on the true and estimated errors in small-sample settings, regardless of the classification rule, feature-selection procedure, and estimation method. The high feature dimensionality and small samples in bioinformatics have brought into play this decorrelating effect of feature selection. The implications are far reaching because there are hundreds of papers in the literature that have employed the methodology analyzed in this paper : high dimensional feature selection, classifier design, and error estimation with very small samples. Since epistemologically the meaning of such papers lies in the accuracy of the error estimation, one must ask the question whether these papers have any meaning whatsoever.

5. Bibliographie

[BRA 04] BRAGA-NETO U., DOUGHERTY E., Is cross-Validation Valid for Small-Sample Microarray Classification, *Bioinformatics*, vol. 20, n^o 3, 2004, p. 374-380.

[DEV 96] DEVROYE L., GYORFI L., *A Probabilistic Theory of Pattern Recognition*, Springer, New York, 1996.

[DOU 01] DOUGHERTY E., Small Sample Issues for Microarray-Based Classification, *Comparative and functinal Genomics*, vol. 2, 2001, p. 28-34.

[HAN 07] HANCZAR B., HUA J., DOUGHERTY E., Decorrelation of the True and Estimated Classifier Errors in High-Dimensional Settings, *EURASIP Journal on Bioinformatics and Systems Biology*, , 2007, page Article ID 38473.

[ISA 08] ISAKSSON A., WALLMAN M., GÖRANSSON H., GUSTAFSSON M. G., Cross-validation and bootstrapping are unreliable in small sample classification, *Pattern Recognition Letters*, vol. 29, n^o 14, 2008, p. 1960–1965.

[MOL 05] MOLINARO A., SIMON R., PFEIFFER M., Prediction Error Estimation : A Comparaison of Re-sampling Methods, *Bioinformatics*, vol. 21, n^o 15, 2005, p. 3301-3307.

Comparaison et classification de séries temporelles via leur développement en ondelettes de Haar asymétriques.

Catherine Timmermans[1], Rainer von Sachs[1], Véronique Delouille[2].

(1) Institut de Statistique, Université Catholique de Louvain, Voie du Roman Pays 20, BE-1348 Louvain-la-Neuve.
(2) Solar Influence Data analysis Center, Observatoire Royal de Belgique, Avenue Circulaire 3, BE-1180 Bruxelles.
Adresse de correspondance : catherine.timmermans@uclouvain.be

RÉSUMÉ. *Développer une série temporelle dans une base orthonormale adaptative constituée d'ondelettes de Haar asymétriques (Fryzlewicz, 2007 ; Girardi et Sweldens, 1997) revient à exprimer cette série sous forme d'une somme de contributions d'importances décroissantes : les premiers termes du développement encodent les caractéristiques majeures de la série et les termes suivants décrivent des motifs de moindre ampleur. Les caractéristiques considérées essentielles ici sont des changements de niveau localement importants (pics) ou affectant un grand ensemble de données (discontinuités du niveau moyen).*
Ce travail étudie la possibilité d'exploiter le caractère « hiérarchique » du développement en ondelettes de Haar asymétriques pour définir une mesure de dissimilarité entre des séries temporelles.

MOTS-CLÉS : *Ondelettes de Haar asymétriques, séries temporelles, mesure de dissimilarité.*

Introduction.

Nous souhaitons comparer et classer des séries temporelles.

La démarche à adopter semble claire si la classification s'effectue visuellement : confronter d'abord les allures globales des séries pour évaluer grossièrement leur ressemblance, puis affiner l'analyse en comparant les motifs de moindre ampleur présents dans la série. Cette approche visuelle pourrait-elle nous inspirer pour élaborer une méthode de classification automatique ? Les résultats présentés dans ce document vont dans ce sens.

Pour comparer des séries temporelles de N observations consécutives effectuées à intervalles réguliers, nous proposons une démarche en deux temps :

- Exprimer d'abord chaque série dans une base dont les premiers vecteurs supportent les caractéristiques majeures de la série, tandis que les vecteurs suivants correspondent à des motifs de moindre ampleur.
  Si les éléments considérés importants pour décrire la série sont des changements de niveau importants localement (pics) ou affectant un grand nombre de données (discontinuité du niveau moyen), alors les bases de Haar asymétriques semblent de bonnes candidates pour notre développement. Cette famille de bases, introduite en 1997 par Girardi et Sweldens [GIR 97], est présentée dans la première section de ce document.

- Mesurer ensuite la dissimilarité entre les séries temporelles, en exploitant le caractère « hiérarchique » du développement en ondelettes de Haar asymétriques, de sorte que les éléments d'importance similaire pour la description de chacune des séries soient comparés entre eux.
  La deuxième section de ce document est consacrée à la définition d'une telle mesure de dissimilarité.

Les troisième et quatrième sections de ce document sont dévolues à la mise en oeuvre de la démarche proposée pour la classification d'une part, et pour la comparaison évolutive de deux séries d'autre part.

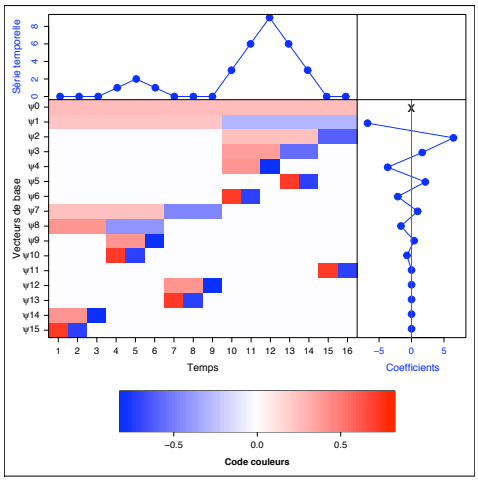


FIG. 1. Représentation d'un développement BUUHWT. La série temporelle est représentée dans la partie supérieure. L'axe temporel correspondant est commun à la figure principale et se trouve donc au bas de celle-ci. Cette partie principale du graphe représente les vecteurs de la base d'ondelettes de Haar asymétriques la mieux ajustée à la série (BUUHWT). Ces vecteurs sont représentés rang par rang, en fonction du temps, au moyen du code couleur indiqué au bas de la figure. Verticalement, de haut en bas sur le côté droit, se trouve la suite de coefficients associés au développement en ondelette. Chaque coefficient est ainsi placé en regard du vecteur de base auquel il correspond. Pour des raisons graphiques, la valeur du coefficient associé au vecteur constant ψ_0 n'est pas représentée.

1. Les bases d'ondelettes de Haar asymétriques.

Les bases d'ondelettes de Haar asymétriques sont des bases orthonormales constituées d'un vecteur constant et d'un ensemble d'ondelettes orthonormées de type Haar (fonctions « marche d'escalier ») dont le point de discontinuité entre parties positives et négatives n'est pas nécessairement situé au milieu du support. Les coefficients associés à ces bases d'ondelettes encodent l'importance des changements de niveau entre deux observations, ou groupes d'observations, consécutifs.
En 2007, Fryzlewicz [FRY 07] a proposé, sous le nom de *Bottom-Up Unbalanced Haar Wavelet Transform* (ci-après BUUHWT), un algorithme de construction de la base d'ondelettes de Haar asymétriques $\{\psi_k\}_{k=0..N-1}$ la mieux adaptée à une série particulière - au sens où les vecteurs de cette base et leurs coefficients associés sont ordonnés selon l'importance des changements de niveau qu'ils encodent pour la description de la forme globale de la série. Le développement qui en résulte est un développement multi-échelles évitant la restriction dyadique propre aux ondelettes classiques. La famille des bases d'ondelettes de Haar asymétriques est donc réellement adaptative.

La figure 1 présente le développement BUUHWT obtenu pour une série particulière. Comme nous le souhaitions, les premiers vecteurs non constants supportent le pic le plus important de la série. On s'en rend compte en regardant la position de la discontinuité entre les parties positives et négatives des différentes ondelettes. Les vecteurs suivants indiquent le petit pic tandis que les quelques derniers vecteurs correspondent à des zones ne présentant pas de changement de niveau. Ils sont ici associés à des coefficients nuls, car la série illustrée n'est pas bruitée.

2. La définition d'une mesure de dissimilarité.

Nous proposons ici une mesure de la dissimilarité de deux séries temporelles $x^{(1)}$ et $x^{(2)}$, fondée sur les développements BUUHWT respectifs de celles-ci.

Notons comme suit le développement BUUHWT de la série $x^{(i)}, i = 1, 2$:

$$x^{(i)} = \sum_{k=0}^{N-1} d_k^{(i)} \psi_k^{(i)} = \sum_{k=0}^{N-1} v_k^{(i)}, \qquad [1]$$

où les $d_k^{(i)}$ sont les coefficients associés aux vecteurs de base $\psi_k^{(i)}$ pour la i^e série. Le vecteur $v_k^{(i)}$ est donc la contribution du rang k à la description de la série temporelle $x^{(i)}$. L'indice de rang k, $k = 1 \ldots N-1$, est d'autant plus petit que les éléments correspondants ont un rôle majeur pour décrire la structure de la série, tandis que le

rang $k = 0$ est associé au vecteur constant $\psi_0^{(i)}$ et encode le niveau moyen de la série. À chacune des ondelettes $\psi_k^{(i)}, k = 1 \ldots N-1$, est par ailleurs associé un indice $s_k^{(i)}$ encodant son point de discontinuité : $s_k^{(i)}$ est l'abscisse de la dernière valeur strictement positive de l'ondelette $\psi_k^{(i)}$.

Ces notations étant fixées, nous proposons de mesurer la dissimilarité d des séries $x^{(1)}$ et $x^{(2)}$ par une somme pondérée de dissimilarités partielles δ_k évaluées rang par rang :

$$d(x^{(1)}, x^{(2)}) = \sum_{k=1}^{N-1} w_k \delta_k, \qquad [2]$$

où w_k est une fonction de poids décroissante car l'importance de la dissimilarité au rang k devrait affecter la mesure de dissimilarité globale d'autant plus fortement que les caractéristiques comparées correspondent à des éléments majeurs de la série. Dans le cadre de cette étude, la fonction de poids utilisée est définie par $w_k = \frac{\log(N+1-k)}{\log(N+1)}$. Comme nous nous intéressons ici aux structures des séries, nous ne considérons pas le rang $k = 0$ encodant leur niveau global et notre mesure de dissimilarité est donc définie à un changement de niveau près.

Nous mesurons la dissimilarité partielle au rang k en combinant les différences de localisations des ondelettes à ce rang et les différences d'amplitude des coefficients associés :

$$\delta_k = |s_k^{(1)} - s_k^{(2)}| + |d_k^{(1)} - d_k^{(2)}|. \qquad [3]$$

Conceptuellement, les fonctions valeurs absolues pourraient ici être remplacées par toute autre fonction monotone croissante en fonction de la valeur absolue de son argument. Les résultats des tests présentés ci-après ne s'en trouvent cependant pas améliorés.

Dans certains cas, une pondération des deux termes de l'expression [3] devrait être introduite. L'obtention d'un poids optimal est aisée par validation croisée si l'on traite un problème de classification supervisée. Dans ce cadre, l'expérience a confirmé qu'un équilibre entre le terme de localisation et le terme d'amplitude est plus performant que la présence seule de l'un de ces deux termes. Notons que l'introduction d'une pondération équivaut visuellement à modifier l'échelle de représentation de la série en abscisse (axe temporel) ou en ordonnée (valeurs observées). Le choix de l'échelle idéale - et donc de la pondération - peut donc le plus souvent se faire graphiquement, car la meilleure échelle pour une bonne discrimination devrait coïncider avec l'échelle où l'on distingue le mieux les séries visuellement.

3. Un exemple de classification supervisée.

Afin d'évaluer les potentialités de la dissimilarité ainsi définie, le test suivant est effectué : 10 séries-type sont définies ; elles sont chacune de longueur 11 et les caractéristiques qui les composent ont des hauteurs variant entre 10 et 100. Ces séries-type sont présentées à la figure 2 (gauche). À partir de chacune des 10 familles, 1000 séries sont simulées : 20% d'entre elles correspondent aux séries initiales décalées d'un pas de temps vers la droite, 20% aux séries décalées d'un pas de temps vers la gauche, 20% sont amplifiées d'un facteur 1.25, 20% d'un facteur 0.75. En outre, toute les séries simulées sont affectées d'un bruit additif gaussien. Ensuite, les dissimilarités de ces 10000 séries aux 10 séries-type sont évaluées. Les séries sont alors classées dans la famille la plus proche. Le pourcentage de mauvaises classifications - qui est ici notre critère pour évaluer la qualité relative de la dissimilarité proposée - est alors calculé. Les résultats présentés à la figure 2 (centre) correspondent à la dissimilarité d introduite à la section précédente, ainsi qu'à la distance euclidienne notée $dEucl$.

Nous observons un taux de mauvaises classifications significativement moindre pour notre mesure de dissimilarité que pour la distance euclidienne, jusqu'au niveau de bruit maximum testé, qui correspond à 100% de la grandeur de la plus petite caractéristique des séries à classer et 10% de la plus grande.

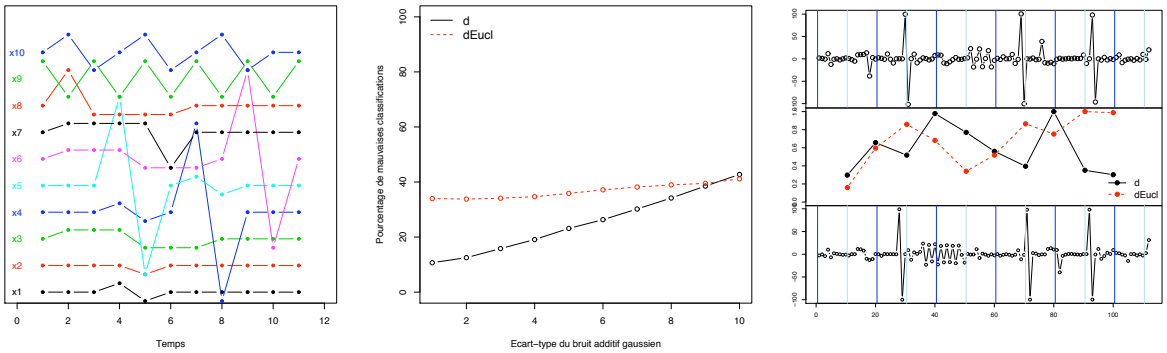


FIG. 2. De gauche à droite : (1) Les **10 familles de séries** considérées pour le test de classification. (2) **Résultat de la classification supervisée :** Pourcentage de mauvaises classifications en fonction de l'écart-type du bruit additif gaussien affectant les séries à classer, pour la dissimilarité d est présentée à la section 2, et la distance euclidienne $dEucl$. L'écart-type du bruit maximum testé (*Ecart-type*$= 10$) correspond à 100% de la grandeur de la plus petite caractéristique des séries présentées en (1) et 10% de la plus grande. Jusqu'à ce niveau de bruit, la dissimilarité d reposant sur le développement BUUHWT des séries est plus performante que la distance euclidienne. (3) **Comparaison évolutive de deux séries temporelles**. En haut et en bas de la figure sont représentées les séries comparées. La bande centrale présente l'évolution de la distance entre les courbes, calculée au moyen de la dissimilarité d introduite à la section 2 ou au moyen de la distance euclidienne $dEucl$. Les valeurs calculées sont ici normalisées afin de permettre la comparaison des résultats obtenus.

4. Comparaison évolutive de séries temporelles longues.

Lorsque nous souhaitons comparer des séries temporelles longues au moyen de la dissimilarité d introduite à la section 2, il s'avère souvent judicieux de privilégier une transformation localisée, afin d'éviter une confusion des caractéristiques comparées deux à deux. Le principe est le suivant. Les séries temporelles étudiées sont subdivisées en sous-séries de longueur T. Un développement BUUHWT est alors obtenu pour chaque sous-série. Les vecteurs de base sont ensuite concaténés rang par rang et les coefficients de même rang rassemblés en un vecteur. La dissimilarité entre les séries est alors calculée section par section.
Bien entendu, les résultats obtenus dépendent de la longueur T choisie pour les sous-séries. T définit dans quels intervalles les caractéristiques sont comparées deux à deux. Cette longueur est donc fonction de la densité de la série en caractéristiques et devrait être déterminée au cas par cas. Par ailleurs, pour limiter la sensibilité au point de départ du découpage en sous-séries, nous proposons de considérer des sous-séries décalées de $\frac{T}{2}$, de sorte que la décomposition soit redondante.

On peut ensuite moyenner ces dissimilarités locales pour obtenir une mesure de dissimilarité globale. Alternativement, on peut s'intéresser à la courbe des dissimilarités entre séries pour identifier automatiquement les moments où deux séries sont proches et et ceux où elles se distancient. Ceci pourrait s'avérer un outil de diagnostique intéressant pour des problèmes de comparaison de courbes à une courbe cible. Pensons par exemple ici à du contrôle de processus ou à la calibration de modèles d'évolution.
Un exemple de telle comparaison évolutive est présenté à la figure 2 (droite). La courbe de dissimilarités obtenue semble visuellement raisonnable.

5. Bibliographie

[FRY 07] FRYZLEWICZ P., Unbalanced Haar technique for non parametric function estimation, *J. Am. Stat. Assoc.*, vol. 102, n^o 480, 2007, p. 1318-1327.

[GIR 97] GIRARDI M., SWELDENS W., A new class of unbalanced Haar wavelets that form an unconditional basis for Lp on general measure spaces, *J. Fourier Anal. Appl.*, vol. 3, n^o 4, 1997, p. 457-474.

Apprentissage de différentes classes de similarité dans les k-PPVs

Ali Mustafa Qamar, Eric Gaussier

LIG (Laboratoire d'Informatique de Grenoble)
Université Joseph Fourier
BP 53 - 38041 Grenoble cedex 9
{ali-mustafa.qamar,eric.gaussier}@imag.fr

RÉSUMÉ. *Beaucoup de travaux se sont intéressés à l'apprentissage de métriques pour des distances de Mahalanobis dans le cadre de la classification de type k-PPV (k plus proches voisins). Or pour certains types d'objets, comme les textes, il est souvent préférable d'utiliser une similarité plutôt que'une distance. Nous nous intéressons dans ce travail à l'apprentissage de métriques pour des similarités (de type cosinus par exemple) dans le cadre de la classification de type k-PPV. En particulier, nous présentons deux algorithmes incrémentaux permettant d'apprendre des similarités qui généralisent les coefficients de Dice, de Jaccard et le cosinus. Nous illustrons ces algorithmes sur des bases de données standard de classification.*

MOTS-CLÉS : *Classification, Apprentissage du similarité, k-PPV.*

1. Introduction

L'algorithme des k-plus proches voisins (k-PPV) est un des plus anciens mais aussi des plus simples algorithme de classification. Il consiste, pour chaque nouvel exemple x à catégoriser, à déterminer les k-plus proches voisins de x déjà catégorisés et à affecter x à la classe la plus représentée dans ces plus proches voisins. Cet algorithme est utilisé dans différentes communautés (base de données, recherche d'information, apprentissage). Plusieurs approches ont été proposées pour améliorer l'algorithme des k-PPV en utilisant la géométrie des données. Ces approches se sont essentiellement concentrées sur des métriques généralisant la distance euclidienne ([WEI 06, DAV 07, SHA 04]). Or, dans beaucoup de situations, ce ne sont pas des distances qui nous intéressent mais plutôt des similarités. Nous abordons précisément ici le problème de l'apprentissage de similarités pour déterminer les k-plus proches voisins d'un exemple donné.

2. Formulation du Problème

Nous considérons ici des similarités de la forme :

$$\mathrm{s}_A(x, y) = \frac{x^T A y}{\mathrm{N}(x, y)} \qquad [1]$$

où x et y sont deux exemples de $\mathbb{R}^p$, T désigne la transposée d'une matrice, A une matrice $(p \times p)$ et $\mathrm{N}(x, y)$ une normalisation qui permet souvent, en pratique, d'obtenir des scores dans l'intervalle $[0, 1]$). Le but du travail que nous avons développé dans [QAM 08] est d'apprendre, à partir d'exemples déjà classifiés, des matrices A symétriques (diagonales ou non) ou quelconques. Dans ces cas, les similarités définies par l'équation 1 peuvent être vues, suivant la normalisation adoptée, comme des généralisations particulières du cosinus, du coefficient de Jacard ou du coefficient de Dice. Toutefois, dans ce genre d'approches, rien ne garantit que la forme $x^T A y$

correspond à une forme bilinéaire symétrique, et donc un produit scalaire, ce qui peut être gênant lorsque l'on cherche à définir une fonction cosinus généralisée. Nous nous concentrons donc ici à l'apprentissage de similarités correspondant à l'équation 1 pour lesquelles la matrice A considérée est semi-définie positive.

3. Apprentissage de similarité

Nous avons introduit dans [QAM 08] un algorithme, applelé *SiLA*, qui est une variante du perceptron par vote proposé dans [FRE 99] et utilisé dans [COL 02], et pour lequel nous avons établi des bornes sur l'erreur de généralisation, à la fois en mode en ligne et en mode batch. Nous proposons ici d'étendre cet algorithme, extension que nous appellerons *eSiLA*, en ajoutant, à chaque mise à jour de la matrice A, une projection orthogonale sur le cône des matrices semi-définies positives, suivant en cela la méthode proposée dans [SHA 04]. Cette projection orthogonale garantit, d'une part, la convergence de l'algorithme, et préserve, d'autre part, les bornes sur l'erreur de généralisation (la présentation de cette convergence et de ces bornes dépasse le cadre du présent résumé, et sera fournie dans une version étendue de l'article).

Soit $(x^{(1)}, c^{(1)}), \cdots, (x^{(n)}, c^{(n)})$ un ensemble d'apprentissage de n exemples étiquetés, avec $x^{(i)} \in \mathbb{R}^p$. Nous introduisons, comme dans [WEI 06], pour chaque $x^{(i)}$, ses k voisins *cible*, qui sont les k éléments de $c^{(i)}$ les plus proches de $x^{(i)}$, selon une similarité de base que l'on veut généraliser (comme par exemple le cosinus). Dans la suite, k-PPV(A, x, s) désigne l'ensemble des k plus proches voisins de l'exemple x dans la classe s avec la fonction de similarité donnée par l'équation 1, $T(i)$ l'ensemble des voisins cibles de $x^{(i)}$. Nous nous plaçons de plus dans le cas d'un problème de classification binaire, l'extension au cadre multi-classes étant réalisée naturellement en considérant une approche *un contre tous*. L'algorithme *eSiLA* est défini avec ces données par :

eSiLA - Apprentissage

Entrée : Ensemble d'apprentissage $((x^{(1)}, c^{(1)}), \cdots, (x^{(n)}, c^{(n)}))$ du n vecteurs dans $\mathbb{R}^p$, nombre d'itérations M ;

Sortie : liste de matrices $(p \times p)$ pondérés $((A^1, w_1), \cdots, (A^q, w_q))$

Initialisation $t = 1, A^{(1)} = 0$ (matrice nulle), $w_1 = 0$

Répéter M **fois**

1. pour $i = 1, \cdots, n$
2. $B(i) =$ k-PPV$(A^{(t)}, x^{(i)}, \bar{c}^{(i)})$
3. si $\sum_{y \in T(i)} s_A(x^{(i)}, y) - \sum_{z \in B(i)} s_A(x^{(i)}, z) \leq 0$
4. $A^{(t+1)} = \hat{A}^{(t)} + \sum_{y \in T(i)} f(x^{(i)}, y) - \sum_{z \in B(i)} f(x^{(i)}, z)$
5. $\hat{A}^{(t+1)} = \sum_{j, \lambda_j > 0} \lambda_j u_j u_j^T$ (où λ et u sont les valeurs propres et vecteurs propres de la matrice A^{t+1})
6. $w_{t+1} = 1$
7. $t = t + 1$
8. alors
9. $w_t = w_t + 1$

Quand un exemple d'entrée est mal catégorisé par la fonction de similarité (ligne 3), la matrice courante A est mise à jour par la différence entre les coordonnées des voisins cible et des plus proches voisins situés dans l'autre classe (ligne 4 de l'algorithme), ce qui correspond à la mise à jour du perceptron standard. Dans ce cas, la matrice mise à jour est projetée sur le cône des matrices semi-définies positives pour obtenir une similarité valide dans le cadre retenu ici. Au contraire, quand la matrice courante $A^{(t)}$ classifie correctement l'exemple d'entrée, alors son poids est augmenté de 1. Le poids final de $A^{(t)}$ correspond donc au nombre d'exemples correctement classifiés par $A^{(t)}$ sur les différentes itérations. Comme dans [QAM 08], $f(x, y)$ est une matrice $(p \times p)$ de similarité entre x et y.

TAB. 1. *Caractéristiques des données*

	Iris	**Wine**	**Balance**	**Ionosphere**	**Glass**	**Soybean**	**Pima**	**News**
#ex. (apprentissage)	96	114	400	224	137	30	492	1824
#ex. (validation)	24	29	100	57	35	8	123	457
#ex. (test)	30	35	125	70	42	9	153	2280
#classes	3	3	3	2	6	4	2	20
#traits	4	13	4	34	9	35	8	20000
#traits (après SVD)	4	13	4	34	9	35	8	200

Les matrices apprises par l'algorithme ci-dessus permettent de construire une matrice $A = \sum_{l=1}^{q} w_l A^l$ qui sera utilisée pour classifier un nouvel exemple[1]. Nous considérons deux règles de classification formée sur A. La première correspond à la règle standard des k-plus proches voisins, alors que la deuxième consiste à affecter un nouvel exemple x à la classe c pour laquelle $\sum_{z \in \text{k-PPV}(A,x,c)} s_A(x,z)$. Nous appellerons cette règle *règle de classification symétrique* dans la suite.

4. Validation expérimentale

Nous présentons le résultats obtenus par *SiLA* et *eSiLA* sur huit collections de test différentes. Les sept première font partie de la base de données UCI ([ASU 07]). Ce sont les collections *Iris*, *Wine*, *Balance*, *Pima*, *Ionosphere*, *Glass* et *Soybean*. La huitième collection est la base 20-newsgroups, composée d'environ 20000 articles. Dans ce dernier cas, nous avons utilisé la bibliothèque *Rainbow* ([MCC 96]) pour segmenter chaque document en ne retenant que les 20000 mots les plus fréquents. Nous avons ensuite appliqué une décomposition en valeurs singulières en utilisant *SVDlibc*[2], ce qui a réduit le nombre de dimensions à 200. Cette dernière étape réduit le nombre de documents non vides à 4561. Le tableau 1 récapitule les informations concernant ces différentes collections.

Bien que les k-plus proches voisins aient été utilisés traditionnellement avec la distance euclidienne sur ces collections (ou avec la distance de Mahanalobis, généralisation de la distance euclidienne, comme [WEI 06, DAV 07]), nos résultats montrent que le cosinus doit être préféré à la distance euclidienne sur la majorité des collections considérés. Dans nos expériences, nous avons utilisé les deux règles de classification décrites plus haut. La première est la règle standard des *k-PPV*, dans laquelle la classification est fondée sur les k plus proches voisins, alors que la deuxième, *Sk-PPV* ou *S* signifie "symétrique", s'appuie sur la différence de la similarité entre les k plus proches voisins dans une même classe et les k plus proches voisins dans les autres classes[3].

Le tableau 2 fournit les résultats obtenus sur toutes les collections. Ce tableau montre que nos méthodes (k-PPV-A et Sk-PPV-A) améliorent les k-plus proches voisins sur la majorité des collections (Balance, Wine, Ionosphere, Pima et News). Sur Soybean, Iris et Glass, toutes les méthodes fournissent à peu près les mêmes résultats (proches de 1). D'un certain point de vue, il y a peu à gagner sur ces collections dans la mesure où les k-PPV standard se comportent déjà remarquablement bien. Enfin, on peut noter que *eSiLA* est meilleur que *SiLA* sur Wine et Pima, et comparable sur Ionosphere et Soybean.

5. Conclusion

Dans ce papier, nous avons dévelopé un algorithme "eSiLA" qui permet d'apprendre des similarités, correspondant à l'équation 1 et fondées sur des matrices semi-définies positives. Cet algorithme étend l'algorithme SiLA

1. Nous renvoyons le lecteur à [HEL 95] pour la justification de la forme de A.
2. Disponible sur http ://tedlab.mit.edu/ dr/svdlibc/
3. On peut trouver dans [NOC 03] une version un peu différente de la règle *symétrique* du k-PPV dans laquelle on considère les k plus proches voisins de l'exemple x mais aussi les points pour lesquels x est un voisin le plus proche

TAB. 2. *Résultats sur toutes les collections*

	k-PPV-eucl	k-PPV-cos	k-PPV-A	Sk-PPV-cos	Sk-PPV-A
Balance	0.883	0.959	0.979	0.969	**0.983**
Wine	0.825	0.905	**0.916**	0.909	**0.916**
Wine (eSiLA)	0.825	0.905	**0.926**	0.909	**0.928**
Iris	0.978	**0.987**	**0.987**	0.982	**0.987**
Ionosphere	0.854	0.871	**0.911**	0.871	**0.911**
Ionosphere (eSiLA)	0.854	0.871	**0.914**	0.871	**0.914**
Soybean	**1.0**	**1.0**	0.994	0.989	0.989
Soybean (eSiLA)	**1.0**	**1.0**	**1.0**	0.989	0.989
Glass	**0.998**	**0.998**	0.997	**0.998**	0.997
Pima	**0.698**	0.652	0.647	0.665	**0.678**
Pima (eSiLA)	**0.698**	0.652	0.659	0.665	**0.673**
News	-	0.929	**0.947**	0.907	0.902

que nous avons développé dans [QAM 08] par une projection orthogonale sur le cône des matrices semi-définie positives. Nous avons validé notre algorithme sur différentes collections standard, issues, pour la plupart, de la base de collections UCI. Les résultats que nous avons obtenus montrent l'intérêt d'un apprentissage de similarité sur la majorité des collections considérées.

6. Bibliographie

[ASU 07] ASUNCION A., NEWMAN D., UCI Machine Learning Repository, 2007.

[COL 02] COLLINS M., Discriminative training methods for hidden Markov models : theory and experiments with perceptron algorithms, *EMNLP '02 : Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, Morristown, NJ, USA, 2002, Association for Computational Linguistics, p. 1–8.

[DAV 07] DAVIS J. V., KULIS B., JAIN P., SRA S., DHILLON I. S., Information-Theoretic Metric Learning, *Proceedings of the 24th International Conference on Machine Learning*, 2007.

[FRE 99] FREUND Y., SCHAPIRE R. E., Large Margin Classification Using the Perceptron Algorithm, *Mach. Learn.*, vol. 37, n^o 3, 1999, p. 277–296, Kluwer Academic Publishers.

[HEL 95] HELMBOLD D. P., WARMUTH M. K., On weak learning, *J. Comput. Syst. Sci.*, vol. 50, n^o 3, 1995, p. 551–573, Academic Press, Inc.

[MCC 96] MCCALLUM A. K., Bow : A toolkit for statistical language modeling, text retrieval, classification and clustering, 1996.

[NOC 03] NOCK R., SEBBAN M., BERNARD D., A Simple Locally Adaptive Nearest Neighbor Rule with Application to Pollution Forecasting, *International Journal for Pattern Reconition and Artificial Intelligence*, vol. 17, n^o 8, 2003.

[QAM 08] QAMAR A. M., GAUSSIER É., CHEVALLET J.-P., LIM J.-H., Similarity Learning for Nearest Neighbor Classification, *ICDM*, 2008, p. 983-988.

[SHA 04] SHALEV-SHWARTZ S., SINGER Y., NG A. Y., Online and batch learning of pseudo-metrics, *ICML '04 : Proceedings of the twenty-first international conference on Machine learning*, New York, NY, USA, 2004, ACM.

[WEI 06] WEINBERGER K., BLITZER J., SAUL L., Distance Metric Learning for Large Margin Nearest Neighbor Classification, WEISS Y., SCHÖLKOPF B., PLATT J., Eds., *Advances in Neural Information Processing Systems 18*, p. 1473–1480, MIT Press, Cambridge, MA, 2006.

Comparaison et évaluation de métriques pour la classification de profils d'expression de gènes

Alpha Diallo, Ahlame Douzal-Chouakria, Françoise Giroud

Université Joseph Fourier Grenoble 1
Lab. TIMC-IMAG, CNRS UMR 5525
Faculté de Médecine
38706 LA TRONCHE Cedex, France
{Alpha.Diallo, Ahlame.Douzal, Francoise.Giroud}@imag.fr

RÉSUMÉ. *La division cellulaire est un processus fondamental dans le monde du vivant puisqu'il représente le mode de multiplication de toute cellule. Le cycle cellulaire est l'ensemble des phases par lesquelles une cellule passe entre deux divisions successives. Le cycle cellulaire comprend 4 phases principales : G1, S, G2 et M. La régulation de la division cellulaire apparaît d'une très grande complexité car il existe des cascades de réactions moléculaires interdépendantes dont il est difficile de situer le point de départ. Nous allons dans cet article nous intéresser au développement de méthodes de classification des gènes différentiellement exprimés au cours du cycle cellulaire. Les gènes étudiés sont décrits par leurs profils d'expression au cours du temps (c'est à dire, par des séries temporelles), données issues de la technologie des biopuces à ADN. Quatre méthodes de mesure de distances de classification ont été testées. Une étude de l'efficacité des méthodes a été menée sur un jeu de données temporelles virtuel généré sur la base d'un modèle "random-periods". Le modèle tient compte de l'amplitude initiale du comportement périodique des profils, de la durée du cycle cellulaire, de l'atténuation des amplitudes suivant les cycles et de la tendance des profils d'expression des gènes.*

MOTS-CLÉS : *Séries temporelles, mesures de distance, classification, profils d'expression de gènes.*

1. Introduction

Les techniques de classification se sont montrées particulièrement efficaces pour comprendre la caractérisation de la fonction des gènes et des voies de régulation. Elles permettent de répondre à des problèmes biologiques importants allant de la recherche des gènes corégulés au cours du cycle cellulaire à l'identification des liens entre les variations d'expression et d'autres données biologiques. Nous distinguons au moins deux principales approches de classification de profils ou de séries temporelles : la paramétrique et la non-paramétrique. Dans cet article nous nous sommes intéressés à l'approche non-paramétrique qui consiste à classer les séries temporelles sur la base de leurs descriptions temporelles. Le défi de cet approche est de savoir comment inclure l'information de dépendance entre les mesures prises à des moments différents. Dans ce contexte, nous proposons d'évaluer l'éfficacité de quatre métriques pour la classification des profils d'expression de gènes. Cette étude a été menée sur un jeu de données temporelles virtuel généré sur la base d'un modèle "random-periods". Le modèle tient compte de l'amplitude initiale du comportement périodique des profils, de la durée du cycle cellulaire, de l'atténuation des amplitudes suivant les cycles et de la tendance des profils d'expression des gènes.

2. Proximité entre profils d'expression de gènes

Pour la classification d'un ensemble de profils d'expression de gènes évoluant dans le temps, le choix de la distance est crucial puisqu'il définit la mesure de ressemblance entre profils de deux gènes. La distance euclidienne

et celles fondées sur le coefficient de Pearson ont souvent été utilisées. Considérons les niveaux d'expression de deux gènes $g_1 = (u_1, ..., u_p)$ et $g_2 = (v_1, ..., v_p)$ observés aux instants $(t_1, ..., t_p)$. La distance euclidienne δ_E entre g_1 et g_2 est définie par : $\delta_E(g_1, g_2) = \left(\sum_{i=1}^{p}(u_i - v_i)^2\right)^{\frac{1}{2}}$. Il ressort de cette définition, que la proximité dépend uniquement de l'écart entre les valeurs d'expression sans tenir compte de la forme des profils d'expression. En d'autres termes, deux profils d'expression de gènes sont dits proches au sens de δ_E si et seulement si les valeurs observées aux mêmes instants sont proches. Cette distance ignore l'information de dépendance entre les valeurs d'expression, elle est invariante à toute permutation des instants d'observations. Einsen et al.[EIS 98] suggèrent des critères fondés sur le coefficient de corrélation de Pearson. Toutefois, les distances basées sur ce coefficient peuvent mener à une surestimation des proximités en forme due aux effets de tendance, comme nous le montrerons dans la section 5. En réponse à ces limites, nous utilisons un indice de dissimilarité couvrant la mesure de proximité en valeurs et en forme des profils d'expression de gènes proposé par Douzal Chouakria et Nagabhushan[DOU 07]. La proximité en forme est mesurée par le coefficient de corrélation temporelle suivant :

$$\text{CORT}(g_1, g_2) = \frac{\sum_{i=1}^{p-1}(u_{(i+1)} - u_i)(v_{(i+1)} - v_i)}{\sqrt{\sum_{i=1}^{p-1}(u_{(i+1)} - u_i)^2}\sqrt{\sum_{i=1}^{p-1}(v_{(i+1)} - v_i)^2}}$$

$\text{CORT}(g_1, g_2) \in [-1, 1]$. La valeur $\text{CORT}(g_1, g_2) = 1$ signifie que dans chaque période d'observation $[t_i, t_{i+1}]$, les expressions des gènes g_1 et g_2 croissent ou décroissent simultanément avec le même taux d'accroissement (formes similaires). Une valeur de $\text{CORT}(g_1, g_2) = -1$ exprime que dans chaque période d'observation $[t_i, t_{i+1}]$ g_1 croit, g_2 décroit ou vice-versa avec un même taux d'accroissement en valeur absolue (formes opposées). Enfin une valeur de $\text{CORT}(g_1, g_2) = 0$ signifie une absence de monotonicité entre les accroissements de g_1 et g_2 et leurs taux d'accroissement sont stochastiquement linéairement indépendants (formes différentes). L'indice de dissimilarité D_k est défini comme suit :

$$D_k(g_1, g_2) \quad = \quad f(\text{CORT}(g_1, g_2))\ \delta_E(g_1, g_2), \quad \text{avec } f(x) = \frac{2}{1 + exp(k\ x)}, \quad k \geq 0.$$

Cet indice de dissimilarité module la proximité en valeurs en fonction de la proximité en forme. La fonction de modulation f devra augmenter la proximité en valeurs si les formes sont opposées (la corrélation temporelle décroit de 0 à -1). À l'inverse, elle diminuera la proximité en valeurs à mesure que les formes sont similaires (la corrélation temporelle évolue de 0 à +1). La dissimilarité résultante correspond à la distance euclidienne si les formes sont différentes (corrélation temporelle nulle).

Nous allons, dans les sections suivantes, comparer et évaluer la distance euclidienne δ_E, une métrique fondée sur le coefficient de corrélation de Pearson classique COR, une mesure de distance basée sur coefficient de corrélation temporelle CORT et l'indice de dissimilarité D_k sur un jeu de données simulées.

3. Simulation des profils d'expression périodiques

Nous utilisons des profils d'expression de gènes virtuels générés sur la base d'un modèle proposé par Liu et al.[LIU 04] pour étudier des gènes périodiquement exprimés. La fonction sinusoïdale caractérisant la périodicité des expressions est :

$$f(t, \theta_g) \quad = \quad a_g + b_g t + \frac{K_g}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} cos(\frac{2\pi t}{T exp(\sigma z)} + \Phi_g) exp(-\frac{z^2}{2}) dz$$

où $\theta_g = (K_g, T, \sigma, \Phi_g, a_g, b_g)$ est spécifique du gène g. Le paramètre K_g représente l'amplitude initiale du profil d'expression périodique du gène g, T est la durée du cycle cellulaire. Le paramètre σ régit le taux d'atténuation d'amplitude au cours des différents cycles, Φ_g correspond à la phase du cycle cellulaire où le gène est le plus exprimé. Les paramètres a_g et b_g (l'ordonnée à l'origine et la pente , respectivement) contrôlent le sens d'évolution (tendance) des profils.

Basé sur le modèle et les paramètres spécifiés dans [LIU 04], quatre expériences sont menées. La première expérience génère des profils d'amplitude initiale K_g variant dans $[0.34, 1.33]$. La seconde expérience inclut une atténuation des amplitudes σ évoluant dans $[0.054, 0.115]$. La troisième expérience inclut les effets de tendance

$b_g \in [-0.05, 0.05]$ et $a_g \in [0, 0.8]$ et enlève les effets de σ. Enfin la quatrième expérience simule des profils avec une variation simultanée des paramètres K_g,σ, a_g, b_g dans les mêmes intervalles que précédemment. Chaque paramètre est pris de manière aléatoire dans son intervalle. Pour chaque expérience $j \in \{1, ..., 4\}$, 10 échantillons S_{ij} $i \in \{1, ..., 10\}$ sont simulés. Chaque échantillon est composé de 500 profils d'expression de gènes avec 100 gènes pour chacune des 5 phases ou inter-phases G_1/S, S, G_2, G_2/M et M/G_1. L'évolution des profils est suivi sur 3 cycles cellulaires, T est fixée à 15 heures pour toutes les simulations et Φ_g prend les valeurs $0, 5.190, 3.823$, 3.278 ou 2.459 pour la génération respective des 5 classes G_1/S, S, G_2, G_2/M ou M/G_1. La figure 1 montre la variation de profils de la classe G_1/S de l'expérience 4.

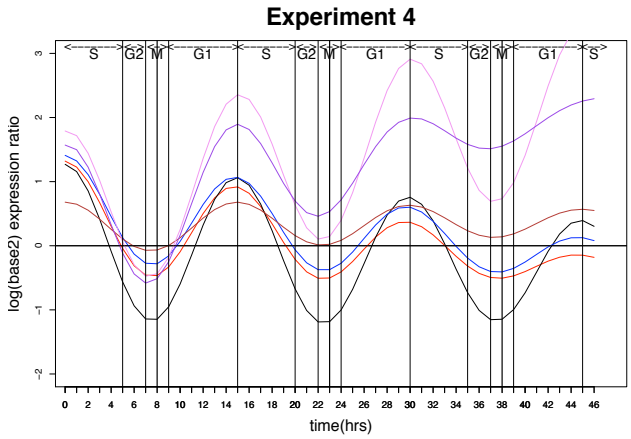


FIG. 1. *Visualisation de quelques profils d'expression simulés de la classe G_1/S pour l'expérience 4.*

4. Evaluation des métriques

Pour la méthode de classification, nous utilisons l'algorithme PAM (Partitioning Around Medoids), proposé par Kaufman et Rousseuw[KAU 90], pour partionner l'ensemble des profils d'expression générés de chaque échantillon S_{ij} en 5 classes (correspondant aux 5 phases et inter-phases du cycle cellulaire). L'algorithme est utilisé pour chaque métrique δ_E, COR et CORT et pour chaque expérience j. Par exemple, pour la métrique δ_E et pour l'expérience j, l'algorithme PAM est appliqué sur les 10 échantillons $S_{1j}, ..., S_{10j}$ afin d'extraire les 10 partitions $P^{1j}_{\delta_E}$,..., $P^{10j}_{\delta_E}$. Trois critères sont retenus pour apprécier la qualité des classes : la *average silhouette width* (asw), le *corrected Rand index* (RI) et le *within/between ratio* (wbr). Pour évaluer la métrique δ_E de l'expérience j, les valeurs moyennes sur les 10 partitions des critères asw, RI et wbr sont alors retenues. Pour ce qui est de l'indice de dissimilarité D_k, une classification adaptative est appliquée. Elle consiste à exécuter l'algoritheme PAM sur l'échantillon S_{ij} pour plusieurs valeurs de k ($k = 0, ..., 6$ avec un pas égale à 0.01). Ceci permet d'apprendre la valeur $k*$ qui fournit la partition optimale $P^{ij}_{D_{k*}}$ selon les critères asw et wbr. La valeur $k*$ donne la meilleure contribution des proximités en valeurs et en forme à l'indice de dissimilarité et par conséquent à D_k. De façon similaire aux métriques δ_E, COR et CORT, l'évaluation de D_{k*} est effectuée sur la base de la moyenne des valeurs asw, RI et wbr obtenues sur les 10 partitions $P^{1j}_{D_{k*}}$,..., $P^{10j}_{D_{k*}}$.

Pour la méthode de classement, l'algorithme 10-NN est utilisé pour chaque métrique δ_E, COR et CORT et pour chaque expérience j. Par exemple, pour la métrique δ_E et pour l'expérience j, l'algorithme 10-NN est appliqué sur les 10 échantillons $S_{1j}, ..., S_{10j}$ pour générer 10 classes $C^{1j}_{\delta_E}$,..., $C^{10j}_{\delta_E}$. Pour chaque classe $C^{ij}_{\delta_E}$ le taux de mal classés est calculé. L'évaluation de la métrique δ_E, pour classer les profils d'expression simulés de l'expérience j, est résumée par la moyenne des taux de mal classés sur les 10 classes $C^{1j}_{\delta_E}$,..., $C^{10j}_{\delta_E}$ obtenues. Pour l'indice de dissimilarité D_k, un classement adaptatif est réalisé. Il consiste à exécuter l'algorithme 10-NN sur l'échantillon S_{ij} pour plusieurs valeurs de k ($k = 0, ..., 6$ avec un pas égale à 0.01) pour apprendre la valeur $k*$ qui minimise le taux de mal classés $C^{ij}_{D_{k*}}$. La moyenne des taux de mal classés, obtenus sur les 10 classes $C^{1j}_{\delta_E}$,..., $C^{10j}_{\delta_E}$, est le paramètre retenu pour évaluer D_k.

5. Résultats et discussion

Donnons d'abord quelques éléments sur les critères considérés. L'asw indique qu'une structure forte (asw proche de 1) ou faible ($asw < 0.5$), des partitions obtenues, est trouvée. Le wbr mesure la compacité (variabilité au sein d'une classe) et séparabilité (variabilité entre les classes). Une bonne partition est caractérisée par une faible valeur de wbr. Enfin, le RI mesure la similarité entre les partitions obtenues après classification et les vraies partitions définies par les S_{ij} (RI=1 pour une forte similarité et RI=0 pas de similarité).

La figure 2 montre que les valeurs moyennes des asw, wbr et RI de la classification baseé sur δ_E se dégradent de l'expérience 1 à 4, montrant l'inadéquation de la distance euclidienne face aux variations complexes. La classification basée sur COR donne, pour les expériences 1 et 2, de fortes structures de partitions avec de très bonnes valeurs des paramètres asw, wbr et RI. Toutefois, cette qualité diminue de façon drastique dans les expériences 3 et 4, montrant la limite du coefficient de corrélation de Pearson face aux variations de tendance ; comme nous l'avions indiqué dans la section 2. Enfin, la meilleure classification et les structures de partitions les plus fortes sont données par CORT et D_k dans les quatre expériences, avec une asw variant dans $[0.8, 1]$, un wbr autour de 0, un RI évoluant dans $[0.83, 1]$.

Pour la méthode de classement : la figure 2, des taux d'erreur, montre que pour les expériences 1 et 2, les quatre métriques sont presque aussi efficaces avec des taux d'erreur de classement autour de 0. Cependant, on note pour les expériences 3 et 4, une très forte augmentation de ce taux d'erreur pour les classements basés sur δ_E, une légère augmentation du taux d'erreur pour les classements basés sur la COR et encore un taux d'erreur de classement autour de 0 pour CORT et D_k. Selon les résultats des 4 expériences, les métriques CORT et D_k peuvent être considérées comme le moyen le plus efficace pour classer les profils d'expression de gènes.

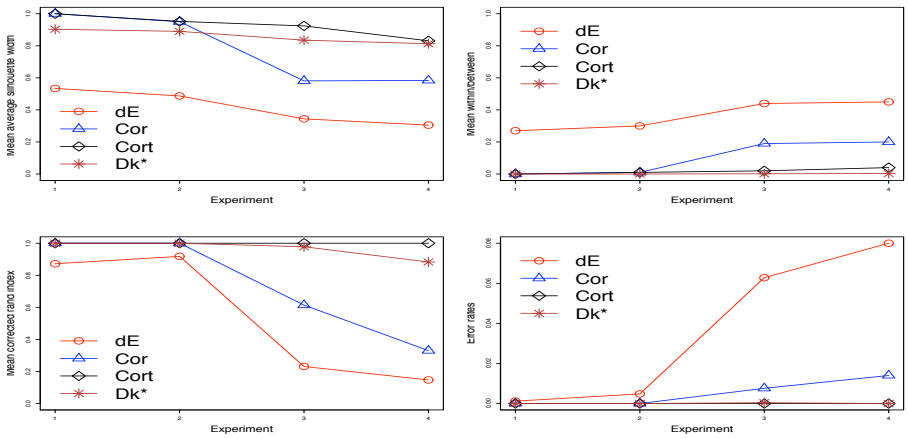


FIG. 2. *Evaluation et comparaison des métriques. En haut, à gauche la figure des valeurs moyennnes des asw et à droite celle des wbr pour les 4 expériences. En bas, à gauche la figure des valeurs moyennnes des RI et à droite celle des taux d'erreur de classement pour les 4 expériences.*

6. Bibliographie

[DOU 07] DOUZAL CHOUAKRIA A., NAGABHUSHAN P., Adaptive dissimilarity index for measuring time series proximity, *Advances in Data Analysis and Classification Journal*, vol. 1, 2007, p. 5–21.

[EIS 98] EISEN M., SPELLMAN P., BROWN P., BOTSTEIN, Cluster analysis and display of genome-wide expression patterns, *Proc.Nalt.Acad.Sci. USA*, , 1998, p. 14863–14868.

[KAU 90] KAUFMAN L., ROUSSEEUW P., *Finding Groups in Data. An Introduction to Cluster Analysis*, John Wiley & Sons, New York, 1990.

[LIU 04] LIU D., UMBACH D. M., PEDDADA S. D., L. L., CROCKETT P. W., WEINBERG C. R., A Random-Periods Model for Expression of Cell-Cycle Genes, *Proc.Natl.Acad.Sci USA*, vol. 101, 2004, p. 7240–7245.

Analyse de la stabilité d'une partition basée sur la cohésion et l'isolation des classes

Lassad El Moubarki[1,3], **Ghazi Bel Mufti**[2], **Patrice Bertrand**[3,4], **Mohamed Limam** [1]

[1] *ISG de Tunis, Larodec*
[2]*ESSEC de Tunis, Cefi*
[3]*Université Paris-Dauphine, Ceremade (UMR CNRS no. 7534)*
[4] *INRIA-Rocquencourt, AxIS*

RÉSUMÉ. *Plusieurs méthodes ont récemment été proposées afin de valider un partitionnement en évaluant sa stabilité. Les mesures de validation utilisant le critère de stabilité diffèrent entre elles non seulement par leurs façons de comparer les partitions, mais aussi par le fait qu'elles sont basées sur différentes définitions de la notion de classe. Partant du principe qu'un manque de stabilité est dû à un défaut de cohésion et/ou d'isolation des classes, nous proposons d'interpréter l'indice de Rand par le degré global d'isolation et de cohésion de la partition. Nous montrons que pour toutes les classes, ces degrés de cohésion et d'isolation se décomposent en fonction des contributions de chaque élément de l'ensemble des données à classer. Nous illustrons notre approche sur des données simulées et sur des données réelles biologiques. En particulier, nous traitons un jeu de données simulées qui illustre le comportement asymptotique des mesures de stabilité qui a été formellement établi dans un article récent de Ben David, von Luxburg et Pál. De plus, nous comparons, sur les jeux de données traités, les différentes mesures de stabilité de partitions obtenues par plusieurs méthodes de classification.*

MOTS-CLÉS : *stabilité d'une partition, isolation, cohésion, échantillonnage aléatoire*

1. Introduction

Récemment, plusieurs méthodes de validation des résultats d'une classification utilisant le critère de stabilité d'une partition ont été proposées (e.g. Levine et Domany [LEV 01], Ben-Hur *et al.* [BEN 02], Tibshirani *et al.* [TIB 01], Bertrand et Bel Mufti [BER 06])). Ces méthodes sont basées sur la comparaison répétitive des partitions obtenues sur des échantillons aléatoires tirés des données de référence. La variabilité entre les partitions obtenues sur les échantillons permet d'évaluer la significativité de la partition obtenue sur les données de référence. Les mesures de stabilité proposées sont souvent utilisées pour valider le choix du nombre de classes k. Néanmoins, des travaux récents de Ben David *et al* [BEN 08] montrent que lorsque la taille des données est importante les méthodes de validation basées sur la stabilité doivent être utilisées avec précaution. En particulier, ces travaux montrent que lorsque la taille des données est importante et lorsque la méthode de classification utilisée possède un optimum unique sur une partition en k classes, alors cette partition est stable même si le nombre de classes n'est pas correct.
Dans ce qui suit, nous définissons nos mesures de cohésion et d'isolation d'une partition. Ensuite, nous présentons une décomposition de l'indice de Rand [RAN 71], qui montre que cet indice mesure conjointement les deux propriétés d'isolation et de cohésion d'une partition. Enfin, nous appliquons notre approche sur un jeu de données artificiel qui illustre les limites des mesures de stabilité soulignées par Ben David *et al.* [BEN 08], ainsi, qu'aux données biologiques *breast-cancer* [MAN 90].

2. Décomposition de l'indice de Rand en composantes mesurant l'isolation et la cohésion d'une partition

Dans ce paragraphe nous commencons par rappeler le principe de définition des indices introduits par Bertrand et Bel Mufti [BER 06] afin d'évaluer la cohésion et l'isolation d'une partition, puis nous adaptons ce principe à l'indice de Rand et décomposons cet indice en deux indices relatifs à la cohésion et l'isolation des classes.

Considérons un ensemble de n objets à classer, noté S. Notons S' l'ensemble de données résultant de la perturbation de S et A_k un algorithme de partionnement en k classes. La partition obtenue en appliquant A_k à S est notée P_k ($A_k(S) = P_k$). Les mesures de stabilité que nous proposons sont basées sur l'estimation de la qualité des deux règles logiques suivantes exprimant respectivement la propriété d'isolation et la propriété de cohésion d'une partition (cf. Bertrand et Bel Mufti [BER 06]) :

(R1) *Règle de cohésion d'une partition* : si deux objects sont classés ensemble dans la partition P_k, alors ils sont classés ensemble dans la partition $A_k(S')$.

(R2) *Règle d'isolation d'une partition* : si deux objets sont séparés dans la partition P_k, alors ils sont séparés dans la partition $A_k(S')$.

Nous proposons ici de mesurer la qualité des règles (R1) et (R2) à l'aide de l'indice de confiance. Rappelons que l'indice de confiance I_{conf} se calcule par la formule $I_{conf} = \mathbb{P}(E|F) = \mathbb{P}(E \cap F)/\mathbb{P}(F)$ si $E \Rightarrow F$ est la règle examinée, la notation $\mathbb{P}$ désignant la loi de probabilité empirique observée sur les données. En notant $I^{co}(P_k,\, A_k(S'))$ (resp. $I^{is}(P_k,\, A_k(S'))$) l'indice de confiance mesurant la qualité de la règle (R1) de cohésion (resp. (R2) d'isolation), nous obtenons :

$$I^{co}(P_k,\, A_k(S')) = \frac{N_{11}}{N_{pb}} \qquad \text{et} \qquad I^{is}(P_k,\, A_k(S')) = \frac{N_{00}}{N_{pw}},$$

où N_{00} est le nombre de paires d'objets séparés dans la partition P_k et dans la partition $A_k(S')$, N_{11} est le nombre de paires d'objets qui sont classés ensemble dans P_k et dans $A_k(S')$, N_{pb} compte le nombre de paires d'objets qui ne sont pas classés ensembles dans P_k et N_{pw} le nombre de paires d'objets qui sont classés ensemble dans P_k et dans $A_k(S')$.

En utilisant comme ensemble de données perturbées un échantillon des données qui est stratifié selon la partition P_k, il est possible de définir de la même manière, des indices évaluant la cohésion et l'isolation de chacune des classes C_i ($i = 1, ..., k$) de la partition P_k. Nous montrons que l'indice I^{co} (resp. I^{is}) évaluant P_k selon le critère de cohésion (resp. isolation), est une moyenne pondérée des indices évaluant les classes de P_k pour la cohésion (resp. isolation).

Considérons à présent l'indice de Rand souvent utilisé pour mesurer la stabilité d'une partition. Il est facile de vérifier que l'indice de Rand de comparaison de deux partitions donné par :

$$Rand(P_k,\, A_k(S')) = N_{00} + N_{11} / \left(\begin{array}{c} n \\ 2 \end{array} \right),$$

est une moyenne pondérée des indices d'isolation et de cohésion. Plus précisément :

$$Rand(P_k,\, A_k(S')) = \frac{N_{pw} I^{co}(P_k,\, A_k(S')) + N_{pb} I^{is}(P_k,\, A_k(S'))}{N_{pw} + N_{pb}}.$$

Les mesures de stabilité d'une partition sont alors obtenues en faisant des comparaisons répététives de la partition de référence P_k à des partitions obtenues sur des échantillons proportionnels stratifiés générés aléatoirement à partir des données de référence S. A noter que les mesures de stabilité sont normalisées par rapport à la mesure de stabilité obtenue sous le modèle nul d'absence de structure en classes dans les données. Plus précisément cette valeur que l'on note $\overline{Stab}_0$ est obtenue en simulant un grand nombre de données de taille n distribuées uniformément dans la plus petite hypersphere contenant les données. Chaque jeu de données simulé est classifié selon le même

algorithme de partionnement A_k en k classes, et une mesure de la stabilité de la partition obtenue est calculée. Enfin nous calculons $\overline{Stab}_0$ comme la moyenne de toutes les mesures de stabilité ainsi obtenues.

Les différentes étapes du processus de réechantillonnage et de classification que nous proposons, sont récapitulées dans le tableau suivant :

ENTRÉES: S, P ou $C \in P$, A_k
SORTIES: $\overline{Stab}_{ref}$, estimation de $Ic(R)$ avec $R \in \{R_1, R_2\}$.

1: **pour** $j = 1$ à N **faire**
2: Tirer un échantillon S'_j de S et calculer $A_k(S'_j)$.
3: Calculer $Ic(R)$ avec $Q_j = A_k(S'_j)$, noté par $stab(P_k, Q_j)$.
4: **fin pour**
5: Calculer la moyenne : $\overline{Stab}_{ref} = \frac{1}{N} \sum_{j=1}^{N} stab(P_k, Q_j)$

Valeur ajustée : $Stab_{ajust} = \dfrac{\overline{Stab}_{ref}(P_k) - \overline{Stab}_0}{1 - \overline{Stab}_0}$

---

3. Expériences

Dans ce paragraphe nous nous proposons d'illustrer notre approche sur un jeu de données artificiel sans structure et sur les données biologiques breast-cancer (cf. [MAN 90]), et de comparer notre approche à la méthode de Ben-Hur *et al.* [BEN 02]. Le jeu de données uniforme est constitué de 500 points simulés uniformement dans le rectangle $[0, 2] \times [0, 1]$. Les résultats montrent que tous les indices ajustés, d'isolation, de cohesion et de Rand, fournissent des valeurs faibles inférieures à 0.5. Ce résultat permet de conclure qu'il n'y a pas de structure en classes dans les données. Le graphique de la Figure 1 nous montre que la méthode single link possède les résultats les plus instables du point de vue du critère de l'isolation de la partition. Cependant, l'indice de Ben Hur *et al.* [BEN 02] ajusté, appliqué avec la méthode *k-means*, possède une valeur maximale proche de 0.9 au niveau de la partition en deux classes. Cette forte stabilité qui est justifiée par l'unicité de l'optimum fourni par la méthode *k-means* (cf. Ben David *et al.* [BEN 08]), semble donc être un artefact de cette méthode de validation par stabilité.

Les données breast-cancer (cf. [MAN 90]) contiennent 683 individus décrits par 10 variables. Cette base de données contient deux classes naturelles de tumeurs, bénin et malin. Les résultats obtenus montrent que l'indice de Rand ajusté, fournit pratiquement la même valeur pour les partitions en deux, en trois et en quatre classes, et ce pour la plupart des algorithmes de classification usuels. Cependant, en observant les courbes d'isolation et de cohésion nous remarquons que l'isolation de P_2 est nettement meilleure que celle de P_3. De même nous observons que la cohésion de P_2 est meilleure que celle de P_4. Ainsi, nous concluons que la partition en deux classes est préférée aux partitions en trois et en quatres classes.

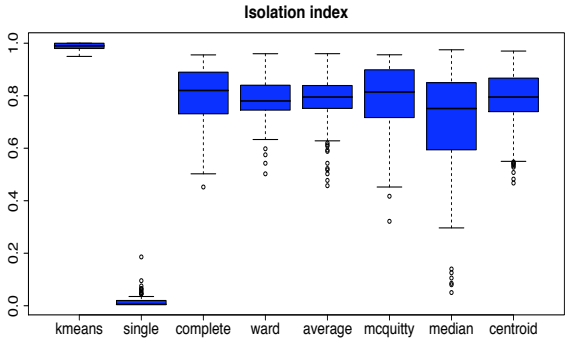


FIGURE 1. *Box-plots décrivant la distribution des valeurs, non ajustées, de l'indice d'isolation de la partition en deux classes pour les différentes méthodes de classification utilisées.*

Enfin, l'approche proposée étant indépendante du choix de la méthode de bruitage, les résultats obtenus pour d'autres types de bruitage, comme l'ajout de bruit seront présentés.

4. Conclusion

Nous avons proposé une décomposition additive de l'indice de Rand qui, dans les études sur la stabilité en classification, est souvent utilisé sous la forme du taux d'éléments bien classés. Nous avons montré à travers des exemples que cette décomposition fournit plus d'information sur la qualité de la partition. L'autre interêt de cette décomposition est qu'elle est de type linéaire, et donc qu'elle est vérifiée pour les moyennes globales calculées sur les N jeux de données perturbées, les perturbations pouvant aussi bien être obtenues par échantillonnage des données de référence que par ajout d'un bruit.
Néanmoins, d'autres simulations intensives sur des jeux de données réelles et artificielles sont à faire, pour mieux évaluer l'approche proposée.

5. Bibliographie

[BEN 02] BEN-HUR A., ELLISSEEFF A., GUYON I., A stability based method for discovering structure in clustered data, *Pacific Symposium on Biocomputing*, 2002, p. 6–17.

[BEN 08] BEN-DAVID S., VON LUXBURG U., Relating clustering stability to properties of cluster boundaries, SERVEDIO R., ZHANG T., Eds., *Proceedings of the 21st Annual Conference on Learning Theory*, Berlin, 2008, Springer, p. 379–390.

[BER 06] BERTRAND P., BEL MUFTI G., Loevinger's measures for assessing cluster stability, *Computational Statistics and Data Analysis*, vol. 50, n^o 4, 2006, p. 992–1015.

[LEV 01] LEVINE E., DOMANY E., Resampling Method for Unsupervised Estimation of Cluster Validity, *Neural Computation*, vol. 13, 2001, p. 2573–2593.

[MAN 90] MANGASARIAN O. L., WOLBERG W. H., Cancer diagnosis via linear programming, *SIAM News*, vol. 23, n^o 5, 1990, p. 1–18.

[RAN 71] RAND W., Objective criteria for the evaluation of clustering methods, *journal of the American Statistical Association*, vol. 66, 1971, p. 846-850.

[TIB 01] TIBSHIRANI R., WALTHER G., BOTSTEIN D., BROWN P., Cluster validation by prediction strength, Ca, 2001, Technical report : Department of Statistics, Stanford University.

Indice de distance sur les structures hiérarchiques semi-floues

Sahondra Ravonialimanana[1], **Ralambondrainy Henri**[2]

1. Université de Fianarantsoa, 2. IREMIA, Université de la Réunion
1.B.P.1486, Fianarantsoa Madagascar, 2. 15 av Cassin, 97715 Saint-Denis Message Cedex 9, Réunion
rafilipojs@yahoo.fr,ralambon@univ-reunion.fr

RÉSUMÉ. *L'indice que nous proposons, dans ce papier, est fondé sur le degré d'imbrication de hiérarchies semi-floues, inspiré par l'indice de Kosko. Nous utilisons cet indice pour comparer des partitions et des hiérarchies .*

MOTS-CLÉS : *Classification, distance, partition,hiérarchie, ensemble flou*

1. Introduction

Les méthodes d'Analyse des données qui structurent les données en une hiérarchie d'ensembles sont nombreuses. Il importe de pouvoir comparer différentes hiérarchies. Après avoir rappelé la définition d'une hiérarchie semi-floue qui est une généralisation d'une hiérarchie classique, nous proposons des indices permettant de mesurer la proximité entre deux hiérarchies semi-floues. Nous faisons une étude comparative, sur des partitions, de notre indice avec celui de Rand, et sur des hiérarchies avec celle proposée par Leclerc.

2. Imbrication de sous-ensembles flous

Soit $\mathcal{X}$ un référentiel donné, de cardinal n. On note par m_h la fonction d'appartenance d'un sous ensemble flou h de $\mathcal{X}$ et par $\mathcal{F}(\mathcal{X})$ l'ensemble des sous ensembles flous de $\mathcal{X}$. La magnitude d'une classe floue h est définie par $M(h) = \sum_{x\in\mathcal{X}} m_h(x)$. Pour un ensemble fini classique H, on a $M(H) = card(H)$. Le degré d'imbrication $S(h,h')$, selon Kosko[1], d'une classe h dans une classe h' est mesuré par la proportion, au sens de la magnitude, des éléments de h appartenant à h'. $S(h,h') = \frac{M(h\bigcap h')}{M(h)}$. Lorsque H et H' sont des sous ensembles finis classiques, l'imbrication de H dans H' s'écrit : $S(H,H') = \frac{M(H\bigcap H')}{M(H)} = \frac{Card(H\bigcap H')}{Card(H)}$. Les propriétés suivantes montrent la compatibilité entre l'indice d'imbrication S et l'ordre d'inclusion.

Propriétés 2.1 Pour tout $h, h', h'' \in \mathcal{F}(\mathcal{X})$ et pour tout $x \in \mathcal{X}$ on a :
1- $S(\{x\},h) = m_h(x)$; **2-** $h \subseteq h' \Leftrightarrow S(h,h') = 1$; **3-** $h' \subset h \Leftrightarrow 0 < S(h,h') < 1$; **4-** $h\bigcap h' = \emptyset \Leftrightarrow S(h,h') = S(h',h) = 0$; **5-** $h' \subseteq h'' \Rightarrow S(h,h') \leq S(h,h'')$.

3. Degré d'imbrication d'un sous-ensemble flou dans une hiérarchie semi-floue

Nous avons défini dans [3] une hiérarchie semi-floue $\mathcal{F}$ comme une famille de sous ensembles flous de $\mathcal{F}(\mathcal{X})$ vérifiant :$1) m_X = 1 \in \mathcal{F}; 2) \forall h,h' \in \mathcal{F} : m_{h\cap h'} = m_h \wedge m_{h'} \in \{m_h, m_{h'}, 0\}; 3) \forall x \in \mathcal{X} : m_{\{x\}} \in \mathcal{F}$. Le support d'une hiérarchie semi-floue est une hiérarchie classique. Soient un sous ensemble flou Y et une hiérarchie semi-floue $\mathcal{F}$. On note $\overline{Y}$ la plus petite classe de $\mathcal{F}$ contenant Y. Nous proposons de mesurer le degré d'imbrication de Y à la hiérarchie semi-floue $\mathcal{F}$ par :
$m_{\mathcal{F}}(Y) = S(Y,\mathcal{F}) = S(\overline{Y},Y)$. On montre facilement que $m_{\mathcal{F}}(Y)$ est une fonction d'appartenance au sens des

ensembles flous. Pour mesurer le degré d'imbrication de Y à la hiérarchie semi-floue $\mathcal{F}$, une autre possibilité est de considérer l'ensemble des classes Y^l de $\mathcal{F}$ qui minorent Y dans $\mathcal{F}$. On note par $\underline{Y^l}$ l'ensemble des minorants maximaux de Y. Le degré d'imbrication d'un sous ensemble flou Y à une hiérarchie semi-floue $\mathcal{F}$ est mesuré par : $m_{\mathcal{F}}(Y) = S'(Y, \mathcal{F}) = \frac{1}{|\underline{Y^l}|} \sum_{y \in \underline{Y^l}} S(Y, y)$, c.-à-d. la moyenne des degrés d'imbrication de Y dans les classes de $\mathcal{F}$ qui minorent Y. La fonction $m_{\mathcal{F}}(Y)$ est aussi une fonction d'appartenance au sens des sous ensembles flous. Ces indices s'étendent aux hiérarchies semi-floues . La magnitude d'une hiérarchie semi-floue $\mathcal{F}$ a pour expression $M(\mathcal{F}) = \sum_{Y \in \mathcal{F}(\mathcal{X})} m_{\mathcal{F}}(Y)$. L'imbrication de $\mathcal{F}_1$ dans $\mathcal{F}_2$ est alors mesurée par : $S(\mathcal{F}_1, \mathcal{F}_2) = \frac{M(\mathcal{F}_1 \bigcap \mathcal{F}_2)}{M(\mathcal{F}_1)}$ où $M(\mathcal{F}_1 \bigcap \mathcal{F}_2) = \sum_{Y \in \mathcal{F}_1 \cup \mathcal{F}_2} m_{\mathcal{F}_1} \bigwedge m_{\mathcal{F}_2}(Y)$. On remarque que l' imbrication entre 2 hiérarchies semi-floues vérifie les mêmes propriétés que celles énoncées dans 2.1.

4. Indice de distance sur des structures hiérarchiques

4.1. *Partitions semi-floues*

Une partition semi-floue $P = \{C_1, ..., C_k\}$ est une famille finie de sous ensembles flous de $\mathcal{X}$ à support deux à deux disjoints et telle que la réunion de tous les supports est égale à X. Si la notion d'hiérarchie floue n'est pas connue, diverses définitions de partitions floues ont été proposées dans la littérature. La structure de partition semi-floue que nous proposons est intermédiaire entre celle de partition classique et celle de partition floue définie par Bezdec par exemple. Une partition semi-floue est une partition classique dans le sens qu'un individu appartient à une seule classe, mais le degré d'appartenance de cet élément à la classe est précisé. Ce degré d'appartenance sera par exemple la qualité de typicité de l'élément dans la classe, mesurée par sa proximité au prototype de la classe considérée. Cependant, une partition semi-floue n'est pas une partition floue selon Bezdec car la somme des degrés d'appartenance n'est pas égale à un.

A chaque partition semi-floue P , nous pouvons associer la hiérarchie semi-floue H_P telle que $H_P = \{\{m_{\{x\}}\}_{x \in \mathcal{X}}, C_1, ..., C_k, \mathcal{X}\}$. Soit Y un sous ensemble flou de $\mathcal{F}(X)$. Nous allons préciser l'expression de la fonction d'appartenance de Y dans H_P suivant S et S'.

4.1.1. *Fonction d'appartenance*

1. Cas de S

La fonction d'appartenance de Y dans H_P est donné par $m_{H_P}(Y) = S(Y, H_P) = S(\overline{Y}, Y)$. On a alors :
Si $\forall i \in \{1, ..., k\}, Y \nsubseteq C_i$ ce qui implique $\overline{Y} = \mathcal{X} \Rightarrow m_{H_P}(Y) = S(\overline{Y}, Y) = \frac{M(Y)}{M(\mathcal{X})}$;
Si $\exists i \in \{1, ..., k\}$, tel que $Y \subset C_i$, alors $\overline{Y} = C_i \Rightarrow m_{H_P}(Y) = S(\overline{Y}, Y) = \frac{M(Y)}{M(C_i)}$;
Si $\exists j \in \{1, ..., k\}$, tel que $Y = C_j$, alors $\overline{Y} = C_j \Rightarrow m_{H_P}(Y) = S(\overline{Y}, Y) = 1$.

2. Cas de S'

La fonction d'appartenance a pour expression : $m_{H_P}(Y) = S'(Y, \mathcal{F}) = \frac{1}{|\underline{Y^l}|} \sum_{y \in \underline{Y^l}} S(Y, y)$.

Si $\exists J \subseteq \{1, ..., k\}$, tel que $Y = \bigcup_{j \in J} C_j$, alors $\underline{Y^l} = \{m_{C_k}\}_{k \in K}$ et $m_{H_P}(Y) = 1$.

Sinon, $\underline{Y^l} = \{m_{\{y\}}\}_{y \in Y}$ et $m_{H_P}(Y) = \frac{1}{|\underline{Y^l}|} \sum_{y \in Y} \frac{m_Y(y) \wedge m_{\{y\}}(y)}{M(Y)} = \frac{1}{|\underline{Y^l}|} \frac{\sum_{y \in Y} m_{\{y\}}(y)}{M(Y)}$
avec $m_{\{y\}}(y) = \inf_k m_{C_k}(y)$.
Dans tous les cas la fonction d'appartenance de Y dans H_P s'obtient en calculant le degré d'imbrication de Y dans H_P. Dans la suite, pour toute partition semi-floue P, on note m_{H_P} par m_P.

4.1.2. *Comparaison de deux partitions semi-floues*

Soient 2 partitions semi-floues C et D de fonctions d'appartenance respectives m_C et m_D.

Distances entre deux partitions semi-floues

Puisque toute classe de $C \cup D$ admet un degré d'imbrication par rapport à chaque partition , alors on peut calculer la distance entre les deux partitions semi-floues C et D. Elle est donnée par : $d(C, D) = \sum_{h \in C \cup D} |m_C(h) - m_D(h)|$.

Imbrication de deux partitions semi-floues

Le degré d'imbrication de la partition semi-floue C dans D est mesurée par la proportion des classes floues de C imbriquée dans D.

$$S(C,D) = \frac{M(H_C \bigcap H_D)}{M(H_C)} = \frac{\sum_{Y \in H_C \cup H_D} m_{H_C} \wedge m_{H_D}(Y)}{M(H_C)} = \frac{\sum_{C_k \in H_C} m_{H_D}(C_k) + \sum_{D_l \in H_D} m_{H_C}(D_l)}{M(H_C)}$$

Applications et comparaison avec la méthode de Rand

On suppose que les 2 partitions semi-floues C et D sont déterminées par les tableaux Table 1. et Table 2.

D	x_1	x_2	x_3	x_4	x_5	x_6	x_7
d_1	1	0	0	0	0	0	0
d_2	0	0.25	0.75	0	0	0	0
d_3	0	0	0	1	0	0	0
d_4	0	0	0	0	0.75	1	0.5

C	x_1	x_2	x_3	x_4	x_5	x_6	x_7
c_1	1	0	0	0	0	0	0
c_2	0	0.5	0	0	0	0	0
c_3	0	0	0.5	0.5	0	0	0
c_4	0	0	0	0	1	0.75	1

TABLE 1. *Les fonctions d'appartenance des éléments dans chaque classe de D et C*

$m_D(c_i)$

	c_1	c_2	c_3	c_4
S	1	0.25	0.14	0.89
S'	1	0.25	0.5	0.73

$m_C(d_j)$

	d_1	d_2	d_3	d_4		
S	1	0.14	0.5	0.75	$d(C,D)$	3.33
S'	1	0.4	0.5	0.89	$d(C,D)$	2.71

TABLE 2. *Les fonctions d'appartenance des classes des deux partitions D et C.*

Considérons les partitions classiques correspondantes à C et D. Notons par c_{x_i} la classe de C contenant x_i et par d_{x_i} celle de D. On dit que deux éléments x_i et x_j sont en accord positif si $x_j \in c_{x_i}$ et $x_j \in d_{x_i}$, et en accord négatif, si $x_j \notin c_{x_i}$ et $x_j \notin d_{x_i}$. Soient a le nombre de paires d'accords positifs, et b le nombre de paires d'accords négatifs. L'indice brut de Rand [4] est déterminé par : $R = \frac{(a+b)}{\frac{n(n-1)}{2}} = 0.85$. Tandis que l'indice de Rand corrigé par Hubert Arabie [5] de C et D est égal à $HA(C,D) = 0.69$. Ces 2 indices représentent le pourcentage global de paires en accord pour les deux partitions.

Le degré d'imbrication de la partition C dans D a pour valeur : $S(C,D) = \frac{M(C \cap D)}{M(C)} = 0.74$. Il mesure la proportion, au sens de la magnitude, des classes de C imbriquées dans D. Le degré d'appartenance d'un élément dans une classe peut-être considéré comme le degré de typicité de l'élément à la classe (par exemple la distance normalisé de l'élément au prototype de la classe). La différence entre le degré d'imbrication proposé et l'indice de Rand consiste à la prise en compte cette typicité, lorsque cette information est disponible, dans le calcul des imbrications de toutes les classes de C dans D. L'indice de Rand considère simplement les paires d'éléments qui sont en accord.

4.2. *Application sur les hiérarchies classiques*

Soient C et D deux hiérarchies classiques (Figure 1). En utilisant le degré d'imbrication, on peut comparer C et D de deux façons différentes :

a- soit en calculant l'imbrication de C dans D : $S(C,D) = \frac{M(C \bigcap D)}{M(C)}$ où $M(C \bigcap D) = \sum_{Y \in C \cup D} m_C \bigwedge m_D(Y)$.

b- ou en calculant la distance de Hamming entre C et D : $d(C,D) = \sum_{h \in C \bigcup D} |m_C(h) - m_D(h)|$.

D'autre part, Leclerc [2] a proposé diverses métriques fondées sur des pondérations des classes de la hiérarchie H. Pour un indice de pondération i, la distance $d_i(H,H') = i(H) + i(H') - 2i(H \cap H')$. Considérons l'indice γ

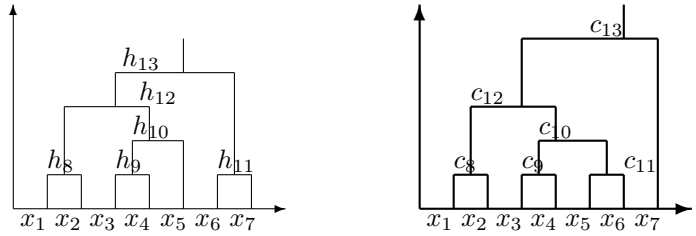


FIGURE 1. *Hiérarchies classiques D et C.*

S	c_8	c_9	c_{10}	c_{11}	c_{12}	c_{13}	h_8	h_9	h_{10}	h_{11}	h_{12}	h_{13}
m_D	1	1	0.57	0.29	0.86	1	1	1	1	1	1	1
m_C	1	1	1	1	1	1	1	1	0.75	0.29	0.71	1
S'	c_8	c_9	c_{10}	c_{11}	c_{12}	c_{13}	h_8	h_9	h_{10}	h_{11}	h_{12}	h_{13}
m_D	1	1	0.5	0.5	0.34	1	1	1	1	1	1	1
m_C	1	1	1	1	1	1	1	1	0.5	0.5	0.5	1

TABLE 3. *Les fonctions d'appartenance des classes des hiérarchies classiques D et C*

tel que $\gamma(H)$ soit le nombre de classes non triviales de l'hiérarchie . Si on note par F l'ensemble des singletons de H et par $I_H = H - F$, l'ensemble des classes intérieures de H. On a : $d_\gamma(D,C) = \gamma(C) + \gamma(D) - 2\gamma(D \cap C)) = |I_D| - 1 + |I_C| - 1 - 2(|I_{D \bigcap C}| - 1)$. Pour les hiérarchies C, D des figures 3 et 4, on obtient : $d_\gamma(D,C) = 6$. En utilisant le degré d'appartenance des classes D et C, selon la table 3, on trouve pour nos indices : **1-** Cas de S : $d(D,C) = 2.54$, **2-** Cas de S' : $d(D,C) = 3.17$. La méthode de Leclerc calcule la valeur de la distance différence symétrique entre les deux hiérarchies, cad le nombre de classes non en commun. Dans ce mode de calcul, deux classes de C et D ayant un, ou plusieurs éléments non en commun, sont traitées de la même manière dans le calcul de la distance. Nos distances par contre, prennent en compte la typicité de chaque élément d'une classe de la hiérarchie pour mesurer les degrés d'imbrication des différentes classes entre elles et synthétiser les différences dans la valeur de la distance. Ce qui explique les valeurs plus faibles de ces distances.

4.3. *Conclusion*

Nous avons proposé des indices de distance sur des structures hiérarchiques semi-floues. Ces dernières peuvent être vues comme des généralisations des structures hiérarchiques classiques : partitions, hiérarchies totales, hiérarchies faibles, pyramides, ... pour lesquelles la typicité des éléments des classes est prise en compte. Les indices proposés sont donc applicables à ces structures hiérarchiques classiques. Les comparaisons que nous avons faites sur les partitions, et hiérarchies classiques, avec des indices classiques montrent la qualité de ces indices qui sont plus précis dans la mesure des différences entre deux structures hiérarchiques. Un autre intérêt de ces indices est qu'ils permettent de comparer des hiérarchies faibles, pyramides, ... structures hiérarchiques pour lesquelles, on ne dispose pas de mesures de comparaison. Une structure hiérarchique semi-floue est un ensemble flou, dans cet article, nous avons choisi la distance de Hamming sur les fonctions d'appartenance mais d'autres distances sur les ensembles flous sont envisageables et seraient intéressantes à étudier.

Bibliographie

[1] B. Kosko. *Neuronal networks and fuzzy systems* . Prentice-hall International Editions,1992

[2] B. Leclerc. *La comparaison des hiérarchies :indices et métriques.* Maths et Sciences humaines, tome 92 (1985) p.5-40

[3] S. Ravonialimanana, H. Ralambondrainy, J. Diatta. *Indice de comparaison de hiérarchies semi-floues : application aux hiérarchies classiques.* RNTI, Réf. 831, 2008.

[4] Genane Youness. *Contributions à une méthodologie de comparaison de partitions.* Thèse de Doctorat en Science de l'Université PARIS 6, 2004

[5] L.Denoeud, H. Garreta, A Guénoche *Comparison of distance indices between partitions* Proceedings of IFCS'2006.

Détermination du nombre de classes d'une partition floue par mesures de séparation et de chevauchement fondées sur des opérateurs d'agrégation

Carl Frélicot, Hoel Le Capitaine

Laboratoire MIA – Université de La Rochelle
Avenue Michel Crépeau, 17042 La Rochelle Cedex 01
{cfrelico,hlecap01}@univ-lr.fr

RÉSUMÉ. *La validation des résultats est une étape cruciale en classification. Dans le cas des méthodes par partition, elle passe par la détermination du nombre de classes. La littérature abonde en indices dont l'optimum est la solution à ce problème, en particulier dans le cas des partitions floues obtenues par exemple par l'algorithme FCM. Ils combinent généralement des mesures de compacité et de séparation calculées à partir de la matrice de partition floue U et/ou des centres des classes et/ou des données elles-mêmes. Nous présentons ici un nouvel indice de validité fondé sur une mesure de séparation et une mesure de chevauchement des classes qui n'utilisent pas les centres. Ces mesures agrègent les seuls éléments de U à l'aide d'opérateurs algébriques adaptés combinant des normes triangulaires. Les propriétés de l'indice proposé sont données. Les résultats obtenus sur des jeux de données artificielles et réelles de nature et structure variées prouvent l'efficacité de l'indice proposé, comparativement à de nombreux indices de la littérature.*

MOTS-CLÉS : *classification automatique, validation, partition floue, algorithme FCM, normes triangulaires*

1. La validation de partition floue

Le partitionnement est une approche de la classification non supervisée ayant pour but de trouver une certaine structure de groupes dans un ensemble $X = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$ de n points en dimension p. Par exemple, le très usité algorithme des *C-Moyennes Floues* (FCM, [BEZ 81]) partitionne itérativement X en $c > 1$ groupes représentés par leur centre $\mathbf{v}_i \in \mathbb{R}^p$ et fournit, pour chaque $\mathbf{x}_k$, un vecteur $\mathbf{u}_k = {}^t(u_{1k}, \cdots, u_{ck})$ de degrés d'appartenance aux groupes, ces degrés étant contraints par $\sum_{i=1}^{c} u_{ik} = 1$ ($\forall k = 1, n$) et $0 < \sum_{k=1}^{n} u_{ik} < n$ ($\forall i = 1, c$). Pour c fixé par l'utilisateur, FCM est une application : $X \mapsto (U, V)$ où les matrices $U = [\mathbf{u}_1, ..., \mathbf{u}_n]$ et $V = [\mathbf{v}_1, ..., \mathbf{v}_c]$ sont obtenues par minimisation itérative de la fonctionnelle $J_m(U, V, X) = \sum_{k=1}^{n} \sum_{i=1}^{c} u_{ik}^m ||\mathbf{x}_k - \mathbf{v}_i||^2$ où $m > 1$ est un paramètre rendant la partition résultante plus ou moins floue. Valider un partitionnement flou (U, V) consiste à évaluer s'il reflète ou non la structure de X. La plupart des travaux concernent le problème de la détermination du nombre optimal c_{opt} de groupes et la littérature regorge d'*Indices de Validation de Classification* (CVI) pour FCM et ses dérivés. Étant donné un CVI, la procédure de validation se réduit généralement à :

(1) choisir les bornes c_{min} et c_{max} d'un intervalle de valeurs possibles pour c
(2) pour $c = c_{min}$ to c_{max} : exécuter FCM et calculer CVI(c) à partir de (X, U, V)
(3) sélectionner c_{opt} tel que CVI(c_{opt}) est optimal et valider le partitionnement correspondant (U, V).

Les CVI sont souvent classifiés selon le type d'information qu'ils manipulent : U, (U, V) ou (U, V, X), ou bien selon la/les propriété(s) de la structure qu'ils privilégient : *compacité* C et/ou *séparation* S des groupes. Si on considère seulement C, la meilleure partition est celle où il y a autant de groupes que de points. Inversement, si seule S est prise en compte, la partition optimale consiste en un seul groupe. Dans le tableau 1, nous reportons les CVIs de la littérature auquel nous comparons celui que nous proposons.

2. L'indice de validation proposé

Il a été établi que les mesures de séparation qui ne prennent en compte que les distances entre centres V conduisent à des difficultés d'interprétation de la structure [KIM 04], c'est pourquoi nous proposons de n'utiliser que U. Pour chaque $\mathbf{x}_k$, nous définissons une mesure de séparation S_k quantifiant le degré de chevauchement

du groupe le plus probable, c'est-à-dire celui correspondant au degré d'appartenance le plus élévé par rapport aux $(c-1)$ autres. De manière évidente, une valeur faible de S_k indique une grande séparation du groupe le plus probable par rapport aux autres. Étant donné un partitionnement (U,V) de X, chaque u_{ik} définit la similarité de $\mathbf{x}_k$ au prototype $\mathbf{v}_i$. L'opérateur max est généralement appliqué aux composantes de $\mathbf{u}_k$ pour sélectionner le groupe auquel $\mathbf{x}_k$ doit être associé. Les valeurs plus petites interagissant avec la plus grande, une telle association exclusive n'est pas efficace et il semble opportun d'évaluer à quel point $\mathbf{x}_k$ peut appartenir à plusieurs groupes et à combien par une mesure de chevauchement C_k. Pour définir à la fois S_k et C_k, nous proposons d'utiliser l'opérateur défini dans [MAS 08] afin de qualifier le rejet d'ambiguité dans un autre contexte, celui de la classification supervisée. Soient $\mathcal{P}$ l'ensemble des parties de $I=\{1,2,\cdots,c\}$ et $\mathcal{P}_l=\{A\in\mathcal{P} : card(A)=l\}$, alors l'opérateur *OU flou d'ordre* l associe à $\mathbf{u}_k$ une valeur $\overset{l}{\perp}(\mathbf{u}_k)\in[0,1]$ définie par :

$$\overset{l}{\underset{i=1,c}{\perp}} u_{ik} = \underset{A\in\mathcal{P}_{l-1}}{\top}\left(\underset{j\in I\setminus A}{\perp} u_{jk}\right), \text{ où } (\top,\perp) \text{ est un couple de normes triangulaires duales.} \quad (1)$$

Rappelons qu'une norme triangulaire (t-norme) est un opérateur binaire $\top : [0,1]^2 \to [0,1]$ commutatif, associatif, non décroissant ayant 1 pour élément neutre. Une conorme triangulaire (t-conorme) est l'opérateur dual $\perp : [0,1]^2 \to [0,1]$ vérifiant les mêmes axiomes excepté que son élément neutre est 0, voir [KLE 05]. Les résultats que nous donnons à la section 3 ont été obtenus avec les normes *Standard*, *Algébrique* et de *Hamacher* ($\gamma\in[0,+\infty[$) définies respectivement par :
• $a_1 \top_S a_2 = min(a_1,a_2)$ et $a_1 \perp_S a_2 = max(a_1,a_2)$,
• $a_1 \top_A a_2 = a_1\, a_2$ et $a_1 \perp_A a_2 = a_1 + a_2 - a_1\, a_2$,
• $a_1 \top_H a_2 = \frac{a_1\, a_2}{\gamma+(1-\gamma)(a_1+a_2-a_1\, a_2)}$ et $a_1 \perp_H a_2 = \frac{a_1+a_2-a_1\, a_2-(1-\gamma)a_1\, a_2}{1-(1-\gamma)a_1\, a_2}$.
Avec les normes standard, il est démontré dans [MAS 08] que (1) vaut la $l^{\text{ème}}$ plus grande valeur de $\mathbf{u}_k$, soit $u_{(l)k}$.

TAB. 1. *Critères de validation : nom et référence, formule, optimalité et type de combinaison de propriétés.*

Critère		Optim.	Type
Coef. de Partition Normalisé [ROU 78]	$NPC(c) = \dfrac{-1+\frac{c}{n}\sum_{k=1}^{n}\sum_{i=1}^{c} u_{ik}^2}{c-1}$	max	C
Entropie de Partition Norm. [DUN 77]	$NPE(c) = -\dfrac{\sum_{k=1}^{n}\sum_{i=1}^{c} u_{ik}\log(u_{ik})}{n-c}$	min	C
Indice de Xie-Beni [XIE 91]	$XB(c) = \dfrac{J_m(U,V)/n}{\min_{i,j=1,c;j\neq i}\ \|\|\mathbf{v}_i-\mathbf{v}_j\|\|^2}$	min	C/S
Indice de Fukuyama et Sugeno [FUK 89]	$FS(c) = J_m(U,V) - \sum_{k=1}^{n}\sum_{i=1}^{c} u_{ik}^m \|\|\mathbf{v}_i-\overline{\mathbf{v}}\|\|^2$	min	$C-S$
Hyper-Volume Flou [GAT 89]	$FHV(c) = \sum_{i=1}^{c}\left(\dfrac{\sum_{k=1}^{n} u_{ik}^m\,(\mathbf{x}_k-\mathbf{v}_i)\,^t(\mathbf{x}_k-\mathbf{v}_i)}{\sum_{k=1}^{n} u_{ik}^m}\right)^{1/2}$	min	C
Indice de Bensaid et al. [BEN 96]	$SC(c) = \sum_{i=1}^{c}\dfrac{\sum_{k=1}^{n} u_{ik}^m\|\|\mathbf{x}_k-\mathbf{v}_i\|\|^2}{\sum_{k=1}^{n} u_{ik}\sum_{j=1}^{c}\|\|\mathbf{v}_i-\mathbf{v}_j\|\|^2}$	min	C/S
Indice de Kwon [KWO 98]	$K(c) = \dfrac{J_m(U,V)+\frac{1}{c}\sum_{i=1}^{c}\|\|\mathbf{v}_i-\overline{\mathbf{v}}\|\|^2}{\min_{i,j=1,c;j\neq i}\ \|\|\mathbf{v}_i-\mathbf{v}_j\|\|^2}$	min	C/S
Indice de Wu et Yang [WU 05]	$WY(c) = \sum_{i=1}^{c}\sum_{k=1}^{n}\dfrac{u_{ik}^2}{\min_{i=1,c}\left(\sum_{k=1}^{n} u_{ik}^2\right)} - \sum_{i=1}^{c}\exp\left(-\min_{j\neq i}\dfrac{\|\|\mathbf{v}_i-\mathbf{v}_j\|\|^2}{\frac{1}{c}\sum_{i=1}^{c}\|\|\mathbf{v}_i-\overline{\mathbf{v}}\|\|^2}\right)$	max	$C-S$
Indice de Pakira et al. [PAK 04]	$PBM(c) = \left(\dfrac{\max_{i,j=1}^{c}\|\|\mathbf{v}_j-\mathbf{v}_i\|\|}{c}\ \dfrac{\sum_{k=1}^{n}\|\|\mathbf{x}_k-\overline{\mathbf{v}}\|\|}{J_m(U,V)}\right)^2$	max	S/C

Par calculs successifs de (1) pour différentes valeurs de l, nous obtenons une combinaison d'ordre de degrés de chevauchement pour un point $\mathbf{x}_k$ donné. Afin de déterminer le degré total de chevauchement C_k, il suffit d'évaluer quel ordre (≥ 2) implique le plus grand chevauchement, par exemple à l'aide d'une disjonction floue. Pour la séparation S_k de chaque $\mathbf{x}_k$, nous proposons d'utiliser aussi (1) avec $l=1$, qui évalue le chevauchement d'un seul groupe, c'est à dire sa séparation vis-à-vis des autres, puisque $\mathbf{u}_k$ est normalisé. Cette agrégation, qui correspond

à la disjonction floue des degrés d'appartenance, sélectionne donc le groupe le plus probable. Finalement, nous définissons le nouveau CVI à valeurs dans $[0,1]$ comme la moyenne arithmétique des rapports C_k/S_k :

$$CSI_{\perp}(c) = \frac{1}{n}\sum_{k=1}^{n} \overset{1}{\underset{l=2,c}{\perp}}\Big(\overset{l}{\perp}(\mathbf{u}_k)\Big) \Big/ \overset{1}{\perp}\Big(\mathbf{u}_k\Big) \tag{2}$$

Comme une faible valeur de C_k et une valeur élevée de S_k indiquent que $\mathbf{x}_k$ appartient à un groupe bien séparé et non chevauché par d'autres, l'indice CSI est à minimiser selon la procédure rappelée à la section 1 et le nombre optimal de classes est donné par $argmin_{c=c_{min},c_{max}} CSI_{\perp}(c)$. Il vient facilement qu'avec les normes standard $C_k = u_{(2)k}$ et $S_k = u_{(1)k}$, puisque $\overset{l}{\perp}(\mathbf{u}_k) = u_{(l)k}$, et donc $CSI_{\perp_S}(c) = \frac{1}{n}\sum_{k=1}^{n} u_{(2)k}/u_{(1)k}$.
Si la partition est stricte, i.e. $u_{ik} \in \{0,1\}$, une valeur vaut 1 et les autres 0 : $u_{(1)k} = 1$ et $u_{(2)k} = \cdots = u_{(c)k} = 0$. Alors, quel que soit le couple $(\top, \perp)$, $C_k = \Big(\overset{2}{\perp}\big(1,0,...,0\big)\Big)\overset{1}{\perp}...\overset{1}{\perp}\Big(\overset{c}{\perp}\big(1,0,...,0\big)\Big) = \overset{1}{\perp}(\underbrace{0,...,0}_{c-1 \text{ fois}}) = 0$ car 0 est absorbant pour $\top$ et $S_k = 1$ car 1 est absorbant pour $\perp$, par conséquent $CSI_{\perp}(c) = 0$.
Inversement, si la partition est totalement floue, i.e. $u_{ik} = \frac{1}{c}$ $(\forall i = 1,c)$, $CSI_{\perp}(c)$ est bornée supérieurement mais la borne dépend des normes car C_k en dépend. Dans le cas des normes standard, $\overset{l}{\perp}(\mathbf{u}_k) = u_{(l)k}$ et alors $C_k = \Big(\overset{2}{\perp}\big(\frac{1}{c},...,\frac{1}{c}\big)\Big)\overset{1}{\perp}...\overset{1}{\perp}\Big(\overset{c}{\perp}\big(\frac{1}{c},...,\frac{1}{c}\big)\Big) = \overset{1}{\perp}(\underbrace{\frac{1}{c},...,\frac{1}{c}}_{c-1 \text{ fois}}) = \frac{1}{c}$, ainsi que S_k, et par conséquent $CSI_{\perp_S}(c) = 1$.

3. Résultats et conclusion

Nous testons l'indice (2) et le comparons à ceux du tableau 1 sur des jeux de données artificielles représentés à la figure 1 et jeux de données réelles variés [BLA 98], dont la plupart sont des références de la littérature :

- D_1 composé de 4 classes de 50 points distribués selon des lois normales dont 2 se chevauchent légèrement, D_2 constitué des points de D_1 plus 100 points d'une distribution uniforme rendant le chevauchement plus important, $X30$ [BEZ 98], $Bensaid$ [BEN 96] composé de 3 classes de volume et cardinalité très différents,
- *Starfield* contenant la position et l'intensité lumineuse de 51 étoiles situées près de Solaris (8 à 9 classes selon les articles), *Iris* composé de 3 classes (dont 2 se chevauchent) de 50 fleurs décrites par 4 attributs physiques, *Wine* consistant en 13 attributs chimiques mesurés sur 178 vins italiens, *Breast* constitué de 699 cellules cancéreuses ou saines décrites par 9 attributs obtenus par imagerie numérique.

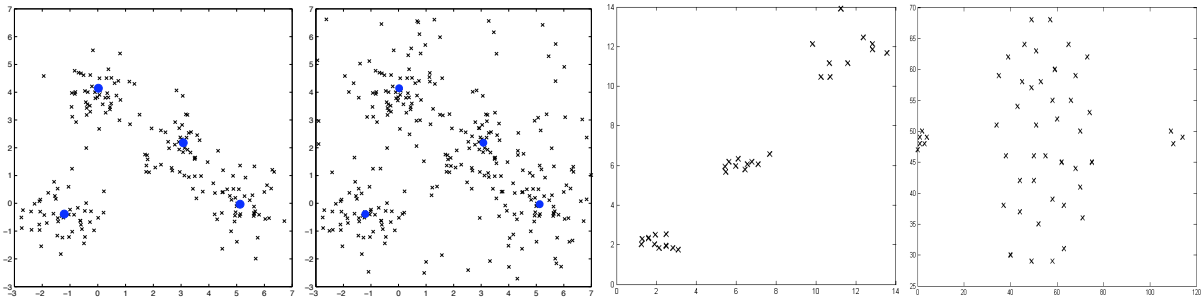


FIG. 1. *Jeux de données artificielles : D_1, D_2, $X30$ [BEZ 98] et $Bensaid$ [BEN 96]*

FCM a été utilisé avec $m = 2$ et un paramètre de terminaison $\varepsilon = 10^{-3}$. Le nombre de classes optimal c_{opt} a été cherché dans l'intervalle [$c_{min} = 2$,c_{max}] avec $c_{max} = 10$ pour les jeux à cardinalité n élevée et un entier proche de $\sqrt{n}$ pour les autres comme il est d'usage. Les résultats sont donnés au tableau 2. Ils montrent que le nouvel indice $CSI_{\perp}$ n'est mis en défaut qu'une fois, sur D_1 avec les normes algébriques, archimédiennes ($a \top a \leq a$). Ceci est d'autant plus remarquable pour les jeux de données avec chevauchement (D_1, D_2) ou faisant débat (*Starfield*, *Iris*) pour lesquels seul FHV fait aussi bien mais échoue sur *Bensaid* car les classes sont de volumes différents.

Le même constat peut être fait pour PBM et d'une manière générale pour tous les CVIs combinant compacité C et séparation S, et/ou utilisant l'information de type (U, V) ou (U, V, X). Enfin, l'indice $CSI_{\perp}$ est bien plus pertinent que les autres CVIs n'utilisant que la matrice de partition floue U, à savoir NPC et NPE.

TAB. 2. *Résultats : jeux de données et caractéristiques (cardinalité n, dimension p, nombre $c^{\star}$ de classes attendu, nombre c_{max} de classes cherché), nombre de classes sélectionné par les critères testés*

Données	dimensions				NPC	NPE	XB	FS	FHV	SC	K	WY	PBM	$CSI_{\perp}$		
	n	p	$c^{\star}$	c_{max}										$\perp_S$	$\perp_A$	$\perp_{H_0}$
D_1	200	2	4	10	4	2	3	4	4	3	3	4	4	4	4	4
D_2	300	2	4	10	3	2	3	5	4	8	3	3	4	4	3	4
$X30$	30	2	3	5	3	3	3	4	3	5	3	3	3	3	3	3
$Bensaid$	49	2	3	7	3	2	3	7	6	5	3	6	6	3	3	3
$Starfield$	51	2	8 ou 9	10	2	2	6	7	9	8	3	3	4	8	8	8
$Iris$	150	4	2 ou 3	10	3	2	2	5	3	3	2	2	3	2	2	2
$Wine$	178	13	3	10	3	2	2	10	3	6	2	3	3	3	3	3
$Breast$	699	9	2	10	2	2	2	3	2	4	2	2	2	2	2	2

Dans un futur proche, nous proposerons une étude de l'influence du choix des normes sur lesquelles l'indice est fondé, qui peuvent conduire à une sur – ou sous – détermination du nombre de classes. Nous montrerons que ceci est lié aux propriétés mathématiques des normes triangulaires.

4. Bibliographie

[BEN 96] BENSAID A. M., HALL L. O., BEZDEK J. C., CLARKE L. P., SILBIGER M. L., ARRINGTON J. A., MURTAGH R. F., Validity-Guided (Re)Clustering with Applications to Image Segmentation, *IEEE Transactions on Fuzzy Systems*, vol. 4, n^o 2, 1996, p. 112-123.

[BEZ 81] BEZDEK J. C., *Pattern recognition with fuzzy objective function algorithm*, Plenum Press, 1981.

[BEZ 98] BEZDEK J., PAL N., Some New Indexes of Cluster Validity, *IEEE Transactions on Systems, Man and Cybernetics*, vol. 23, n^o 3, 1998, p. 301-315.

[BLA 98] BLAKE C., MERZ C., Repository of Machine-Learning databases, University of California at Irvine, 1998, http ://www.ics.uci.edu/mlearn/MLRepository.html.

[DUN 77] DUNN J. C., *Fuzzy Automata and Decision processes*, Chapitre Indices of partition fuzziness and the detection of clusters in large data sets, Elsevier, NY, 1977.

[FUK 89] FUKUYAMA Y., SUGENO M., A new method for choosing the number of clusters for the fuzzy c-means method, *Proc. 5th Fuzzy Systems Symposium*, 1989, p. 247-250.

[GAT 89] GATH I., GEVA A. B., Unsupervised Optimal Fuzzy Clustering, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, n^o 7, 1989, p. 773-780.

[KIM 04] KIM D.-W., LEE K., LEE D., On cluster validity index for estimation of the optimal number of fuzzy clusters, *Pattern Recognition*, vol. 37, n^o 10, 2004, p. 2009-2025.

[KLE 05] KLEMENT E., MESIAR R., *Logical, Algebraic, Analytic, and Probabilistic Aspects of Triangular Norms*, Elsevier, 2005.

[KWO 98] KWON S. H., Cluster validity index for fuzzy clustering, *Electronic Letters*, vol. 34, n^o 22, 1998, p. 2176-2177.

[MAS 08] MASCARILLA L., BERTHIER M., FRÉLICOT C., A K-order Fuzzy OR Operator for Pattern Classification with K-order Ambiguity Rejection, *Fuzzy Sets and Systems*, vol. 159, n^o 15, 2008, p. 2011-2029.

[PAK 04] PAKHIRA M. K., BANDYOPADHYAY S., MAULIK U., Validity index for crisp and fuzzy clusters, *Pattern Recognition*, vol. 37, n^o 3, 2004, p. 487-501.

[ROU 78] ROUBENS M., Pattern classification problems and fuzzy sets, *Fuzzy Sets and Systems*, vol. 1, n^o 4, 1978, p. 239-253.

[WU 05] WU K., YANG M., A cluster validity index for fuzzy clustering, *Pattern Recognition Letters*, vol. 26, n^o 9, 2005, p. 1275-1291.

[XIE 91] XIE X. L., BENI G., A Validity Measure for Fuzzy Clustering, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, n^o 8, 1991, p. 841-847.

Distance de compression et classification prétopologique

Vincent Levorato [*] ,Thanh Van Le[*] , Michel Lamure[**] , Marc Bui [*]

[*] *Laboratoire ERIC EPHE-Sorbonne, 41 rue Gay Lussac 75005 Paris*
vincent.levorato@ephe.sorbonne.fr, than-van.le@univ-lyon1.fr, marc.bui@ephe.sorbonne.fr

[**] *Laboratoire ERIC - Université Claude Bernard Lyon 1, 43 boulevard du 11 novembre 1918, 69622 Villeurbanne Cedex*
michel.lamure@univ-lyon1.fr

RÉSUMÉ. *Nous présentons dans cet article des algorithmes de classification prétopologique récemment développés, ainsi que l'introduction d'un nouvel algorithme original. Les algorithmes prétopologiques,principalement basés sur la fonction d'adhérence prétopologique et le concept de fermés permettent de classifier des données complexes non représentables dans un espace métrique. Après quelques rappels sur les notions de classification, nous proposons un algorithme exploitant la notion de distance basée sur la complexité de Kolmogorov dans un modèle prétopologique afin d'introduire un indice de similarité. Nous illustrerons celui-ci par une application permettant de saisir le degré de finesse que l'on peut obtenir avec une telle approche.*

MOTS-CLÉS : *Classification, Partitionnement, Prétopologie, Modélisation, Kolmogorov.*

1. Introduction

Le problème de la classification peut être posé sous forme d'une interrogation : étant donné un ensemble d'observations dont on connaît une ou plusieurs caractéristiques, comment peut-on les regrouper en un certain nombre de groupes de manière à ce que les groupes obtenus soient constitués d'observations semblables et que les groupes soient les plus différents possible entre eux ?
Il existe de multiples méthodes de partitionnement des données, comme le regroupement hiérarchique, ou l'algorithme des k-moyennes. Nous allons nous pencher sur quelques méthodes existantes, puis nous nous focaliserons sur les récentes recherches de classification prétopologiques développées par Thanh Van Le [LE 07].
Dans la suite de l'article, nous montrerons comment il est possible d'utiliser la complexité de Kolmogorov dans un modèle prétopologique afin d'introduire un indice de similarité.

2. Classification de données par partitionnement

La classification d'objets par partitionnement en différents groupes, ou si l'on se place dans un contexte mathématique, en sous-ensembles, dépend des caractéristiques que ces objets pourraient avoir en commun, et principalement en effectuant une mesure de la distance qui les sépare. Le partitionnement de données est une technique connue de l'analyse de données en statistique que l'on utilise dans divers domaines que ce soit en fouille de données ou en reconnaissance d'images par exemple. Habituellement, on scinde les méthodes de classification en deux classes [JAR 68] : les méthodes de classification *hiérarchiques* et *non-hiérarchiques*. Mais tout d'abord, il faut poser le problème de la mesure de la distance.

2.1. *Mesure de similarité et distance*

L'objectif de l'analyse de données par partitionnement est de « séparer »un ensemble E composé de n objets en m partitions $\{C_i | i \in I\}$ tel que : $\forall i \in I, \cup C_j = E$ et $\forall (i,j) \in (I \times I), i \neq j, C_i \cap C_j = \varnothing$

Les objets d'une même partition C_i ont une forte similitude entre eux, et peu de similitude avec les objets des autres partitions. Pour mesurer cette similarité ou dissimilarité entre objets, une notion de distance, selon la nature des données, est nécessaire. Si ces objets sont exprimés par des variables numériques tel que l'âge, le nombre de sujets, ..., des distances telles que la distance Euclidienne, la distance de Manhattan (city block), la distance de Chebyshev, la distance de Minkowski, ..., sont utilisées de préférence. Cependant, pour représenter de « simples » distances entre variables de même catégorie comme les couleurs, les familles d'animaux, le sexe, ..., on se tournera vers la distance de Jaccard ou de Hamming par exemple [KAU 90].

Pourtant, dans la réalité, un objet peut être caractérisé par plusieurs types de variables. Dans ce cas, il faut utiliser une distance capable de mesurer la ressemblance entre des objets caractérisés par des variables de différents types. Pour cela, il existe le coefficient de similarité de Gower qui a été inventé pour mesurer la proximité entre données par intervalles, de manière nominale ou binaire. Il y aussi quelques distances proposées par Ralambondrainy [RAL 95] ou par Wilson et Martinez [WIL 97] permettant de traiter des variables de type quantitatif et qualitatif.

La plupart de ces distances requiert un processus de pré-traitement qui calcule la proximité entre chaque objet selon l'indice de similarité ou dissimilarité employé pour chaque variable permettant la comparaison. D'une manière différente, la distance de compression normalisée (Normalized Compression Distance) entre valeurs de différentes natures peut être calculée en se basant sur la complexité de Kolmogorov [CIL 05] :

$$NCD(x,y) = \frac{C(xy) - min\{C(x), C(y)\}}{max\{C(x), C(y)\}}$$

où $C(x)$ représente la taille binaire de x compressé, $C(y)$ représente la taille binaire de y compressé, et $C(xy)$ représente la taille binaire de la concaténation de x et y, le tout compressé. C peut utiliser des algorithmes de compression divers sans pertes comme l'algorithme GZip de Ziv et Lempelon ou BZip de Burrows et Wheeler.

Ces distances sont employées dans des méthodes de classification afin de mesurer la ressemblance des objets d'un ensemble. Le choix de ces algorithmes dépend des types de données, mais aussi de l'utilisation que l'on veut en faire. Nous pouvons résumer deux types de méthodes de classification de données : les méthodes de partitionnement hiérarchique soit construisent des amas d'objets similaires (ascendante), soit divisent un ensemble d'objets en sous-ensembles (descendante) ; et les méthodes de partitionnement non-hiérarchiques qui assignent chaque objet au sous-ensemble dont le centre est le plus proche de l'objet en question.

3. Méthodes de classification prétopologiques

3.1. *Rappels des concepts prétopologiques*

La théorie de la prétopologie est un outil mathématique de modélisation du concept de proximité. Celle-ci permet des opérations sur les ensembles telles que l'adhérence ou la fermeture [BEL 93]. Une application d'adhérence $a(.)$ de $\mathcal{P}(E)$ dans $\mathcal{P}(E)$ est appelée *adhérence* ssi $\forall A \in \mathcal{P}(E) : a(\emptyset) = \emptyset$ et $A \subseteq a(A)$. On définit soit les couples (E, a) ou soit le triplet (E, i, a) comme étant des *espaces prétopologiques*. A partir de ces propriétés, on obtient des espaces prétopologiques plus ou moins complexes, du plus général jusqu'à l'espace topologique. Les espaces les plus intéressants appartiennent au type $\mathcal{V}$, ($\forall A \in \mathcal{P}(E), \forall B \in \mathcal{P}(E), A \subseteq B \Rightarrow a(A) \subseteq a(B)$) lesquels nous utiliserons dans la suite de l'article. Le processus de dilatation généré par l'adhérence s'arrête à un instant donné et n'évolue plus. Dans ce cas, on a $a^{k+1}(A) = a^k(A)$ avec $k \in \mathbb{N}$. On nomme A comme étant un sous ensemble *fermé*.

On appelle fermé élémentaire et on note F_x la fermeture d'un singleton $\{x\}$ de E. On note $\mathcal{F}_e$, l'ensemble des fermés élémentaires de E tel que : $\mathcal{F}e(E, a) = \{F_x, x \in E\}$.

On appelle fermé minimal de E, tout élément de $\mathcal{F}_e$, minimal au sens de l'inclusion. L'ensemble des fermés minimaux est noté : $\mathcal{F}_m$.

3.2. *Méthodes MCP et MCPR*

Ces algorithmes ont été développés par Michel Lamure et Thanh Van Le, et apparaissent dans la thèse de cette dernière [LE 07], dont une écriture en pseudo-code a été faite dans [LEV 08].

L'algorithme MCPR se base l'algorithme des *k-moyennes*. Il fournit la possibilité de classifier des objets dans k classes par les k centres choisis a priori et qui sont employés comme germes afin de déterminer les k classes. Cependant, une des limites de l'algorithme *k-moyennes* est que le nombre de classes doit être prédéterminé et le procédé de choix des centres initiaux doit être effectué d'une certaine manière.

La méthode de classification MCP se fait selon deux processus :

1. *Processus de structuration :* il ressemble à celui de MCPR. Il consiste à chercher les familles des fermés et à déterminer des noyaux initiaux pour tous les fermés minimaux.

2. *Processus de partitionnement :* pour chaque élément qui n'est pas encore classé, nous choisissons une classe dans laquelle cet élément sera assigné de telle façon que deux contraintes ci-dessous soient satisfaites à la fois : la distance entre cet élément et le noyau du groupe est la plus proche ; il y a des liens entre cet élément et tous les autres du groupe.

4. Application de MCP grâce à la complexité de Kolmogorov

Dans cette partie, nous allons utiliser la méthode de classification prétopologique (MCP) en se basant sur un indice de similarité, lequel sera calculé à partir de la description élémentaire des singletons de l'espace grâce la complexité de Kolmogorov.

Voici comment est construit le modèle sur lequel nous voulons appliquer MCP : nous avons un ensemble d'auteurs, ainsi que l'intitulé des articles que ceux-ci ont écrits. L'objectif du modèle est de retrouver les auteurs en relation en fonction du titre de l'article qu'ils ont écrits, donc du sujet qu'ils ont abordé, puis de les classer selon leur domaine. Chaque auteur correspond au moins un article, et pour chaque article correspond au moins un auteur. Pour connaître la "$distance$" entre chaque article, on calcule toutes les distances de compression normalisées pour chaque paire d'articles en prenant comme mot à compresser la correspondance binaire de leur titre (pré-traitement). Soient E et F deux ensembles non vides.

(1) On appelle multiapplication ou correspondance de E dans F toute application de E dans $\mathcal{P}(F)$, ensemble des parties de F. Si X est une multiapplication de E dans F on notera :$X \ : E \rightarrow \mathcal{P}(F)$ dans ce cas, pour tout $x \in E$, $X(x)$ est une partie de F. (2) Soit X une multiapplication de E dans F. Si $\forall x \in E$, $X(x) \neq \emptyset$, on dit que X est une multiapplication à valeurs non vide ; si $\forall x \in E$, $X(x)$ est constante on dit que X est une multiapplication constante. (3) Soit X une multiapplication de E dans F, et A une partie de E ; on appelle image de A par X la partie de F notée $X(A)$ définie par : $X(A) = \bigcup_{x \in A} X(x)$.

On définit l'ensemble E comme l'ensemble des auteurs et l'ensemble G comme l'ensemble des articles. Comment savoir si deux articles sont "$proches$" ou pas ? En utilisant la définition de la distance de compression normalisée, on construit un espace prétopologique avec des relations valuées (à la manière d'un graphe complet), ces relations représentant la distance entre deux articles (calculée par la distance de compression normalisée appliquée au titre de l'article). Ainsi, les articles qui contiennent des mots similaires auront une distance qui les séparent plus faible que des articles qui ne possèdent aucun terme en commun, même s'il existe un risque de biais. Voici comment on le formalise :

$$Soit \ \ G \times G \rightarrow \mathbb{R}_+^*, (x, y) \rightarrow v(x, y)$$

$$\forall B \in \mathcal{P}(G), s \in \mathbb{R}_+^*, a_G(B) = \{y \in G - B, \sum_{x \in B} v(x, y) \leq s\} \cup B$$

L'application adhérence $a_G(.)$ nous permet d'exprimer la proximité entre articles. Le couple (G, a_g) correspond donc à l'espace prétopologique des articles. Nous allons nous servir de cette adhérence pour construire celle de l'espace prétopologique des auteurs. Définissons tout d'abord les deux multiapplications nous permettant de passer de E à G et de G à E :

Soit f une multiapplication de E dans G, A une partie de E, on définit l'image de A par la multiapplication f la partie de G notée $f(A)$ par $f(A) = \{y \in G; x \in E, y = f(x)\}$. Soit g une multiapplication G dans E, B une partie de G, on définit l'image de B par la multiapplication g la partie de E notée $g(B)$ par $g(B) = \{y \in E; x \in G, y = g(x)\}$.

En somme, f nous permet de déterminer quels articles ont été écrits par un ensemble d'auteurs et g quels auteurs ont écrits un ensemble d'articles. Nous pouvons à présent construire l'adhérence sur E :

$$\forall A \in \mathcal{P}(E), a(A) = \{x \in E, g(a_G(f(A)))\}$$

Le couple (E, a) correspond à l'espace prétopologique qui nous intéresse, c'est-à-dire celui portant sur les auteurs. On peut remarquer que celui-ci est de type $\mathcal{V}$.

On peut constater par cet exemple de modélisation que la prétopologie peut aisément concevoir des représentations de réseaux complexes. Les méthodes de classifications vues plus haut s'intègrent parfaitement dans ce type de modèle, étant basé sur l'adhérence.

5. Conclusion

Dans cet article, on a pu constater que des méthodes de classification pouvaient être utilisées dans les espaces prétopologiques, et qu'elles pouvaient être appliquées sur tout type de données. Etant basées sur l'adhérence comme la plupart des algorithmes prétopologiques car définissant la proximité, on a montré qu'il était possible d'utiliser la définition de cette adhérence pour définir différentes sortes de similarité, permettant de donner des résultats de classification différents selon le modèle étudié, faisant de la prétopologie un outil complet pour construire des modèles au comportement complexe. La modélisation d'auteurs présentée comme un cas particulier peut être utilisée de manière générique pour d'autres types de problèmes. Une des perspectives envisageables serait une comparaison de l'aspect théorique à l'aspect qualitatif d'une telle modélisation, ouvrant une direction vers des recherches intéressantes.

6. Bibliographie

[BEL 93] BELMANDT Z., *Manuel de prétopologie et ses applications : Sciences humaines et sociales, réseaux, jeux, reconnaissance des formes, processus et modèles, classification, imagerie, mathématiques*, Hermes Sciences Publications, 1993.

[CIL 05] CILIBRASI R., VITANYI P., Clustering by compression, *IEEE Transactions on information theory*, vol. 51, n^o 4, 2005.

[JAR 68] JARDINE N., SIBSON R., The construction of hierarchic and non-hierarchic classifications, *The Computer Journal*, vol. 11, n^o 2, 1968.

[KAU 90] KAUFMAN L., ROUSSEEUW P. J., *Finding groups in data : An introduction to cluster analysis*, WILEY-Interscience, 1990.

[LE 07] LE T. V., Classification prétopologique des données. Application à l'analyse des trajectoires patients, PhD thesis, Université Lyon 1, 2007.

[LEV 08] LEVORATO V., Contributions à la Modélisation des Réseaux Complexes : Prétopologie et Applications, PhD thesis, Université de Paris 8, 2008.

[RAL 95] RALAMBONDRAINY H., A conceptual version of the k-means algorithm, *Pattern Recognition Letters*, vol. 16, Elsevier Science, 1995, p. 1147-1157.

[WIL 97] WILSON D. R., MARTINEZ T. R., Improved heterogenous distance function, *Journal of Artificial Intelligence Researche*, vol. 6, 1997.

Les composantes connexes associées aux graphes des plus proches voisins

Maurice ROUX

IMEP(Case 462),
Université Paul Cézanne (Aix-Marseille 3),
Faculté des Sciences de Saint-Jérome,
13397 Marseille Cedex 20, France.

RÉSUMÉ : Cet exposé passe en revue quelques graphes couramment utilisés en analyse des données, notamment le classique arbre de longueur minimale. Toutefois il faut souligner que la plupart de ces graphes sont utilisés comme auxiliaires ou comme étapes préliminaires à d'autres méthodes de classification. On s'intéresse ici plus particulièrement aux graphes basés sur les K plus proches voisins de chaque individu de l'ensemble E étudié. Ces graphes ne sont pas nécessairement connexes et suggèrent ainsi une série de partitions liées au paramètre K, dont le choix s'avère crucial. On montre facilement que ces partitions sont emboîtées et constituent donc des hiérarchies de parties sur E. L'objet de ce travail est de montrer que ces hiérarchies peuvent être obtenues par une construction agglomérative usuelle basée sur une distance "non paramétrique" dérivée de la distance initiale entre les individus. Les graphes basés sur les plus proches voisins apparaissent ainsi comme des méthodes de classification pleines et entières.

MOTS-CLÉS : arbre de longueur minimale, graphe des K plus proches voisins, graphes des K voisins mutuels.

1. Introduction

Les graphes basés sur la liste des plus proches voisins de chaque objet ont intéressé les chercheurs depuis longtemps, par exemple pour donner quelques bases théoriques aux méthodes de classification ([GOW 78], [WON 83]). Plus récemment ces listes de plus proches voisins ont été utilisées pour déterminer les points de densité élevée ([TRA 03], [GUE 04]) ou pour détecter les individus extrêmes (« outliers », [HAU 93]). Toutefois les méthodes envisagées ne proposent que des règles empiriques pour déterminer le nombre K de voisins à prendre en compte.

Notre objectif dans le présent papier est double. Premièrement nous voulons passer en revue quelques graphes classiques pour représenter des dissimilarités entre objets étudiés et leurs relations d'inclusion. Deuxièmement nous voulons montrer qu'en construisant des dissimilarités basées sur les rangs des dissimilarités initiales, on peut obtenir les composantes connexes de certains graphes, basés sur les listes de plus proches voisins, comme les classes associées aux niveaux hiérarchiques d'un algorithme d'agrégations successives.

Ce travail a été motivé par la constatation que, dans les jeux de données réelles, on a souvent des différences importantes dans les densités des groupes d'objets présents dans l'échantillon étudié. D'où l'intérêt de certains chercheurs pour des « distances contextuelles » dérivées des distances interindividuelles initiales ([ZHA 07]), prenant en compte les voisinages des objets étudiés.

Dans le prochain paragraphe (no 2) nous rappelons les définitions et propriétés de certains graphes classiques en analyse des données. Dans le paragraphe suivant (no 3) on définit deux dissimilarités basées sur les rangs des voisins des objets étudiés. Puis on applique une construction ascendante hiérarchique à chacune de ces deux

distances ; on montre ensuite leurs relations avec les composantes connexes des graphes basés sur les plus proches voisins.

2. Quelques graphes utilisés couramment en analyse des données

Dans ce paragraphe on suppose que l'ensemble E, d'effectif n, des objets étudiés est fixé et constitue l'ensemble des sommets des divers graphes examinés. Les arêtes des graphes sont munies de poids égaux aux distances ou dissimilarités d données entre les objets.

Après avoir rappelé les définitions de l'arbre de longueur minimale (ALM, [KRU 56]), du graphe de Gabriel (GG, [GAB 69]) et du graphe des voisins relatifs (GVR, [TOU 80]) on souligne que ces 3 graphes sont connexes et inclus les uns dans les autres comme suit :

$$ALM \subseteq GVR \subseteq GG$$

2.1. Les graphes des K plus proches voisins (K-PPV)

Une autre série de graphes est composée des graphes des K plus proches voisins, où K doit être inférieur ou égal à n – 1. Une arête (i, j) d'un tel graphe est telle que, soit i appartient à la liste des K plus proches voisins de j, soit j apparaît dans la liste des K plus proches voisins de i (soit les deux). On a évidemment les inclusions suivantes :

$$1\text{-PPV} \subseteq 2\text{-PPV} \subseteq ... \subseteq \text{K-PPV} \subseteq ... \subseteq (n-1)\text{-PPV}$$

Certains de ces graphes ne sont pas connexes (faibles valeurs de K) tandis que d'autres le sont (grandes valeurs de K).

2.2 Les graphes des K voisins mutuels (K-VM)

On définit d'une façon analogue une autre série de graphes dite des K voisins mutuels (ou des K voisins réciproques), où K doit être, comme ci-dessus, inférieur ou égal à n – 1. Dans ces graphes une arête (i, j) est telle que l'on a simultanément i dans les K plus proches voisins de j ET j dans la liste des plus proches voisins de i. Comme précédemment on a :

$$1\text{-VM} \subseteq 2\text{-VM} \subseteq ... \subseteq \text{k-VM} \subseteq ... \subseteq (n-1)\text{-VM}$$

Et certains de ces graphes ne sont pas connexes (faibles valeurs de K) tandis que d'autres le sont (fortes valeurs de K). De plus, les conditions relatives aux arêtes étant plus restrictives pour les K-VM que pour les K-PPV, on a les inclusions suivantes :

$$1\text{-VM} \subseteq 1\text{-PPV} \ ; \ 2\text{-VM} \subseteq 2\text{-PPV} \ ; \ \ldots \ ; \ \text{K-VM} \subseteq \text{K-PPV} \ ; \ \ldots \ ; \ (n-1)\text{-VM} = (n-1)\text{-PPV}$$

La dernière inclusion est une égalité car les graphes des (n – 1)-VM et des (n – 1)-PPV ne sont autres que le graphe complet basé sur l'ensemble E, où tous les points sont reliés deux à deux par une arête.

3. Dissimilarités de rangs et constructions ascendantes hiérarchiques

3.1. Une nouvelle dissimilarité basée sur les rangs de voisinage

A partir de la dissimilarité initiale d(i, j) entre les objets i et j on construit deux nouvelles dissimilarités, Dmin et Dmax, basées sur les rangs des voisins de i et sur les rangs des voisins de j. On suppose évidemment que les voisins des objets sont rangés par ordre des dissimilarités d croissantes. On note Rang(i | j) le rang de i dans la liste des voisins de j et Rang(j | i) le rang de j dans la liste des voisins de i.

$$Dmin(i, j) = Min\{ Rang(i \mid j) ; Rang(j \mid i) \}$$

$$Dmax(i, j) = Max\{ Rang(i \mid j) ; Rang(j \mid i) \}$$

Par convention on suppose que Dmin (i, i) = 0 et Dmax(i, i) = 0.

Ces dissimilarités ont les propriétés de réflexivité et de symétrie mais ne satisfont pas nécessairement à l'inégalité triangulaire. Il faut noter qu'en général les valeurs de ces dissimilarités sont des nombres entiers situés dans l'intervalle [0, n – 1]. Il y a cependant un problème lorsque certaines dissimilarités d initiales sont égales. Ce cas ne se présente pas souvent en pratique, mais s'il se présente il est clair qu'il faut affecter les deux valeurs égales du même rang.

3.2. Application de la construction hiérarchique du lien simple

Remarquons, tout d'abord, que si l'on a Dmin(i, j) = 1 cela signifie sur soit i est le plus proche voisins de j, soit j est le plus proche voisin de i, ce qui n'exclut pas le cas où i et j sont plus proches voisins l'un de l'autre (voisins mutuels). De même si Dmin(i, j) = 2, cela signifie que soit i est dans la liste des 2 plus proches voisins de j, soit j est dans la liste des 2 plus proches voisins de i (soit les 2 simultanément). Et ainsi de suite avec Dmin(i, j) = 3, etc …

Les premières étapes de la méthode ascendante du lien simple consisteront à agréger toutes les paires d'objets pour lesquels Dmin = 1. Ce qui veut dire que les classes formées au niveau 1 sont les composantes connexes du graphe 1-PPV. De façon analogue, les étapes suivantes agrégeront tous les groupes dont la distance minimum sera réalisée par une paire (i, j) telle que Dmin(i, j) = 2. On formera ainsi les composantes connexes du graphe 2-PPV. Et ainsi de suite jusqu'à ce que l'on n'ait plus qu'une seule composante connexe associée au sommet de la hiérarchie ainsi construite.

Cette construction appelle quelques remarques. Tout d'abord l'arbre hiérarchique construit par le saut minimum est unique ; en effet il peut arriver que deux paires de classes, ayant un élément commun, disons la paire (A, B) et la paire (B, C) soient à agréger au même niveau. Si on choisit de fusionner la paire (A, B) en premier alors le saut minimum entre C et la nouvelle classe $A \cup B$ est toujours égal au saut minimum entre B et C. De sorte que la classe C doit être agrégée au groupe $A \cup B$ avec la même distance qu'entre A et B. Du point de vue graphique cela se traduit par un nœud ternaire à ce niveau, donc une seule nouvelle classe égale à la réunion $A \cup B \cup C$.

D'autre part il peut arriver que certaines valeurs entières de l'intervalle [0, n – 1] ne correspondent à aucune valeur de Dmin, dans ce cas certains niveaux entiers de la hiérarchie n'existent pas. En général tous les objets sont réunis en une seule classe à un niveau inférieur à n – 1. Enfin on peut tenir le même raisonnement avec la dissimilarité Dmax, mais dans ce cas les classes de la hiérarchie correspondent aux composantes connexes des graphes K-VM.

3.3. Que se passe-t-il avec l'agglomération par le lien complet ?

Considérons la dissimilarité Dmin et la série des graphes K-PPV associée. Supposons que A et B soient deux classes devant être fusionnées au niveau K de la méthode agrégative du lien complet. Cela veut dire que la plus grande distance entre un élément de A et un élément de B est exactement égale à K. Donc toutes les distances au sein de la réunion $A \cup B$ sont inférieures ou égales à K, ce qui se traduit en disant que $A \cup B$ est une clique du graphe K-PPV. De plus c'est une clique maximale, car, si ce n'était pas le cas elle serait incluse strictement dans une clique de diamètre K, elle aurait donc un diamètre strictement inférieur à K ce qui contredit la supposition initiale.

Le même raisonnement s'applique au cas de la dissimilarité Dmax et des graphes K-VM. Il faut cependant noter que les cliques maximales sont en général fortement empiétantes les unes sur les autres, et que seule une partie d'entre elles sont découvertes par l'algorithme du lien complet. De plus, compte tenu des nombreuses valeurs

identiques dans les dissimilarités Dmin et Dmax, suivant l'ordre choisi pour agréger les paires de classes, on est conduit à des arbres hiérarchiques sensiblement différents. Ce n'est pas le cas pour l'algorithme du lien simple pour lequel l'ordre des agrégations ne change pas la forme de l'arbre hiérarchique final.

3. Conclusion

L'idée d'utiliser des dissimilarités basées sur les rangs des listes de voisins n'est pas nouvelle. On peut qualifier ces dissimilarités de « non-paramétriques » par analogie avec les tests statistiques basés sur les rangs. Elles ont l'avantage d'être beaucoup moins sensibles aux variations des distances au sein des classes d'objets que l'on veut mettre en évidence. En effet il est fréquent que pour un échantillon donné certaines classes soient très compactes et d'autres soient moins denses. Au sein des premières les distances seront faibles et elles seront plus fortes au sein des secondes. Ces différences sont relativement effacées par passage aux rangs des listes de voisins qui permettent de prendre en compte la densité locale des objets.

Dans notre méthode on évite la difficulté du choix du nombre K de voisins à prendre en compte. En effet la construction hiérarchique, basée sur l'agglomération par le lien simple, permet de visualiser simultanément les composantes connexes associées à toutes les valeurs de K, aussi bien pour les graphes des plus proches voisins que pour les graphes des voisins mutuels. La construction hiérarchique du lien complet implique une partie des cliques maximales de ces mêmes graphes mais l'interprétation de l'arbre hiérarchique final est plus délicate compte tenu de la non-unicité de cet arbre.

5. Bibliographie

[GAB 69] GABRIEL K.R., SOKAL R.R. "A new statistical approach to geographic variation analysis". *System. Zool.* 18, 1969, pp 259-278.

[GUE 04] GUÉNOCHE A. "Clustering by vertex density in a graph". Meeting of the International Federation of the Classification Societies, IFCS'2004, in *"Classification, Clustering and Data Mining"*, D. Banks et al. (Eds.), Springer, 2004, pp 15-23.

[GOW78] GOWDA K.C. AND KRISHNA G. (1978). "Agglomerative clustering using the concept of mutual nearest neighbourhood". Pattern Recognition, 10, 1978, pp 105-112.

[KRU 56] KRUSKAL J.B. "On the shortest spanning subtree of a graph and the traveling salesman problem". *Proc. Amer. Math. Soc.* 7, 1956, pp 48-50.

[TOU 80] TOUSSAINT G.T. "The relative neighborhood graph of a planar set". *Pattern recognition* 12, 1980, pp 261-268.

[TRA 03] TRAN T.N., WEHRENS W., BUYDENS L.M.C. "K-nearest neighbour density-based clustering for high dimensional multispectral images". Proc. 2nd GRSS/ISPRS, *Joint Workshop on Remote Sensing and Data Fusion over Urban Areas,* URBAN 2003, May 2003, Berlin, Germany.

[WON 83] WONG M.A. AND LANE T. "A k-th nearest neighbour clustering procedure". J.R. Statist. Soc. Ser. B, 45(3), 1983, pp 362-368.

[ZHA 07] ZHAO D-L., LIN Z., TANG X., "Contextual Distance for Data Perception,"IEEE 11th International Conference on Computer Vision, 2007, ICCV, pp.1-8,

Une méthode de partitionnement pour la classification de variables qualitatives

Marie Chavent[1,2], **Vanessa Kuentz**[1,2], **Jérôme Saracco**[1,2,3]

[1] *Université de Bordeaux, IMB, 351 cours de la Libération 33405 Talence Cedex*
{chavent,kuentz}@math.u-bordeaux1.fr
[2] *INRIA Bordeaux Sud-Ouest, Equipe CQFD*
[3] *Université Montesquieu Bordeaux IV, GREThA, Avenue Léon Duguit 33608 Pessac Cedex*
jerome.saracco@u-bordeaux4.fr

RÉSUMÉ. *Les méthodes de classification de variables regroupent ensemble les variables qui sont fortement liées et qui par conséquent apportent la même information. Elles sont utiles pour la réduction de dimension, la sélection de variables ou dans certaines études de cas. Dans cet article nous proposons un algorithme des nuées dynamiques pour la classification de variables qualitatives. La proximité entre les variables qualitatives d'une classe et leur prototype ou variable latente est mesurée à l'aide du rapport de corrélation. L'originalité de l'approche tient dans la définition du prototype qui est une variable quantitative. Le bon comportement de l'approche proposée est montrée à l'aide d'un jeu de données simulées. Le potentiel de la méthode est également décrit à l'aide d'une application sur des données réelles.*

MOTS-CLÉS : *Algorithme des nuées dynamiques, variable latente, rapport de corrélation.*

1. Introduction

Les techniques de classification de variables permettent de regrouper des variables fortement liées qui apportent la même information. Il est ensuite possible d'extraire dans chaque classe une variables synthétique ou latente qui sera plus facile à utiliser lors d'analyses ultérieures. Ainsi elles peuvent être une solution pour la dimension de réduction et la sélection de variables, problèmes courants avec l'émergence de bases de données de plus en plus grandes. Dans certaines études de cas, l'objectif est de classifier des variables et non des individus, comme par exemple en analyse sensorielle pour l'identification de groupes de descripteurs.

L'objectif est de trouver une partition des variables en classes homogènes. Il existe pour cela de nombreuses mesures de ressemblance entre variables quantitatives ou qualitatives. Une première approche simple est de calculer une matrice de dissimilarités entre ces variables et d'appliquer l'une des méthodes classiques de classification sur matrice de dissimilarités. Si l'objectif poursuivi est la suppression des redondances entre les variables par la réduction de la dimension du tableau de données, des variables synthétiques doivent être calculées dans une seconde étape pour chaque classe. Une approche naturelle dans ce cas est donc d'utiliser un algorithme de classification qui fournit simultanément des classes et des prototypes de ces classes qui sont des variables synthétiques de ces classes. C'est le cas de l'algorithme des nuées dynamiques. Dans cet esprit, deux méthodes de partitionnement ont été proposées simultanément (voir [VIG 03] et [DHI 03]) pour la classification de variables quantitatives. Lorsqu'on souhaite regrouper des variables fortement corrélées sans tenir compte du signe de la corrélation, un critère de partitionnement basé sur des corrélations au carré est maximisé. Le prototype d'une classe est défini en optimisant le critère d'homogénéité de la classe. Il s'agit de la première composante principale issue de l'Analyse en Composantes Principales de la matrice dont les colonnes sont formées par les variables de la classe considérée.

Dans cet article nous étendons cette méthode de partitionnement au cas de variables qualitatives. Le critère d'homogénéité d'une classe est mesuré à l'aide de rapports de corrélation entre les variables qualitatives de la classe et leur représentant. Nous donnons la définition de la variable quantitative qui maximise le critère d'homogénéité et qui est utilisé comme prototype dans l'algorithme des nuées dynamiques proposé.

2. L'algorithme des nuées dynamiques pour la classification de variables qualitatives

Notations. Nous considérons une matrice $\mathbf{X}$ de données qualitatives où $x_{ij} \in \mathcal{M}_j$ avec $\mathcal{M}_j$ l'ensemble des modalités de x_j. Soit $\mathbf{O}_k = (o_{is})_{n \times q_k}$ la matrice des indicatrices des modalités des p_ℓ variables de C_k. Cette matrice possède n lignes et q_k colonnes, avec q_k le nombre total de modalités des variables de C_k. Comme $\mathbf{O}_k$ peut être considérée comme une sorte de table de contingence, on peut construire la matrice des fréquences $\mathbf{F}_k = (f_{is})_{n \times q_k}$ où $f_{is} = \frac{o_{is}}{np_k}$ car $\sum_{i,s} o_{is} = np_k$. On définit $\mathcal{G}_s = \{e_i \in \mathcal{E} | o_{is} = 1\}$ et $n_s = \text{card}(\mathcal{G}_s)$. Les sommes marginales de $\mathbf{F}_k$ sont utilisées pour définir les poids des lignes et des colonnes : le vecteur $\mathbf{r} = (f_{1.}, ..., f_{i.}, ..., f_{n.})^t$ avec $f_{i.} = \frac{1}{n}$ donne le poids des lignes et le vecteur $\mathbf{c} = (f_{.s}, s \in \cup_{x_j \in C_k} \mathcal{M}_j, s \text{ in increasing order})^t$ avec $f_{.s} = \frac{n_s}{np_k}$donne le poids des colonnes. Soient $\mathbf{D}_n = \text{diag}(\mathbf{r})$, $\mathbf{D}_k = \text{diag}(\mathbf{c})$ et $\widetilde{\mathbf{F}}_k$ la matrice $\mathbf{F}_k$ dont les éléments sont divisés par $\sqrt{f_{i.} f_{.s}}$:

$$\widetilde{\mathbf{F}}_k = \mathbf{D}_n^{-1/2} \mathbf{F}_k \mathbf{D}_k^{-1/2}. \tag{1}$$

L'algorithme des nuées dynamiques. Soit $\mathbf{P} = (C_1, ..., C_K)$ une partition de $\mathcal{V}$ en K classes et soit $\mathbf{Y} = (y_1, \ldots, y_k, \ldots, y_K)$ un ensemble de K variables quantitatives appelées variables latentes ou prototypes. Chaque variable latente y_k est décrite sur les n objets de $\mathcal{E}$ par un vecteur $\mathbf{y}_k \in \mathbb{R}^n$. L'ojectif est de trouver un couple $(\mathbf{P}, \mathbf{Y})$, qui optimise le critère de partitionnement suivant :

$$g(\mathbf{P}, \mathbf{Y}) = \sum_{k=1}^{K} \sum_{x_j \in C_k} \eta^2(\mathbf{x}_j, \mathbf{y}_k). \tag{2}$$

où $\eta^2(\mathbf{x}_j, \mathbf{y}_k)$ est le rapport de corrélation entre une variable qualitative $\mathbf{x}_j$ et la variable latente quantitative y_k de la classe C_k.

L'algorithme des nuées dynamiques comprend les étapes suivantes :

(a) *Initialisation* : On définit une partition initiale $\{C_1, \ldots, C_K\}$ de $\mathcal{V}$.

(b) *Représentation* : Pour tout $k = 1, ..., K$, on définit le prototype de C_k comme $\mathbf{y}_k^* = \arg \max_{\mathbf{y}_k \in \mathbb{R}^n} f(C_k, \mathbf{y}_k)$.
Cette définition est donnée dans la partie suivant.

(c) *Affectation* : Pour tout $j = 1, ..., p$, on cherche k tel que $\ell = \arg \max_{k=1,...,K} \sum_{x_j \in C_k} \eta^2(\mathbf{x}_j, \mathbf{y}_k)$.

(d) *Arrêt* : Si aucun changement dans (c) alors *stop*, sinon on retourne à (b)

Définition du prototype. Définir le prototype y_k de la classe C_k revient à résoudre le problème d'optimisation suivant :

$$\max_{\mathbf{y} \in \mathbb{R}^n} \sum_{x_j \in C_k} \eta^2(\mathbf{x}_j, \mathbf{y}). \tag{3}$$

On peut montrer que le premier vecteur propre normé de $\widetilde{\mathbf{F}}_k \widetilde{\mathbf{F}}_k^t$ est solution de (3), où $\widetilde{\mathbf{F}}_k$ est défini dans (1). Il s'en suit que la première composante principale de l'AFCM (sur la matrice des profils lignes) est solution de (3). Ces deux vecteurs peuvent donc être choisis comme prototype de la classe C_k. Le détail de ces résultats est disponible dans [CHA 08].

3. Applications

Dans cette section, nous présentons des applications de la méthode proposée pour le partitionnement de variables qualitatives. Dans un premier temps, la performance numérique de l'approche est montrée sur un jeu de données simulées. Puis une application sur des données réelles issues d'une enquête de satisfaction de plaisanciers montre le potentiel de la méthode.

4. Bibliographie

[CHA 08] Chavent, M., Kuentz, V., Saracco, J., (2008), A center-based method for the clustering of qualitative variables, *Submitted paper*.

[DHI 03] Dhillon, I.S, Marcotte, E.M., Roshan, U., (2003), Diametrical clustering for identifying anti-correlated gene clusters, *Bioinformatics*, **19**(13), 1612-1619.

[VIG 03] Vigneau, E., Qannari, E.M., (2003), Clustering of variables around latent components, *Communications in statistics Simulation and Computation*, **32**(4), 1131-1150.

Classification hiérarchique de données ordinales

F.-X. Jollois, M. Nadif

LIPADE – UFR Math-Info
Université Paris Descartes
45 rue des Saints-Pères
75006 PARIS
{francois-xavier.jollois,mohamed.nadif}@parisdescartes.fr

RÉSUMÉ. *Dans le cadre d'une classification d'un ensemble d'individus ou d'objets, les variables ordinales sont régulièrement considérées soit comme continues, soit comme nominales. Dans les deux cas, ce choix implique un certain nombre d'inconvénients. Dans ce travail, nous traitons la classification hiérarchique des données ordinales sous l'approche modèle de mélange. Pour ce faire, nous utilisons un modèle de mélange multinomial contraint, respectant ainsi le caractère ordinal des modalités de réponse.*

MOTS-CLÉS : *Classification Ascendante Hiérarchique, Données ordinales, Modèle de mélange*

1. Introduction

Les données ordinales sont présentes dans plusieurs situations : en particulier, elles sont utiles pour représenter les degrés de satisfaction ou les préférences d'individus vis-à-vis de services ou de produits. Plusieurs travaux ont déjà été effectués sur la mise en place de modèles probabilistes ou d'outils statistiques pour décrire les processus de classement ou de notation [FLI 93, MAR 95]. Lorsque l'objectif est de classifier un ensemble d'objets, l'utilisation des modèles de mélange est devenue une approche classique et puissante [BAN 93, CEL 95].

L'utilisation des modèles de mélanges pour la classification de données ordinales a déjà été étudiée [D'E 05, GOU 06]. Les premiers considèrent, dans un contexte de comparaisons appariés, que la note donnée par un juge à un item est une réalisation d'une variable binomiale *décalée*. Gouget [GOU 06] compare différentes approches à partir de modèles multinomiaux contraints : linéaire, polynomial et euclidien. Nous nous intéressons ici au modèle polynomial.

Dans ce travail, nous présentons l'approche par modèle de mélange dans la section 2, et l'algorithme hiérarchique dans la section 3. Dans la section 4, nous présentons le modèle polynomial adapté aux données ordinales. Enfin, dans la section 5, nous concluons et présentons les travaux en cours que nous développerons lors de notre présentation.

2. Modèle de mélange

Dans l'approche modèle de mélange, les objets $\mathbf{x}_1, \ldots, \mathbf{x}_n$ (décrits par d variables ordinales) à classifier sont supposés provenir d'un mélange de s densités dans des proportions inconnus $p_1, \ldots, p_s$. Chaque objet $\mathbf{x}_i$ est ainsi une réalisation d'une densité de probabilité (p.d.f.), décrite par

$$\varphi(\mathbf{x}_i; \boldsymbol{\theta}) = \sum_{k=1}^{s} p_k \varphi_k(\mathbf{x}_i; \boldsymbol{\alpha}_k)$$

où $\varphi_k(\mathbf{x}_i;\boldsymbol{\alpha}_k)$ représente la densité de $\mathbf{x}_i$ de paramètre $\boldsymbol{\alpha}_k$. Le vecteur des paramètres à estimer $\boldsymbol{\theta}$ est composé de $\mathbf{p} = (p_1,\ldots,p_k)$ et $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1,\ldots,\boldsymbol{\alpha}_k)$. La log-vraisemblance des données observées $\mathbf{x} = (\mathbf{x}_i,\ldots,\mathbf{x}_n)$ est donnée par

$$L(\mathbf{x},\boldsymbol{\theta}) = \sum_{i=1}^{n} \log\left(\sum_{k=1}^{s} p_k \varphi_k(\mathbf{x}_i;\boldsymbol{\alpha}_k)\right).$$

Sous l'approche classification, la recherche d'une partition $\mathbf{z} = (z_1,\ldots,z_s)$ peut être effectuée par la maximisation de la log-vraisemblance classifiante, déterminée par

$$\begin{aligned} L_C(\mathbf{x},\mathbf{z},\boldsymbol{\theta}) &= \sum_{k=1}^{s}\sum_{i\in z_k} \log\left(p_k\varphi_k(\mathbf{x}_i;\boldsymbol{\alpha}_k)\right) \\ &= \sum_{k=1}^{s} \#z_k \log(p_k) + \sum_{k=1}^{s}\sum_{i\in z_k} \log\left(\varphi_k(\mathbf{x}_i;\boldsymbol{\alpha}_k)\right). \end{aligned}$$

Si nous supposons les proportions égales, le premier terme devient constant. La maximisation de la log-vraisemblance classifiante restreinte revient donc à maximiser la quantité suivante

$$\begin{aligned} L_C(\mathbf{x},\mathbf{z},\boldsymbol{\theta}) &= \sum_{k=1}^{s}\sum_{i\in z_k} \log\left(\varphi_k(\mathbf{x}_i;\boldsymbol{\alpha}_k)\right) \\ &= \sum_{k=1}^{s} L_{C_k}(\mathbf{x},z_k,\boldsymbol{\theta}_k). \end{aligned} \tag{1}$$

3. Classification Ascendante Hiérarchique (CAH)

La classification ascendante hiérarchique a comme principe de base le regroupement deux à deux des classes les plus proches, jusqu'à l'obtention d'une et une seule classe. Le point de départ initial est la définition de n classes singletons. Ainsi, à chaque étape, nous allons chercher à maximiser $L_C(\mathbf{x},\mathbf{z},\boldsymbol{\theta})$ en regroupant deux classes z_k et z'_k, pour obtenir la partition $\mathbf{z}^*$. Il faut noter que la log-vraisemblance classifiante restreinte décroît lorsque le nombre de classes décroît. Donc, le regroupement de deux classes conduira à une log-vraisemblance classifiante inférieure. Il suffit de minimiser la quantité

$$\delta(\mathbf{z},\mathbf{z}^*) = L_C(\mathbf{x},\mathbf{z},\boldsymbol{\theta}) - L_C(\mathbf{x},\mathbf{z}^*,\boldsymbol{\theta}),$$

avec $\mathbf{z}^*$ représentant la partition où z_k et z'_k sont regroupées en une seule classe. Cette quantité nous permet donc de déterminer une dissimilarité entre les deux classes z_k et z'_k, qui peut se définir à l'aide de la décomposition de $L_C(\mathbf{x},\mathbf{z},\boldsymbol{\theta})$ (1), et qui s'écrit

$$\Delta(z_k,z'_k) = L_{C_k}(\mathbf{x},z_k,\boldsymbol{\theta}_k) + L_{C_{k'}}(\mathbf{x},z_{k'},\boldsymbol{\theta}_{k'}) - L_{C_{kk'}}(\mathbf{x},z_{kk'},\boldsymbol{\theta}_{kk'})$$

où $z_{kk'}$ représente la classe obtenue par fusion des classes z_k et z'_k.

Dans la suite nous allons considéré l'approche mélange pour définir un critère d'agrégation Δ adapté à la nature des données ordinales.

4. Données ordinales

Lorsqu'on ne tient pas compte de l'ordre des modalités, on considère que les données sont constitués d'un échantillon $(\mathbf{x}_1,\ldots,\mathbf{x}_n)$, où $\mathbf{x}_i = (x_i^{je}; j = 1,\ldots,d; e = 1,\ldots,c_j)$, avec c_j le nombre de modalités de la variable

j et $x_i^{je} = 1$ si l'individu i prend la modalité e pour la variable j, et 0 sinon. On utilise alors le modèle des classes latentes [LAZ 68], dont le principe de base est la supposition d'une variable qualitative latente à c_j modalités dans les données. Dans ce modèle, les associations entre chaque paire de variables disparaissent, si la variable latente est constante. C'est le modèle basique de l'analyse des classes latentes, avec l'hypothèse d'indépendance locale. Cette hypothèse est couramment choisie quand les données sont de type qualitatif ou binaire [CEL 92, CHE 96]. Ainsi, la densité d'une observation $\mathbf{x}_i$ peut se décrire comme suit :

$$\varphi_k(\mathbf{x}_i; \boldsymbol{\alpha}_k) = \prod_{j=1}^{d} \prod_{e=1}^{c^j} \left(\alpha_k^{je}\right)^{x_i^{je}}, \text{ avec } \sum_{e=1}^{c^j} \alpha_k^{je} = 1$$

L'hypothèse d'indépendance locale permet d'estimer les paramètres séparément. Cette hypothèse simplifie grandement les calculs, principalement quand le nombre de variables est grand. Bien que cette affirmation est toujours fausse dans la pratique, l'indépendance locale est généralement très performante pour la classification. Ce modèle noté $\lfloor p_k, \alpha_k^{je} \rfloor$, conduit malheureusement à des classes dont les centres ne sont pas de même nature que les données initiales. En imposant une contrainte sur les probabilités associées aux modalités, on peut considérer le modèle noté $\lfloor p_k, \varepsilon_k^j \rfloor$ [NAD 93] conduisant à des classes formées par des modalités majoritaires notées a_k^j. Dans ce modèle, nous considérons que la probabilité α_k^{je} associée à une modalité e de j est égale à $1-\varepsilon_k^j$ si $e = a_k^j$ et égale à $\frac{\varepsilon_k^j}{c_j - 1}$ sinon. Malheureusement, ces deux modèles ne prennent pas en compte l'aspect ordinal des données. Pour ce faire, il est nécessaire d'imposer des contraintes d'ordre sur les probabilités des modalités : pour chaque classe k et chaque variable j, les probabilités α_k^{je} de la modalité e vont en décroissant à partir de la modalité *centrale* a_k^j (il est important de noter que ce n'est pas forcément la modalité majoritaire).

En premier lieu, le plus simple est de considérer cette décroissance constante entre deux modalités. Dans le cas de variables avec un nombre de modalités important, ce modèle s'apparente au modèle multinomial $\lfloor p_k, \varepsilon_k^j \rfloor$. Une solution pour éviter cet écueil est de considérer que la croissance n'est pas linéaire, mais polynomial de degré q. Ainsi, dans ce modèle [GOU 06], noté ici $\lfloor p_k, \alpha_k^{je}, q \rfloor$, la probabilité d'une modalité e pour la variable j dans la classe k peut s'exprimer par :

$$\alpha_k^{je} = \begin{cases} p_{a_k^j} & \text{si } e = a_k^j \\ \left(1 - p_{a_k^j}\right) \frac{(1+\max(a_k^j - 1; c_j - a_k^j) - |e - a_k^j|)^q}{\sum_{u \neq a_k^j} (1+\max(a_k^j - 1; c_j - a_k^j) - |u - a_k^j|)^q} & \text{sinon} \end{cases}$$

Notons que dans ce modèle, on peut introduire une paramétrisation supplémentaire en considérant que q est variable et dépend à la fois de la classe k et de la variable j, dans ce cas le modèle est noté $\lfloor p_k, \alpha_k^{je}, q_k^j \rfloor$. Nous nous concentrons, dans ce travail, sur le modèle simple $\lfloor p_k, \alpha_k^{je}, q \rfloor$.

Il est de plus nécessaire de s'assurer que la probabilité associée à la modalité a_k^j choisie soit supérieure à toutes les autres probabilités des modalités $e \neq a_k^j$. Ainsi, nous définissons le seuil $\beta_q^{c_j}$ comme étant la probabilité telle que, pour tout $e = 1, \ldots, c^j, e \neq a_k^j$, avec

$$\left(1 - \beta_q^{c_j}\right) \frac{(1 + \max(a_k^j - 1; c_j - a_k^j) - |e - a_k^j|)^q}{\sum_{u \neq a_k^j} (1 + \max(a_k^j - 1; c_j - a_k^j) - |u - a_k^j|)^q} \leq \beta_q^{c_j}$$

Ce seuil ne dépend que du nombre de modalités c_j et du paramètre q. Dans le cas où $p_{a_k^j}$ est inférieure à ce seuil $\beta_q^{c_j}$, nous affectons ce seuil à la probabilité $p_{a_k^j}$.

L'algorithme 1 décrit les deux étapes de la classification ascendante hiérarchique utilisant le critère d'agrégation issu du modèle décrit ci-dessus.

Algorithme 1 CAH pour données ordinales

Démarrer avec chaque objet dans sa propre classe
Calcul des dissimilarités entre les classes à l'aide $\Delta(z_k, z'_k), \forall k, k' = 1, \ldots, n$ tel que $k \neq k'$
Tant que Nombre de classes strictement supérieur à 1 **Faire**
  Agrégation des deux classes les plus proches z_k et z'_k en une nouvelle classe z^*_k
  Pour Chaque variable $j = 1, \ldots, d$ **Faire**
    Recherche des modalités principales $a^j_{k^*}$ (en respectant $p_{a^j_{k^*}} \geq p_e \forall e = 1, \ldots, c_j$)
  Fin Pour
  Calcul des nouvelles dissimilarités $\Delta(z_\ell, z^*_k)$ entre chaque classe z_ℓ (avec $\ell \neq k, k'$) et la nouvelle classe créée z^*_k
Fin Tant que

5. Conclusion

Nous abordons dans ce travail la classification hiérarchique de données ordinales sous l'approche modèle de mélange. Le critère d'agrégation utilisé est issu d'un modèle de mélange multinomial contraint, permettant de prendre en compte le caractère ordinal des données.

Des expériences à partir de données simulées seront présentées en tenant compte de plusieurs situations. Celles-ci dépendront du nombre de classes, du degré de mélange et du paramètre q. Nous comparerons les résultats de l'algorithme 1 et l'algorithme obtenu lorsqu'on considère le modèle multinomial sans contraintes (les variables étant considérées comme qualitatives nominales) et le modèle gaussien (les variables étant considérées comme quantitatives). Cette étude sera illustrée par une application sur des données réelles.

6. Bibliographie

[BAN 93] BANFIELD J. D., RAFTERY A. E., Model-based Gaussian and non-Gaussian Clustering, *Biometrics*, vol. 49, 1993, p. 803–821.

[CEL 92] CELEUX G., GOVAERT G., A Classification EM Algorithm for Clustering and Two Stochastic Versions, *Computational Statistics & Data Analysis*, vol. 14, 1992, p. 315–332.

[CEL 95] CELEUX G., GOVAERT G., Gaussian Parcimonious Clustering Methods, *Pattern Recognition*, vol. 28, 1995, p. 781–793.

[CHE 96] CHEESEMAN P., STUTZ J., Bayesian Classification (AutoClass) : Theory and Results, FAYYAD U., PIATETSKY-SHAPIRO G., UTHURUSAMY R., Eds., *Advances in Knowledge Discovery and Data Mining*, AAAI Press, 1996, p. 61–83.

[D'E 05] D'ELIA A., PICCOLO D., A mixture model for preferences data analysis, *Computational Statistics & Data Analysis*, vol. 49, 2005, p. 917–934.

[FLI 93] FLIGNER M., VERDUCCI J., *Probability models and statistical analysis of ranking data*, Springer, New-York, 1993.

[GOU 06] GOUGET C., Utilisation des modèles de mélange pour la classification automatique de données ordinales, PhD thesis, Université de Technologie de Compiègne, December 2006.

[LAZ 68] LAZARFELD P., HENRY N., *Latent Structure Analysis*, Houghton Mifflin, Boston, 1968.

[MAR 95] MARDEN J., *Analyzing and modeling rank data*, Chapman & Hall, London, 1995.

[NAD 93] NADIF M., MARCHETTI F., Classification de donnés qualitatives et modèles, *Revue de Statistique Appliquée*, vol. XLI, n^o 1, 1993, p. 55–69.

Structure des réseaux phylogénétiques de niveau borné

Philippe Gambette, Vincent Berry, Christophe Paul

Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier.
C.N.R.S., Université Montpellier 2.
161 rue Ada, 34392 Montpellier Cedex 5 France

RÉSUMÉ. *Les* réseaux phylogénétiques *généralisent les arbres phylogénétiques en représentant des échanges de matériel génétique entre espèces par des branches qui se rejoignent pour former des parties réticulées. Le* niveau *est un paramètre introduit sur les réseaux phylogénétiques enracinés pour décrire la complexité de leur structure par rapport à un arbre [JAN 04]. Des algorithmes polynomiaux ont récemment été proposés pour reconstruire un réseau de niveau borné compatible avec un ensemble de triplets fournis en entrée [IER 08, TO 09]. Nous étudions la structure d'un réseau de niveau borné pour montrer qu'il peut être décomposé en un arbre de générateurs choisis parmi un ensemble fini. Nous nous intéressons alors à la pertinence du paramètre de niveau dans le cadre d'un modèle d'évolution avec recombinaisons : le modèle coalescent.*

MOTS-CLÉS : *Combinatoire, Décomposition, Graphe, Réseau phylogénétique.*

1. Introduction et définitions

Un *arbre phylogénétique* est un arbre binaire enraciné avec des arcs (orientés, donc) et des feuilles étiquetées bijectivement par un ensemble X de *taxons*, qui représentent le plus souvent des espèces ou des gènes. Un *réseau phylogénétique explicite* est une généralisation d'arbre phylogénétique qui permet de prendre en compte les échanges de matériel génétique entre espèces, qui sont très fréquents entre les bactéries [DOO 99] mais aussi présents chez les végétaux ou même les animaux [HUB 55]. Ces échanges peuvent correspondre à divers événements biologiques : hybridation, recombinaison, transferts horizontaux…

On peut définir formellement un réseau phylogénétique explicite comme un multigraphe orienté acyclique, contenant : exactement un sommet a degré entrant 0 et degré sortant 2 (la *racine*) ; des sommets de degré entrant 1 et de degré sortant 2 (*sommets de spéciation*) ; des sommets de degré entrant 2 et de degré sortant au plus 1 (*sommets hybrides*) ; des sommets étiquetés bijectivement par un ensemble X de taxons, de degré entrant 1 et de degré sortant 0 (*feuilles*). Dans la Figure 2(a) est représenté un réseau phylogénétique explicite N de racine ρ et d'ensemble de taxons $X = \{a, b, c, d, e, f, g, h, i\}$. Les sommets h_i sont des sommets hybrides et ceux non étiquetés sont des sommets de spéciation.

Notons que parler de multigraphe, c'est à dire autoriser la présence de plusieurs arcs entre deux sommets, est un détail technique qui permet la présence de cycles à deux sommets dans le réseau phylogénétique, comme celui contenant h_1 en figure 2(a).

Un graphe orienté est dit *biconnexe* s'il ne contient aucun sommet d'articulation (dont la suppression déconnecte le graphe). Une *composante biconnexe* (ou *blob*) d'un réseau phylogénétique N est un sous-graphe biconnexe maximal de N. Pour tout arc (u, v) de N, on appelle u un père de v, et v un fils de u.

Un réseau phylogénétique explicite est dit de *niveau* k [JAN 04] si toute composante biconnexe contient au plus k sommets hybrides. Un réseau de niveau k qui n'est pas de niveau k-1 est dit strictement de niveau k. Par exemple, dans la Figure 2(a), la composante biconnexe de N qui contient le plus de sommets hybrides est située dans la zone grise (elle contient h_3 et h_4), donc N est strictement de niveau 2.

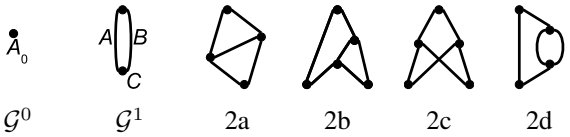


FIGURE 1. *Le générateur $\mathcal{G}^0$ de niveau 0 (a), le générateur $\mathcal{G}^1$ de niveau 1 (b), et les générateurs de niveau 2, appelés 2a, 2b, 2c et 2d [IER 08]. Tous les arcs sont dirigés vers le bas (l'orientation n'est pas indiquée pour un souci de lisibilité.)*

Ce paramètre exprime à quel point le réseau est proche d'un arbre : un réseau de niveau 0 est un arbre phylogénétique, un réseau de niveau 1 est communément appelé *galled tree*. De nombreux problèmes NP-complets peuvent être résolus en temps polynomial sur ces classes de réseaux phylogénétiques [GAM], ce qui motive l'étude des niveaux supérieurs.

En section 2, on étudie la structure de ces réseaux en montrant qu'ils peuvent avoir une grande complexité intrinsèque. Nous considérons ensuite, en section 3, un ensemble de réseaux simulés selon le modèle coalescent avec recombinaison pour montrer que dans ce contexte, les réseaux ont un niveau élevé, ce qui réduit l'application pratique des algorithmes de reconstruction existants.

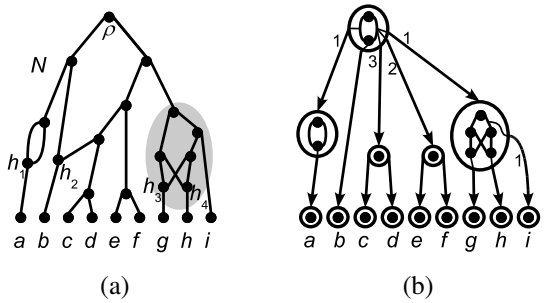
2. Décomposition des réseaux de niveau k

Définition 1 ([IER 08]) *Un* générateur de niveau k *est un réseau phylogénétique biconnexe strictement de niveau k (voir figure 1).*

Ces générateurs ont été introduits à l'origine dans le contexte d'une sous-classe de réseaux phylogénétiques dits *simples*, contenant une seule composante biconnexe. Nous montrons dans le théorème suivant qu'ils permettent de décomposer tout réseau de niveau k en un arbre de générateurs.

Théorème 1 (décomposition des réseaux de niveau k) *Tout réseau N de niveau k peut être décomposé de façon unique en un arbre de générateurs de niveau au plus k.*

L'arbre de décomposition en générateurs d'un réseau de niveau k est illustré en figure 2(b). Il correspond globalement à l'arbre de décomposition en composantes biconnexes du graphe, avec un intérêt supplémentaire dans notre cas : pouvoir étiqueter les noeuds de l'arbre de décomposition par un générateur extrait d'un ensemble fini (à niveau fixé).



(a) (b)

FIGURE 2. *Un réseau phylogénétique de niveau 2 (a) et son arbre de décomposition en générateurs (b) : la numérotation sur les arcs de l'arbre de décomposition indique dans quel ordre les générateurs sont attachés aux arcs du générateur du noeud père.*

Ce théorème de décomposition en générateurs demande donc une étude un peu plus précise de la structure des générateurs. Nous fournissons quelques propriétés sur leur taille, leur nombre, ainsi qu'un algorithme pour construire tous les générateurs de niveau k à partir des générateurs de niveau $k-1$.

Propriété 1 *Pour $k \geq 1$, un générateur de niveau k a au plus $3k-1$ sommets et $4k-2$ arcs.*

Propriété 2 *Le nombre g_k de générateurs de niveau k est compris entre 2^{k-1} et $k!^2 50^k$.*

Ces bornes très peu fines servent essentiellement à montrer que le nombre de générateurs est exponentiel en fonction du niveau. Ceci indique que la vision d'un réseau phylogénétique comme un arbre de blobs cache derrière l'apparente simplicité de l'arbre une grande complexité de structure à l'intérieur des blobs. Toutefois, elles permettent aussi de noter qu'il semble possible de construire automatiquement l'ensemble des générateurs de niveau 4, alors que jusqu'alors seuls ceux de niveau au plus 3 avaient été construits [KEL].

Théorème 2 *Un algorithme polynomial permet de construire l'ensemble de tous les générateurs de niveau $k+1$ à partir de l'ensemble S_k^* de tous les générateurs strictement de niveau k fourni en entrée.*

A la base de cet algorithme, et des deux propositions précédentes, se trouvent deux règles permettant d'attacher un nouveau sommet hybride à l'intérieur d'un générateur de niveau k. L'algorithme consiste donc à construire progressivement l'ensemble S_{k+1}^* en considérant à tour de rôle chaque générateur de l'ensemble S_k^*, et en appliquant une règle d'insertion d'un nouveau sommet hybride, et en vérifiant si le générateur de niveau $k+1$ ainsi créé est isomorphe à un des générateurs déjà ajoutés à S_{k+1}^*. Il faut noter que bien que la complexité du test d'isomorphisme de graphes soit encore indéterminée, nous pouvons le faire théoriquement en temps polynomial car nous travaillons sur des graphes orientés de degré maximum 3 [LUK 82, MIL 77]. En fait, l'algorithme de Luks est peu utilisable en pratique, et nous utilisons un algorithme exact exponentiel dans notre implémentation disponible à l'adresse `http://www.lirmm.fr/~gambette/ProgGenerators.php`.

Cette implémentation a permis de déterminer qu'il existait 1993 générateurs de niveau 4. On a ainsi pu vérifier que la séquence du nombre de générateurs de niveau k, 1,4,65,1993, n'était pas présente dans l'Encyclopédie en ligne des séquences d'entiers [SLO 08], alors que deux séquences de cette base de données contenaient 1,4,65.

3. Niveau de réseaux simulés

Arenas, Valiente et Posada ont étudié les propriétés de réseaux phylogénétiques simulés selon le modèle coalescent avec recombinaison [ARE 08], en mesurant la proportion parmi ces réseaux de ceux appartenant à certaines sous-classes, en particulier les arbres, et les réseaux "galled tree", c'est à dire les réseaux de niveau 0 et 1. Nous avons prolongé leur étude en calculant le niveau de tous les réseaux phylogénétiques générés par leur simulation qui a utilisé le programme Recodon [ARE 07]. L'implémentation en Java d'un algorithme basique de décomposition en composantes biconnexes pour calculer le niveau est également disponible à l'adresse `http://www.lirmm.fr/~gambette/ProgGenerators.php`.

Pour de petites valeurs du niveau, les résultats obtenus sont réunis dans la Table 1. Nous observons que les réseaux phylogénétiques avec un petit niveau, comme les autres restrictions étudiées dans la référence [ARE 08], ne couvrent qu'une portion réduite des réseaux phylogénétiques correspondant au modèle coalescent avec de forts taux de recombinaison. En fait, les réseaux simulés selon ce modèle n'ont pas vraiment la structure arborée exprimée dans le Théorème 1, mais sont le plus souvent constitués d'une grosse composante biconnexe qui contient tous les sommets hybrides. Ce phénomène apparaît même pour de faibles taux de recombinaison.

Ainsi, dans ce contexte, de nouvelles structures et techniques algorithmiques doivent être étudiées. Mentionnons toutefois que ce modèle ne convient pas pour décrire tous les cas d'évolution réticulée, et que d'autres peuvent être plus appropriés, comme celui qui insère des transferts horizontaux selon une loi de Poisson [GAL 07], ou ceux utilisés pour la simulation de réseaux phylogénétiques dans NetGen [MOR 06].

r	arbre	niveau 1	niveau 2	niveau 3	niveau 4	niveau 5
0	1000	1000	1000	1000	1000	1000
1	139	440	667	818	906	948
2	27	137	281	440	582	691
4	1	21	53	85	136	201
8	0	1	1	6	7	12
16	0	0	0	0	0	0

TABLE 1. *Nombre de réseaux simulés selon le modèle coalescent avec recombinaison sur 10 feuilles, ayant niveau 0, 1, 2, 3, 4, 5, en fonction du taux de recombinaison $r = 0, 1, 2, 4, 8, 16$.*

4. Conclusion

Devant l'engouement récent pour les réseaux de niveau k, qui permettent d'obtenir des algorithmes efficaces pour la reconstruction phylogénétique de réseaux à partir de triplets, nous avons présenté des résultats qui permettent de mieux comprendre ces objets : simples par la structure arborée qui apparaît, complexes à l'intérieur des parties réticulées, puisqu'un choix exponentiel de structures y est possible, en fonction du niveau.

La validation des méthodes de reconstruction de réseaux de niveau k sur des données biologiques est en cours, et il est intéressant de voir si les nuances théoriques que nous apportons à propos de leur utilisation seront confirmées en pratique. Ces résultats relancent aussi l'intérêt pour d'autres paramètres sur les réseaux qui permettraient d'obtenir des algorithmes rapides, et quelques pistes semblent déjà prometteuses dans cette optique.

5. Bibliographie

[ARE 07] ARENAS M., POSADA D., Recodon : coalescent simulation of coding DNA sequences with recombination, migration and demography, *BMC Bioinformatics*, vol. 8, 2007, page 458.

[ARE 08] ARENAS M., VALIENTE G., POSADA D., Characterization of Reticulate Networks based on the Coalescent, *Molecular Biology and Evolution*, vol. 25, 2008, p. 2517-2520.

[DOO 99] DOOLITTLE W., Phylogenetic Classification and the Universal Tree, *Science*, vol. 284, 1999, p. 2124-2128.

[GAL 07] GALTIER N., A Model of Horizontal Gene Transfer and the Bacterial Phylogeny Problem, *Systematic Biology*, vol. 56, 2007, p. 633-642.

[GAM] GAMBETTE P., Who is Who in Phylogenetic Networks : Articles, Authors and Programs, `http://www.lirmm.fr/~gambette/PhylogeneticNetworks`.

[HUB 55] HUBBS C., Hybridization Between Fish Species in Nature, *Systematic Zoology*, vol. 4, 1955, p. 1-20.

[IER 08] VAN IERSEL L., KEIJSPER J., KELK S., STOUGIE L., HAGEN F., BOEKHOUT T., Constructing Level-2 Phylogenetic Networks from Triplets, *RECOMB'08*, vol. 4955 de *LNCS*, Springer Verlag, 2008, p. 450-462.

[JAN 04] JANSSON J., SUNG W.-K., Inferring a Level-1 Phylogenetic Network from a Dense Set of Rooted Triplets, *COCOON'04*, vol. 3106 de *LNCS*, Springer Verlag, 2004, p. 462-471.

[KEL] KELK S., `http://homepages.cwi.nl/~kelk/lev3gen/`.

[LUK 82] LUKS E., Isomorphism of Graphs of Bounded Valence Can Be Tested in Polynomial Time, *Journal of Computer and System Sciences*, vol. 25, n^o 1, 1982, p. 42-65.

[MIL 77] MILLER G., Graph Isomorphism, General Remarks, *STOC'77*, 1977, p. 143-150.

[MOR 06] MORIN M., MORET B., NETGEN : Generating Phylogenetic Networks with Diploid Hybrids, *Bioinformatics*, vol. 22, n^o 15, 2006, p. 1921-1923.

[SLO 08] SLOANE N., The On-Line Encyclopedia of Integer Sequences, 2008, Published electronically at `http://www.research.att.com/~njas/sequences/`.

[TO 09] TO T.-H., HABIB M., Level-k Phylogenetic Network can be Constructed from a Dense Triplet Set in Polynomial Time, *CPM'09*, 2009, à paraître.

Résumés de textes par extraction de phrases, algorithmes de graphe et énergie textuelle

Silvia Fernández, Eric SanJuan, Juan-Manuel Torres-Moreno

LIA & IUT STID, Université d'Avignon et des Pays de Vaucluse, France
{silvia.fernandez,eric.sanjuan,juan-manuel.torres}@univ-avignon.fr

RÉSUMÉ. *Lorsqu'il s'agit de résumer automatiquement une large quantité de texte, l'approche actuellement la plus répandue consiste à pondérer les phrases selon leur représentativité. Les calculs sont généralement effectués sur la matrice mots $\times$ phrases. Le résumé reprend alors les n phrases les plus lourdes dans l'ordre de leur occurrence. Comme il s'agit de matrices creuses, il est naturel de les représenter par des graphes et cela a conduit à appliquer aux phrases les mêmes algorithmes qu'au Web. Ainsi deux approches LexRank et TextRank ont été dérivées du très populaire algorithme PageRank. Cependant la même matrice peut être interprétée comme un système magnétique, et la pondération comme un calcul d'énergie. Cela conduit à des calculs de pondérations bien plus simples et qui pourtant produisent presque les mêmes classements. Il apparait alors que l'élément déterminant à la production de ces classements sont les chemins d'ordre 2 dans le graphe d'intersection des phrases.*

MOTS-CLÉS : *Résumé automatique par extraction, Page Rank, Algorithmes de graphes, Systèmes magnétiques, Traitement automatique de la langue naturelle écrite*

1. Introduction

Le résumé automatique par extraction de phrases (RAEP) est une des techniques de la fouille de données textuelles [IBE 07]. Il consiste à pondérer les phrases selon leur représentativité dans le texte et à afficher celles de poids le plus fort dans l'ordre de leur apparition dans le texte et dans la limite de la taille du résumé. Généralement les systèmes combinent une large variété de fonctions de pondération (tel que le système CORTEX [TOR 02]) qui assure que le vocabulaire du résumé corresponde bien au contenu informatif du texte initial. Du point de vue de ce seul critère informatif ce type de systèmes ne sont pas de si "mauvais élèves" [FER 08]. Cependant, pour améliorer la lisibilité du résumé produit, et en particulier la résolution des anaphores les plus visibles, des post-traitements s'avèrent nécessaires.

Quoiqu'il en soit, la majorité des méthodes de résumé automatique évaluées lors des conférences DUC/TAC[1] menées par l'agence NIS[2] utilisent une représentation des phrases par sac de mots. Ainsi le texte est assimilé à une matrice S phrases$\times$mots qui code les occurrences des mots dans les phrases sans tenir compte de leur position.

Contrairement au cas des matrices textes$\times$termes utilisées en Recherche d'information (RI) ou l'analyse de données textuelles, le fait de travailler au niveau des phrases et non à celui plus large de textes entiers, fait que les valeurs $S_{i,j}$ sont généralement 0 ou 1. En effet, les mots qui apparaissent plus d'une fois dans une même phrase sont rarement informatifs et correspondent plutôt à des articles, des adjectifs non qualificatifs ou des verbes auxiliaires. Ainsi les matrices S considérées pour le résumé automatique par extraction de phrases sont majoritairement des matrices binaires symétriques creuses qui peuvent être avantageusement représentées par des graphes non orientées.

1. http ://duc.nist.gov/pubs.html et http ://www.nist.gov/tac/
2. http ://www.nist.gov/

2. Représentation du texte sous forme d'un graphe

Les phrases étant ramenées à des ensembles de mots, un texte T de P phrases peut être représenté par un hypergraphe H_T, c'est à dire un famille d'ensembles de mots : $\Phi_T = \{\varphi_1, ..., \varphi_P\}$ où chaque φ_i est l'ensemble de mots $\{w_{i,1}, ..., w_{i,l_i}\}$ de la i° phrase du texte, l_i étant sa longueur. De H_T on dérive le graphe valué G_T d'intersection des phrases. G_T est un triplé (V_T, E_T, L_T) tel que :

1. V_T est l'ensemble Φ_T des phrases du texte T,
2. E_T est l'ensemble de paires $\{\varphi_i, \varphi_j\}$ d'éléments de Φ_T tel que $\varphi_i \cap \varphi_j \neq \emptyset$
3. L_T est une fonction définie sur E_T par $L_T(\varphi_i, \varphi_j) = |\varphi_i \cap \varphi_j|$

G_T correspond simplement à la matrice carrée $S \times S^t$. Outre son adaptation à la représentation informatique de larges matrices creuses, l'intérêt de cette représentation sous forme de graphes est aussi de suggérer l'utilisation d'une large famille de calcul bien connus de centralité[3] de nœuds dans un graphe issus de l'analyse des réseaux sociaux (ARS) (*Social Network Analysis*[4]). Bien sûr, la fonction de valuation des arêtes L_T peut être remplacée par toute mesure de similarité entre deux ensembles cependant la structure du graphe, c'est à dire le couple (V_T, E_T), reste inchangée contrairement au cas général des graphe seuils considérés en analyse de similarité[5]. Nous allons montrer que c'est cette structure joue un rôle fondamental en REAP.

Cependant l'algorithme TEXTRANK[MIH 04] le plus répandu de REAP qui repose sur un parcours du graphe G_T n'a pas été inspiré par l'ARS, mais par l'algorithme PAGERANK[6] [PAG 98] utilisé par le moteur de recherche Google pour calculer l'importance des pages web liées par des hyperliens. De façon intuitive, une page aura un score PAGERANK haut s'il existe plusieurs pages qui la signalent ou s'il y a quelques unes mais avec un score élevé. PAGERANK prend en compte le comportement d'un surfeur aléatoire qui, à partir d'une page choisie au hasard, commence à cliquer sur les liens contenues dans ce site. Éventuellement il peut sortir de ce chemin et recommencer aléatoirement dans une autre page. Vu d'un autre angle, l'algorithme PAGERANK correspond à une variante de la méthode de calcul par puissance successives (*Power Iteration*[7]) pour calculer le premier vecteur propre de la matrice correspondant au graphe des liens entre pages. Le vecteur des scores R correspond ainsi aux composantes du premier vecteur propre de la matrice carrée M de liens entre pages [PAG 98]. L'algorithme PAGERANK a été transposé au traitement de textes par [MIH 04]. L'auteur a assimilé les phrases aux pages web et les liens aux ensembles de termes partagés. Leur système, connue sous le nombre de TEXTRANK, calcule les rangs des phrases dans les documents. Différentes variantes de PAGERANK existent selon la définition de la fonction L_T de valuation et l'initialisation de la méthode. Cependant le principe reste le calcul approximatif du premier vecteur propre d'une matrice creuse dont le graphe correspondant a la même structure que le graphe G_T défini précédemment. Or, si le calcul par puissances successives du premier vecteur propre reste pertinent pour de très larges matrices creuses, dans le cas de textes relativement courts, tels que ceux qui ont été utilisés dans les conférences DUC/TAC, on peut directement procéder au calcul de ce vecteur.

3. L'Énergie textuelle

Si l'on voit maintenant la matrice S comme un système magnétique de spins tel que en [FER 07], alors cette nouvelle analogie conduit à calculer les énergies E d'interaction, ce qui dans le modèle le plus simple correspond au carré de la matrice de $G_T : E = (S \times S^T)^2$. Cette matrice E donne le nombre de chemins de longueur au plus deux entre deux sommets de G_T.

Dans ce cas on prend comme pondération des phrases la vecteur $E.\vec{1}$. Cette nouvelle approche a été appelée ENERTEX. Par rapport à TEXTRANK elle est plus simple puisque elle se limite aux deux premières itérations,

3. http ://en.wikipedia.org/wiki/Centrality
4. http ://en.wikipedia.org/wiki/Social_network_analysis
5. Un graphe seuil (V_s, E_s) découle d'une fonction de similarité f définie sur toute les paires de V_s et d'un seuil s avec $E_s = \{\{u, v\} : f(u, v) > s\}$. Dans ce cas la structure du graphe dépend du seuil s.
6. Marque déposée de la société Google
7. http ://en.wikipedia.org/wiki/Power_iteration

alors que TEXTRANK décrit un processus itératif (de 30 pas approximativement) basé sur le calcul du premier vecteur propre de la matrice de liens entre phrases. Il se trouve que d'après les résultats présentés en [FER 07], ENERTEX atteint les mêmes performances que TEXTRANK du moins en ce qui concerne les mesures ROUGE sur la distribution des mots et des bi-grames couramment utilisées lors des campagnes du NIST. Comme TEXTRANK procède par ailleurs à des post-traitements pour améliorer la lisibilité du résumé, traitements qui peuvent avoir un impact sur ces mesures, une comparaison directe s'avère intéressante. Pour cela nous avons réalisé deux types d'expériences : sur un ensemble de matrices aléatoires en calculant directement le vecteur propre principal de ces matrices et sur un ensemble de textes réels, en calculant les scores TEXTRANK. Dans les deux cas, les rangs obtenus seront comparés à ceux induits par $E.\vec{1}$.

3.1. *Comparaison théorique sur des matrices aléatoires*

Pour comparer les classements issus du vecteur propre principal avec ceux obtenus par l'énergie textuelle, nous avons réalisé l'expérience suivante. En utilisant le logiciel statistique R [8], nous avons défini un ensemble de matrices entières positives M de taille arbitraire P comme le produit matriciel $S \times S^T$ où S est une matrice binaire de D lignes et N colonnes. Nous supposons que pour quelque $0 < i \leq P, 0 < j \leq N$, la probabilité d'avoir $S_{i,j} = 1$ est une constante p. Il est clair que : P est le nombre de phrases à pondérer, N est le nombre de mots différentes et p est la probabilité qu'un terme t se trouve dans une phrase μ. Pour chaque matrice nous avons calculé la matrice d'énergie $E = (S \times S^T)^2$ et le vecteur $\overrightarrow{e} = E.\overrightarrow{1}$, où $\overrightarrow{1} = (1, ..., 1)$ et $\overrightarrow{e}$ est le score donné par l'énergie textuelle. Additionnellement nous avons obtenu le vecteur propre principal $\overrightarrow{r}$ de M. Enfin, nous avons comparé les scores induits pour chaque vecteur en utilisant le test τ de Kendall[9]. Nous avons expérimenté les triplets suivants de valeurs (P, N, p) : (100 ;100 ;0,01), (500 ;100 ;0,01), (500 ;100 ;0,001), (1000 ;100 ;0,01) et (1000 ;100 ;0,001) en répétant le processus 30 fois pour chaque triplet. Nous avons alors calculé la valeur minimale obtenue pour le τ de Kendall. Nous avons obtenu $|\tau| > 0,8$, ce qui induit une p valeur inférieure à 10^{-5}.

Ce résultat indique que pour des matrices entières aléatoires, les rangs basés sur le calcul de l'énergie textuelle ($E = (S \times S^T)^2$), sont fortement corrélés avec les rangs induit par le vecteur propre principal de la matrice $S \times S^T$.

3.2. *Comparaison sur des textes*

Pour mener une comparaison sur des textes réels nous avons implémenté l'algorithme TEXTRANK [MIH 04]. Nous avons utilisé les deux systèmes, TEXTRANK et énergie textuelle, pour classer les phrases d'une vingtaine de documents issus du corpus DUC 2002[10] choisis aléatoirement. Les classements obtenues sont très similaires surtout dans l'assignation des premières places. Un exemple est montré au tableau 1. Il correspond au document sur l'ouragan *Gilbert* utilisé par [MIH 04] pour illustrer le fonctionnement du système. À gauche, les scores normalisés des 25 phrases du texte obtenus pour la méthode d'énergie et TEXTRANK. À droite, les quatre phrases les plus pertinentes qui seraient sélectionnées pour produire, par exemple, un résumé d'environ 100 termes.

4. Discussion

L'explication de ce phénomène semble se trouver dans le degré de connectivité du graphe. Plus ce facteur est grand, plus efficacement sera calculé la relation entre sommets depuis les premières itérations. En fait, un texte mono-thématique est un système où le haut degré de corrélation entre les phrases, produit du partage de termes, donne lieu à un graphe avec un haut degré de connectivité. La connectivité est clairement plus forte entre les

8. http ://www.r-project.org
9. Le coefficient τ de Kendall est proche de 0 dans le cas d'une indépendance totale entre les scores, proche de 1 pour une concordance parfaite et -1 pour des rangs opposés. La p-value donne la probabilité de l'hypothèse nulle d'indépendance statistique.
10. Les conférences DUC 2002 ont concerné aux tâches de résumé automatique monodocument. Le corpus contient $\approx$ 600 documents généralistes d'une trentaine de phrases chacun.

Rang	Phrase	Score Energie	Phrase	Score TextRank
1	**9**	1,00	**9**	1,00
2	**15**	0,89	**16**	0,90
3	**18**	0,83	**18**	0,86
4	**16**	0,73	**15**	0,74
5	20	0,32	5	0,66
6	14	0,31	14	0,60
7	10	0,31	21	0,56
8	17	0,28	10	0,54
9	5	0,23	12	0,51
10	13	0,21	20	0,46
11	21	0,20	23	0,44
12	11	0,19	13	0,42
13	22	0,18	4	0,39
14	4	0,17	8	0,38
15	23	0,13	17	0,38
16	24	0,12	22	0,38
17	12	0,11	11	0,31
18	8	0,10	24	0,27
19	7	0,03	6	0,08
20	19	0,02	7	0,08
21	3	0,01	19	0,08
22	6	0,00	3	0,00
23	2	0,00	2	0,00
24	1	0,00	1	0,00
25	0	0,00	0	0,00

Les deux systèmes classent en premières places les même quatre phrases :

9 Hurricaine Gilbert Swept towrd the Dominican Republic Sunday, and the Civil Defense alerted its heavily populated south coast to prepare for high winds, heavy rains, and high seas.

15 The National Hurricaine Center in Miami reported its position at 2 a.m. Sunday at latitude 16.1 north, longitude 67.5 west, about 140 miles south of Ponce, Puerto Rico, and 200 miles southeast of Santo Domingo.

16 The National Weather Service in San Juan, Puerto Rico, said Gilbert was moving westard at 15 mph with a ”broad area of cloudiness and heavy weather” rotating around the center of the storm.

18 Strong winds associated with the Gilbert brought coastal flooding, strong southeast winds, and up to 12 feet to Puerto Rico ?s south coast.

TAB. 1. *Score des 25 phrases d'un des documents du corpus DUC 2002 obtenus par le calcul de l'énergie textuelle et le système* TEXTRANK. *Les résultats sont similaires, surtout pour les phrases classées en premières places.*

phrases d'un même texte qu'entre les hyperliens entre pages web, ce qui explique les bonnes performances de notre méthode.

5. Bibliographie

[FER 07] FERNÁNDEZ S., SANJUAN E., TORRES-MORENO J. M., Energie textuelle des mémories associatives, ET PHILIPPE MULLER N. H., Ed., *Actes de TALN 2007*, Toulouse, June 2007, ATALA, IRIT, p. 25–34.

[FER 08] FERNÁNDEZ S., VELÁZQUEZ P., MANDIN S., SANJUAN E., MANUEL J. T.-M., Les systèmes de résumé automatique sont-ils vraiment des mauvais élèves ?, *Actes de Journées internationales d'Analyse statistique des Données Textuelles JADT 2008*, 2008.

[IBE 07] IBEKWE-SANJUAN F., *Fouille de textes : méthodes, outils et applications*, Paris , Hermes Science Publications, Lavoisier, Paris, France, 2007.

[MIH 04] MIHALCEA R., Graph-based ranking algorithms for sentence extraction, applied to text summarization, *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*, Morristown, NJ, USA, 2004, Association for Computational Linguistics, page 20.

[PAG 98] PAGE L., BRIN S., MOTWANI R., WINOGRAD T., The PageRank Citation Ranking : Bringing Order to the Web, rapport, 1998, Stanford Digital Library Technologies Project.

[TOR 02] TORRESMORENO J.-M., VELÁZQUEZMORALES P., MEUNIER J., Condensés de textes par des méthodes numériques, *JADT*, vol. 2, 2002, p. 723–734.

Analyse de graphes de données textuelles et règles d'association

Bangaly Kaba[*]**, Eric SanJuan**[**]

[*]*LIMOS, Université Blaise Pascal Clermont 2, France*
kaba@isima.fr
[**]*LIA & IUT STID, Université d'Avignon, France*
eric.sanjuan@univ-avignon.fr

RÉSUMÉ. *Les matrices de données textuelles sont par nature très creuses et il est courant de préférer les représenter sous forme de graphes. Dans ce formalisme, la classification à lien simple ou ses variantes consistent à sélectionner un sous-ensemble d'arêtes (sous-graphe) et à calculer les composantes connexes de ce dernier. Le concept d'atome de graphe permet de désarticuler une composante connexe en une famille de sous graphes non-disjoints dont les intersections correspondent à des cliques maximales. Nous présentons ici des résultats expérimentaux sur une large variété de données textuelles qui montrent les propriétés de ces atomes vis-à-vis des ensembles d'items fréquents et des règles d'association. La désarticulation des graphes en atomes induit alors une nouvelle méthode de classification en classes non disjointes compatible avec les règles d'association.*

MOTS-CLÉS : *Algorithmes de graphes, fouille de données textuelles, classification non supervisée, désarticulation de graphes, règles d'association*

1. Introduction

Depuis quelques années, de nombreuses méthodes de décomposition de graphes sont proposées. Il existe les méthodes basées sur les calculs de coupe minimale pour partitionner de façon récursive un graphe valué en composantes tel que celle utilisée par Shamir et al. ([RSS00]). Voy et al. ([VSP06]) explorent par contre toutes les cliques maximales d'un graphe. Seno et al. ([STT04]) définissent la notion de sous graphes p-quasi complets. Tous ces travaux récents partagent la recherche des parties fortement connectées d'un graphe et les résultats montrent l'importance de ces parties. Cependant, l'un des problèmes rencontrés reste le grand nombre de cliques ou de p-cliques qu'il peut y avoir dans un graphe. Ainsi, ces méthodes sont coûteuses et demandent des heuristiques pour assurer un bon choix.

En suivant Tarjan [TA85], Berry [ABER98] et les travaux de thèse en [KABA08], nous proposons d'utiliser les séparateurs minimaux complets pour décomposer un graphe. Les groupes de sommets définis par la décomposition appelés atomes ne sont pas disjoints, les séparateurs minimaux complets étant recopiés dans le but de préserver la structure du graphe. Les séparateurs minimaux complets sont définis de façon unique et il en existe par définition moins que de sommets. Les algorithmes de triangulation permettent de les calculer de façon efficace et la décomposition qui en résulte est unique : pour un graphe donné, les atomes sont les mêmes et indépendants de l'algorithme qui les calcule. Par contre un graphe quelconque, même très grand, peut n'avoir qu'un seul atome, lui-même, c'est le cas par exemple de tout cycle sans corde. Après avoir défini exactement la notion de graphes d'atomes, nous montrons ici des exemples de graphes de données textuelles pour lesquels non seulement les atomes existent en grand nombre, mais de plus s'avèrent être stables vis-à-vis des règles d'association.

2. Graphe des atomes

Pour caractériser cet objet nous avons besoin au préalable de quelques définitions générales sur les graphes. Un **graphe non orienté** est défini par un ensemble fini de sommets $V = \{v_i\}$ et un ensemble fini d'arêtes $E = \{e_k\}$. Chaque arête est caractérisée par une paire $\{v_i, v_j\}$ de sommets, appelés extrémités. On note $G = (V, E)$. Deux sommets x et y sont dits adjacents lorsqu'ils sont reliés par une arête, x voit y. L'arête est dite incidente aux deux sommets. Un graphe est dit **complet** si tous ses sommets sont deux à deux adjacents. On dit que $G' = (V' E')$ est un sous graphe de $G = (V, E)$ si $V' \subseteq V$ et $E' \subseteq E$. Une **clique** est un sous-graphe complet. Dans un graphe, une chaîne est constituée d'une suite de sommets adjacents. Un cycle est une chaîne simple fermée d'un graphe non orienté. Une corde dans un cycle est une arête entre deux sommets non consécutifs. Un graphe $G = (V, E)$ est dit **connexe** lorsque pour tout $(x, y) \in V^2$, il existe une chaîne les reliant x et y. Une composante connexe d'un graphe est un sous graphe connexe non vide, maximal au sens du nombre de sommets.

Soient donc $G(V, E)$ un graphe connexe et x, y deux sommets non adjacents. Un ensemble de sommets S est un xy**-séparateur** si la suppression des sommets de S place x et y dans deux composantes connexes différentes du graphe $G[V - S]$. S est un xy-séparateur minimal s'il n'existe pas de xy-séparateur S' proprement inclus dans S. S est un **séparateur minimal** s'il existe x et y pour lesquels S est un xy-séparateur minimal. Nous avons alors les propriétés suivantes pour tout $S \subset V$ et $x, y \in V - S$. Si S est un séparateur d'un graphe $G = (V, E)$ et et C une composante connexe de $G = (V, E)$, alors C est dit **composante pleine** pour S quand $N(C) = S$. En fait, si $G = (V, E)$ est un graphe alors un séparateur $S \in V$ de G est un séparateur minimal si et seulement si $G(V - S)$ admet au moins deux composantes connexes pleines. De plus si C_1 et C_2 sont des composantes connexes pleines induites par S alors tout sommet de S a un voisin dans C_1 et un voisin dans C_2.

Un séparateur minimal S est appelé **séparateur minimal complet** si le sous graphe induit par S est une clique. On appelle **décomposition par séparateurs minimaux complets** d'un graphe G les sous-graphes obtenus en répétant l'étape de décomposition précédente sur chacun des sous-graphes engendrés, jusqu'à ce qu'aucun de ces sous-graphes n'admette de séparateur minimal complet.

Définition 2.1 *Nous appellerons* **atome** *de $G = (V, E)$ un sous-ensemble de sommets A de V qui induit un sous-graphe $G(A)$ connexe, sans séparateur minimal complet, et maximal pour ces deux propriétés. Pour un atome A de G on dira aussi que $G(A)$ est un atome.*

La décomposition par séparateurs minimaux complets a la propriété d'être cohérente, en ce sens que chaque séparateur minimal complet choisi dans l'un des sous-graphes obtenus en cours de décomposition est aussi un séparateur minimal complet du graphe de départ quelque soit l'ordre dans lequel on procède à la décomposition. Berry dans sa thèse [ABER98] a démontré qu'une décomposition par séparateurs minimaux complets était unique (voir aussi [LE93]). Apparu dans la littérature comme un résultat de la décomposition d'un graphe triangulé, le graphe des atomes n'a pas constitué jusqu'ici l'objet d'une étude systématique, encore moins lorsqu'il résulte d'une décomposition d'un graphe non triangulé. Pour nous par contre, le graphe des atomes, structure unique sous-jacente à tout graphe, constitue un outil d'analyse de graphes réels issus en particulier de matrices creuses de corrélations, lorsque l'on se donne un seuil en dessous duquel la corrélation est ignorée.

Définition 2.2 *Le graphe des atomes que nous notons $G_{At} = (V_{At}, E_{At})$ est défini comme suit : Les sommets de G_{At} sont les atomes obtenus après la décomposition du graphe G. Une arête $e = (w_1, w_2)$ est définie entre deux atomes A_1 et A_2 s'il existe un séparateur minimal complet S dans G qui sépare l'atome A_1 de l'atome A_2.*

G_{At} a donc autant de sommets que le graphe d'intersection des atomes mais moins d'arêtes. Il y a une relation forte entre séparateurs minimaux complets et triangulation de graphes. Un graphe $G = (V, E)$ est triangulé s'il ne contient pas de cycle sans corde de longueur supérieure à trois. Il s'ensuit que les seuls séparateurs minimaux d'un graphe triangulé sont complets. On appelle **triangulation minimale** de G tout graphe triangulé $H = (V, E + F)$, tel que pour toute partie propre F' de F, le graphe $H' = (V, E + F')$ n'est pas triangulé. Il se trouve que [PS95] un séparateur minimal d'un graphe connexe G est un séparateur minimal complet si et seulement si il est un séparateur

minimal de toute triangulation minimale de G. Il est alors possible de calculer les séparateurs minimaux d'un graphe en $O(|V||E|)$ en calculer une triangulation quelconque H de G, puis calculer les séparateurs minimaux de H et terminer en testant parmi les séparateurs minimaux de H ceux qui étaient déjà complets dans G.

3. Applications à des graphes de données textuelles

Nous montrons comment la décomposition de graphes permet de calculer et de mettre en évidence une famille de clusters non disjoints particulièrement stables vis à vis des règles d'association. Nous évaluons notre approche en utilisant des corpus de notices extraits du WebOfScience. Le premier porte sur le texte mining et le second sur le terrorisme. Pour le premier nous utilisons directement les listes de mots clefs fournies par le WebOfScience. Pour le second nous utilisons une extraction de termes produite par le système TermWatch [SI06].

Nous considérons d'abord le corpus "datamining" utilisé en [PS07] contenant $3,671$ enregistrements extraits de la base de données SCI (Science Citation Index) accessible via le WebOfScience, traitant du datamining et du textmining sur la période 2000-2006 indexée par un ensemble de 8040 mots clés. La moyenne des mots clés par enregistrement est de 5. Le nombre de mots clés d'une fréquence supérieure à 1 est de $1,524$ et indexent $2,615$ enregistrements. Chaque notice est représentée par l'ensemble de ses mots clefs. Cela constitue un hypergraphe H de $3,171$ hyperarêtes (documents indexés par au moins un mot clé) sur 3.671 hyper-sommets (les mots clés). Cet hypergraphe conduit à 7.082 règles d'association ayant un support supérieur à $0,1\%$ et ayant une confiance supérieure à 80%. Toutes les règles d'association sont de la forme $k_1, ..., k_n \rightarrow c$ avec $1 < n \leq 5$ et signifie que plus de 80% des arêtes de H contenant $k_1, ..., k_n$, contiennent aussi c, $k_1, ..., k_n, c$ étant tous de mots clefs différents.

De l'hypergraphe dual de H (les mots-clefs sont représentés par les ensembles des notices qu'ils indexent), nous dérivons le graphe d'association G_1 des mots clefs apparaissant ensemble dans au moins deux notices. Ce graphe a 645 sommets, 1057 arêtes et est divisé en une principale composante connexe avec 413 sommets et 15 petites composantes connexes avec moins de 16 sommets. Nous avons basé notre étude sur la principale composante que nous avons assimilé au graphe entier G_1. G_1 est un graphe petit monde (SWG) avec un coefficient de clustering moyen de 0.47. Cette valeur est loin de la valeur attendue pour un graphe aléatoire ayant le même degré moyen qui est la moyenne sur l'ensemble des arêtes (4.43/2335). Comme dans les graphes aléatoires, le chemin moyen est faible : 2.321.

Le graphe des atomes $G_1(At)$ a été calculé en moins d'une minute sur un PC standard. Il a 404 atomes dont un central contenant 298 sommets. Les autres 403 atomes ont moins de 13 sommets. Donc G_1 est clairement divisé en un noyau central et une périphérie. Le résultat remarquable que nous obtenons est que 96% de ces 403 petits atomes sont stables vis à vis des règles d'association.

Les règles d'association sont calculées à partir des ensembles fréquents d'items (EFI) $k_1, ..., k_n$ (sous-ensembles de mots clefs apparaissant ensemble dans plus de $0,1\%$ de notices). Lorsque le seuil utilisé pour construire le graphe d'association G_1 correspond ou est proche du support minimal d'un EFI, alors chaque arête correspond à un EFI de taille 2 et tout EFI de taille n génère une clique dans G_1. Il y a donc une relation directe entre EFI et cliques. En particulier, on peut s'attendre à ce que la majorité des séparateurs minimaux corresponde à des EFI et aient des propriété de stabilité vis à vis des règles d'association. Ce que nous constatons expérimentalement et qui est plus surprenant, est que les atomes de ces graphes quelconques qui ne se réduisent pas à des cliques soient aussi stables vis à vis de ces règles.

Nous avons aussi appliqué cette méthode de décomposition au corpus issu de Chaomei et al.[CZZV07] extrait de la même base de données bibliographiques de SCI (Science Citation Index). Ce corpus a été choisi afin d'étudier l'évolution structurelle des réseaux de recherche sur le terrorisme. Nous avons utilisé le systeème TermWatch [1] pour extraire et sélectionner 57.855 syntagmes nominaux à partir de 3.366 résumés bibliographiques. Ces syntagmes nominaux ont été regroupés en 3.293 composantes de termes variants ayant au moins 2 syntagmes nominaux presques synonymes. La taille maximale de ces composantes de termes est de 30. 8.357 termes isolés mais

1. http ://index.termwatch.es

apparaissant dans au moins deux notices distinctes ont été ajoutés à l'ensemble des composantes de termes. A cette liste de termes nous avons encore ajouté le nom de tous les auteurs des articles. Nous avons alors considéré l'hypergraphe induit par la réunion des relations document-terme et document-auteur.

Le graphe d'associations auteurs-termes qui en découle a $16,258$ arêtes. Sa principale composante a $9,324$ arêtes et $1,070$ sommets. La décomposition de cette composante nous donne 489 atomes. L'atome central a $2,070$ arêtes et 307 sommets. Comme dans le cas du graphe des mots clés extraits du corpus sur le datamining, nous obtenons après la décomposition un atome central et une périphérie constituée de multiples petits atomes de moins de 20 éléments chacun. Bien que ce graphe d'association ait été obtenu par une toute autre méthode d'indexation, on retrouve une proportion similaire d'atomes 95% qui sont stables vis à vis des règles d'association.

4. Conclusion

Cette étude est une première tentative d'application de la décomposition de graphes à la cartographie des domaines de connaissance. L'avantage de la décomposition en atomes est qu'elle est unique. Elle est basée sur la seule structure du graphe. Son principal inconvénient est que les petits atomes n'existent pas toujours dans un graphe. Les deux expérimentations présentées ici tendent à montrer la décomposition en atomes s'adapte bien au corpus de thématiques bibliographiques puisque on trouve un large ensemble de petits atomes qui s'avèrent être stables à plus de 95% vis à vis des règles d'association de confiance élevée (80%). D'autres expérimentations sur des corpus plus artificiels de thématiques bibliographiques traitant de : Information Retrieval, genomics ou Organic Chemistry ont confirmé ces résultats.

5. Bibliographie

[ABER98] A.Berry. Désarticulation d'un graphe. Thèse de doctorat, LIRMM, Montpellier, décembre 1998.

[CZZV07] C. Chen, W. Zhu, B. Tomaszewski, A. MacEachren (2007). Tracing conceptual and geospatial diffusion of knowledge. HCI International 2007. Beijing, China. July 22-27, 2007. LNCS, 4564. pp. 265-274.

[KABA08] B.Kaba. Décomposition de graphes comme outil de clustering et de visualisation en fouille de données. Thèse de doctorat, LIMOS, Clermont Ferrand, novembre 2008.

[LE93] H. G.Leimer. Optimal Decomposition by Clique separators, Discrete Mathematics archive, 113(1-3), pp 99-123, 1993.

[PS07] X.Polanco,E.SanJuan. Hypergraph modelling and graph clustering process applied to co-word analysis. In : 11th biennial International Conference on Scientometrics and Informetrics, 2007.

[PS95] A.Parra and P. Scheffler. How to use the minimal separators of a graph for its chordal triangulation. Proc.22nd International colloquium on automata, Languages and Programming (ICALP'95) ; Lecture Notes in Computer Science, 944 : 123-134,1995.

[STT04] S.Seno, R.Teramoto, Y.Takenaka and H.Matsuda. A Method for Clustering Gene Expression Data Based on Graph Structure, Genome Informatics 2004, 15(2), pp 151-160.

[TA85] R.E.Tarjan.Decomposition by clique separators. Discrete Math :55 : 221-232, 1985.

[RSS00] R.Sharan and R.Shamir. CLICK : A Clustering Algorithm with Applications to Gene Expression Analysis, Proc. ISMB'00, AAAI Press, Menlo Park (CA, USA), pp 307-316, 2000.

[SI06] E.SanJuan, F.Ibekwe-SanJuan. Text mining without document context. Information Processing and Management, 42 1532-1552, 2006.

[VSP06] B.H.Voy, J. A.Scharff, A. D.Perkins, A.M.Saxton, B. Borate, E.J. Chesler, L.K.Branstetter and M.A.Langston. Extracting Gene Networks for Low-Dose Radiation Using Graph Theoretical Algorithms, PLOS Computational Biology, 2006.

Estimation des paramètres d'une loi de Weibull bivariée par la méthode des moments
Application à la séparation Monophonie / Polyphonie

Hélène Lachambre, Régine André-Obrecht et Julien Pinquier

IRIT, 118 route de Narbonne, 31062 Toulouse Cedex 9
{lachambre, obrecht, pinquier}@irit.fr

RÉSUMÉ. *Ce travail a pour contexte l'indexation de contenus musicaux. Afin de distinguer la musique monophonique, dans laquelle à chaque instant une seule note est jouée, de la musique polyphonique, nous avons exploité des paramètres caractéristiques de l'harmonicité, dans une approche statistique. Leur répartition est modélisée à l'aide de lois de Weibull bivariées, la difficulté réside dans l'estimation de leurs paramètres. Nous en présentons un approche originale basée sur la méthode des moments. L'évaluation de cette classification sur un corpus très varié ouvre des perspectives intéressantes en pré-traitement des analyses de musique.*

MOTS-CLÉS : *Loi de Weibull bivariée, méthode des moments, estimation des paramètres, monophonie, polyphonie.*

1. Introduction

Dans le cadre général de l'analyse automatique de la musique, nos travaux portent plus précisément sur la séparation entre la monophonie et la polyphonie. Ce problème a déjà été exploré par Smit et Ellis [SMI 07] pour la détection de la voix chantée solo. La monophonie est définie comme « une seule source harmonique », une source étant un chanteur ou un instrument de musique jouant une seule note. Ainsi, une monophonie est soit un chanteur *a capella*, soit un instrument de musique solo. La polyphonie, est définie comme « plusieurs sources harmoniques simultanées », ce qui regroupe les sous-classes suivantes : plusieurs instruments, plusieurs chanteurs, ou un (des) instrument(s) et un (des) chanteur(s). Notons qu'un même instrument pourra être, selon les cas, classé dans l'une ou l'autre catégorie : un piano jouant une seule note est monophonique, alors qu'un piano jouant plusieurs notes simultanément est polyphonique. Au cours de cette étude, nous avons été amenés à modéliser la répartition bivariée des paramètres caractéristiques.

En statistiques, de nombreuses lois de probabilité ont été proposées, la plus connue (et utilisée) en traitement de la musique et de la parole est la loi normale [SCH 97]. La loi de Weibull est surtout utilisée en fiabilité, notamment pour modéliser des taux de panne. Elle présente l'avantage de pouvoir avoir des formes très diverses (et proches de lois connues) en faisant varier ses paramètres. Toutes ces lois ont été étendues au cas bivarié au moyen des copules. Dans notre étude, nous avons choisi de modéliser la répartition bivariée des paramètres caractéristiques par des lois de Weibull bivariées.

Dans le cas bivarié, la question de l'estimation des paramètres de ces lois se pose. Il y a bien sûr toujours la possibilité d'utiliser des méthodes d'optimisation. Cependant, ces méthodes sont parfois gourmandes en temps de calcul, et surtout demandent de bien les initialiser pour qu'elles convergent. Une autre méthode pour l'estimation des paramètres est la méthode des moments. Les paramètres sont alors exprimés, pour des lois à 5 paramètres, en fonction des moments d'ordre 1, d'ordre 2 et du moment joint d'ordre (1,1).

Dans la partie suivante, nous présentons l'estimation des paramètres de la loi de Weibull bivariée par la méthode des moments. Dans la partie 3, nous présentons notre application, la séparation des deux classes audio : musique monophonique et musique polyphonique.

2. Une loi de Weibull bivariée

La loi que nous utilisons ici a été proposée par Hougaard, qui donne sa fonction de répartition dans [HOU 86] : $F(x,y) = 1 - \exp\left(-\left[\left(\frac{x}{\theta_1}\right)^{\frac{\beta_1}{\delta}} + \left(\frac{y}{\theta_2}\right)^{\frac{\beta_2}{\delta}}\right]^{\delta}\right)$, pour $(x,y) \in \mathbb{R}^+ \times \mathbb{R}^+$, avec $(\theta_1, \theta_2) \in \mathbb{R}^+ \times \mathbb{R}^+$ les paramètres d'échelle, $(\beta_1, \beta_2) \in \mathbb{R}^+ \times \mathbb{R}^+$ les paramètres de forme et $\delta \in]0,1]$ le paramètre de corrélation.

2.1. *Estimation des paramètres*

L'estimation des cinq paramètres est effectuée par la méthode des moments. Les moments de la loi que nous étudions sont donnés par Lu et Bhattacharyya [LU 90]. Des équations des moments d'ordre 1 et d'ordre 2, nous obtenons facilement les valeurs de β_1, β_2, θ_1 et θ_2. Nous les considérerons donc connus dans la suite.

δ semble par contre plus difficile à obtenir à partir de l'équation (1) du moment joint. Nous allons montrer que δ est le zéro d'une fonction $f(\delta)$ relativement simple et strictement décroissante.

$$Cov(X,Y) = \theta_1\theta_2 \frac{\Gamma\left(\frac{\delta}{\beta_1}+1\right)\Gamma\left(\frac{\delta}{\beta_2}+1\right)\Gamma\left(\frac{1}{\beta_1}+\frac{1}{\beta_2}+1\right) - \Gamma\left(\frac{1}{\beta_1}+1\right)\Gamma\left(\frac{1}{\beta_2}+1\right)\Gamma\left(\frac{\delta}{\beta_1}+\frac{\delta}{\beta_2}+1\right)}{\Gamma\left(\frac{\delta}{\beta_1}+\frac{\delta}{\beta_2}+1\right)} \quad (1)$$

Nous montrons tout d'abord que $(1) \Leftrightarrow f(\delta) = \delta B\left(\frac{\delta}{\beta_1}, \frac{\delta}{\beta_2}\right) = C$, avec $B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ la fonction Beta.

En posant $C_1 = \Gamma\left(\frac{1}{\beta_1}+1\right)\Gamma\left(\frac{1}{\beta_2}+1\right)$ qui ne dépend pas de δ, et en utilisant la propriété suivante des fonctions Gamma : $\Gamma(a+1) = a\Gamma(a)$, l'équation (1) peut s'écrire :

$$\frac{Cov(X,Y)}{\theta_1\theta_2} + C_1 = \frac{\frac{\delta}{\beta_1}\frac{\delta}{\beta_2}\Gamma\left(\frac{\delta}{\beta_1}\right)\Gamma\left(\frac{\delta}{\beta_2}\right)\Gamma\left(\frac{1}{\beta_1}+\frac{1}{\beta_2}+1\right)}{\frac{\delta\beta_1+\delta\beta_2}{\beta_1\beta_2}\Gamma\left(\frac{\delta}{\beta_1}+\frac{\delta}{\beta_2}\right)} \quad (2)$$

$$\frac{Cov(X,Y)}{\theta_1\theta_2} + C_1 = \frac{\Gamma\left(\frac{1}{\beta_1}+\frac{1}{\beta_2}+1\right)}{\beta_1+\beta_2}\delta\frac{\Gamma\left(\frac{\delta}{\beta_1}\right)\Gamma\left(\frac{\delta}{\beta_2}\right)}{\Gamma\left(\frac{\delta}{\beta_1}+\frac{\delta}{\beta_2}\right)} \quad (3)$$

Nous remarquons que $\frac{\Gamma\left(\frac{\delta}{\beta_1}\right)\Gamma\left(\frac{\delta}{\beta_2}\right)}{\Gamma\left(\frac{\delta}{\beta_1}+\frac{\delta}{\beta_2}\right)} = B\left(\frac{\delta}{\beta_1}, \frac{\delta}{\beta_2}\right)$ et que $C_2 = \frac{\Gamma\left(\frac{1}{\beta_1}+\frac{1}{\beta_2}+1\right)}{\beta_1+\beta_2}$ est independent de δ. Ainsi, l'équation (1) est équivalente à l'équation suivante (avec C constant) :

$$f(\delta) = \delta B\left(\frac{\delta}{\beta_1}, \frac{\delta}{\beta_2}\right) = \frac{\frac{Cov(X,Y)}{\theta_1\theta_2} + C_1}{C_2} = C \quad (4)$$

δ est donc solution de l'équation $f(\delta) - C = 0$.

2.2. *Dérivée de $f(\delta)$ et son signe*

Nous allons montrer que $f(\delta)$ est strictement décroissante en montrant que sa dérivée $f'(\delta)$ est strictement négative pour tout triplet $(\beta_1, \beta_2, \delta) \in \mathbb{R}^+ \times \mathbb{R}^+ \times]0,1]$. Ainsi, estimer δ est équivalent à déterminer l'unique zéro de $f(\delta) - C$.

La dérivée $f'(\delta)$ de $f(\delta)$ est (avec $\psi_0(x) = \frac{\mathrm{d}\ln\Gamma(x)}{\mathrm{d}x}$ la fonction digamma) :

$$f'(\delta) = B\left(\frac{\delta}{\beta_1}, \frac{\delta}{\beta_2}\right)\left[1 + \frac{\delta}{\beta_1}\left(\psi_0\left(\frac{\delta}{\beta_1}\right) - \psi_0\left(\frac{\delta}{\beta_1} + \frac{\delta}{\beta_2}\right)\right) + \frac{\delta}{\beta_2}\left(\psi_0\left(\frac{\delta}{\beta_2}\right) - \psi_0\left(\frac{\delta}{\beta_1} + \frac{\delta}{\beta_2}\right)\right)\right] \quad (5)$$

Nous savons que $B\left(\frac{\delta}{\beta_1}, \frac{\delta}{\beta_2}\right) > 0$, donc pour prouver que $f'(\delta) < 0$, nous devons montrer :

$$\frac{\delta}{\beta_1}\left(\psi_0\left(\frac{\delta}{\beta_1}\right) - \psi_0\left(\frac{\delta}{\beta_1} + \frac{\delta}{\beta_2}\right)\right) + \frac{\delta}{\beta_2}\left(\psi_0\left(\frac{\delta}{\beta_2}\right) - \psi_0\left(\frac{\delta}{\beta_1} + \frac{\delta}{\beta_2}\right)\right) < -1,\ \forall(\beta 1, \beta 2, \delta) \in \mathbb{R}^+ \times \mathbb{R}^+ \times]0, 1] \quad (6)$$

En posant $a = \frac{\delta}{\beta_1}$ et $b = \frac{\delta}{\beta_2}$, l'équation (6) ne dépend plus que de deux variables :

$$(6) \Leftrightarrow a(\psi_0(a) - \psi_0(a+b)) + b(\psi_0(b) - \psi_0(a+b)) < -1,\ \forall(a, b) \in \mathbb{R}^{+*} \times \mathbb{R}^{+*} \quad (7)$$

Pour prouver (7), nous utilisons les fonctions polygamma : $\psi_n(x) = \frac{\mathrm{d}^{n+1}\ln x}{\mathrm{d}x^{n+1}} = \frac{\mathrm{d}^n\psi_0(x)}{\mathrm{d}x^n}$ dont une propriété connue est : $\psi_m(x+1) = \psi_m(x) + (-1)^m m! x^{-(m+1)}$. Pour $m = 1$, nous avons : $\psi_1(x) = \psi_1(x+1) + \frac{1}{x^2}$.
Comme $\psi_1(x) > 0,\ \forall\ x > 0$ (car $\Gamma(x)$ est convexe), alors :

$$\psi_1(x) > \frac{1}{x^2},\ \forall x > 0 \quad (8)$$

En intégrant l'équation (8) entre a et $a+b$ (avec $a+b > a$), nous obtenons :

$$\psi_1(x) > \frac{1}{x^2} \Rightarrow \int_a^{a+b} \psi_1(x)\mathrm{d}x > \int_a^{a+b} \frac{1}{x^2}\mathrm{d}x \quad (9)$$

$$\psi_0(a+b) - \psi_0(a) > -\frac{1}{a+b} + \frac{1}{a},\ \text{puis}\ a(\psi_0(a+b) - \psi_0(a)) > \frac{b}{a+b} \quad (10)$$

Symétriquement, nous avons $b(\psi_0(b) - \psi_0(a+b)) < -\frac{a}{a+b}$, ce qui nous mène à l'inégalité que nous cherchions :

$$a(\psi_0(a) - \psi_0(a+b)) + b(\psi_0(b) - \psi_0(a+b)) < -1 \quad (11)$$

3. Une application : la séparation monophonie / polyphonie

3.1. *Les paramètres*

Dans leur algorithme YIN [CHE 02] d'estimation de la fréquence fondamentale (le « pitch »), de Cheveigné et Kawahara donnent un indice de confiance sur la valeur estimée du pitch, qu'on appelle ici $cmnd(t)$ (avec t l'indice de la trame de signal considérée).

Dans le cas d'un son monophonique, le résultat donné par l'estimateur de pitch est fiable, alors que dans le cas d'une polyphonie, ce résultat ne correspond *a priori* à aucune fréquence réelle. Ainsi, $cmnd(t)$ est faible et varie peu pour des sons monophoniques ; il est élevé et varie plus pour des sons polyphoniques. Ces considérations nous conduisent à utiliser comme paramètres caractéristiques la moyenne à court terme $cmnd_{moy}(t)$ et la variance à court terme $cmnd_{var}(t)$ de $cmnd(t)$. Celles-ci sont calculées sur une fenêtre glissante de 5 échantillons.

3.2. *Modélisation et classification*

La classification est faite toutes les secondes. La fréquence fondamentale étant estimée toutes les 10 ms, $cmnd(t)$, $cmnd_{moy}(t)$ et $cmnd_{var}(t)$ sont également calculés toutes les 10 ms, soit 100 valeurs par seconde.

Nous modélisons la répartition bivariée du couple $(cmnd_{moy}, cmnd_{var})$ pour chaque classe avec des modèles de Weibull bivariés (voir figure 1). La classification est réalisée en calculant la vraisemblance de 100 couples (1 s) par rapport à chacune des lois de Weibull estimées. Le résultat est la classe qui maximise la vraisemblance.

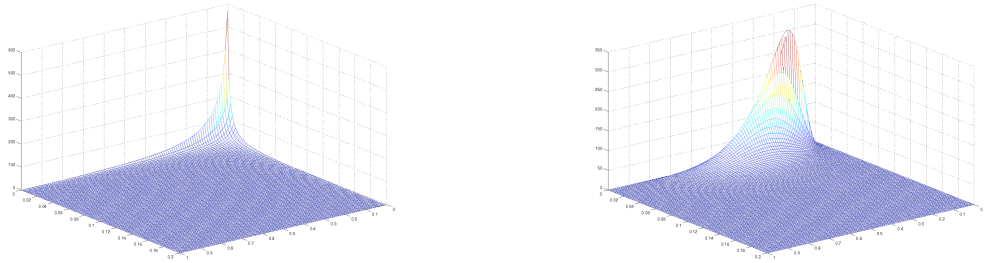


FIGURE 1. *Lois de Weibull estimées pour les classes Monophonie (gauche) et Polyphonie (droite).*

3.3. *Résultats*

Nous avons réalisé deux expériences. Dans la première, nous considérons deux classes : monophonie et polyphonie ; deux lois de Weibull sont estimées. La vraisemblance est calculée sur chaque modèle. Les lois sont apprises avec 50 s (monophonie) et 75 s (polyphonie) de signal, soit 5000 et 7500 couples $(cmnd_{moy}, cmnd_{var})$. Le taux d'erreur est de **8,5 %**.

Pour la deuxième expérience, nous subdivisons les deux classes en 5 sous-classes : un instrument ou un chanteur pour la classe monophonie ; plusieurs instruments, plusieurs chanteurs, ou instrument(s) et chanteur(s) pour la classe polyphonie. Cinq lois de Weibull sont estimées. La vraisemblance est calculée sur chaque modèle, et le résultat est attribué à la classe contenant la sous-classe qui maximise la vraisemblance. Chaque modèle est appris avec 25 s de signal, soit 2500 couples. Ce découpage permet un gain relatif de 25 % : le taux d'erreur est de **6,3 %**.

4. Conclusion et perspectives

Dans cet article, nous avons décrit une méthode de séparation monophonie / polyphonie. La modélisation des paramètres étant faite avec des lois de Weibull bivariées, nous avons présenté une méthode d'estimation des paramètres par la méthode des moments. Les résultats sont très bons : 6,3 % d'erreur.

Nous envisageons d'utiliser cette méthode dans deux applications : d'une part, comme prétraitement afin d'améliorer notre système de détection de la voix chantée [LAC 07] ; d'autre part, nous aimerions élargir cette méthode à d'autres classes de sons. Nos travaux actuels se penchent sur la résolution du problème des zones de parole, qui comportent des zones non voisées (sans fréquence fondamentale) et donc classées comme polyphoniques.

5. Bibliographie

[CHE 02] DE CHEVEIGNÉ A., KAWAHARA H., YIN, a fundamental frequency estimator for speech and music, *Journal of the Acoustical Society of America*, vol. 111, n^o 4, 2002, p. 1917-1930.

[HOU 86] HOUGAARD P., A class of multivariate failure time distributions, *Biometrika*, vol. 73, n^o 3, 1986, p. 671-678.

[LAC 07] LACHAMBRE H., ANDRÉ-OBRECHT R., PINQUIER J., Singing voice characterization for audio indexing, *15th European Signal Processing Conference (EUSIPCO)*, 2007, p. 1563-1540.

[LU 90] LU J., BHATTACHARYYA G., Some new constructions of bivariate Weibull models, *Annals of Institute of Statistical Mathematics*, vol. 42, n^o 3, 1990, p. 543-559.

[SCH 97] SCHEIRER E., SLANEY M., Construction and Evaluation of a Robust Multifeature Speech/Music Discriminator, *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 1997.

[SMI 07] SMIT C., ELLIS D., Solo voice detection via optimal cancelation, *IEEE WASPAA*, 2007.

Vers une discrétisation locale pour les treillis dichotomiques

N. Girard, K. Bertet et M. Visani

L3i - Laboratoire Informatique, Image et Interaction
Pôle Sciences et Technologie Avenue Michel Crépeau
17042 La Rochelle Cedex 1
{ngirar02, kbertet, mvisani}@univ-lr.fr

RÉSUMÉ. *Dans cet article, nous rappelons la méthode de classification supervisée Navigala, que nous avons développée pour de la reconnaissance de symboles détériorés. Elle repose sur une navigation dans un treillis de Galois similaire à une navigation dans un arbre de décision. Les treillis manipulés par Navigala sont des treillis dits dichotomiques, dont nous décrivons dans ce papier les propriétés et les liens structurels avec les arbres de décision. La construction du treillis de Galois oblige à une étape préalable de discrétisation des données continues (discrétisation globale), ce qui n'est généralement pas le cas de l'arbre de décision qui procède à cette discrétisation au cours de sa construction (discrétisation locale). Utilisée comme prétraitement, la discrétisation détermine les concepts et la taille du treillis, lorsque l'algorithme de génération est directement appliqué sur ces données discrétisées. Nous proposons donc un algorithme de discrétisation locale pour la construction du treillis dichotomique ce qui pourrait nous permettre de mettre en œuvre une méthode d'élagage en cours de génération et ainsi d'améliorer les performances du treillis et éviter le sur-apprentissage.*

MOTS-CLÉS : *Treillis de Galois ; treillis dichotomique ; classification ; arbre de classification ; reconnaissance de symboles bruités.*

1. Introduction

La reconnaissance d'objets dans des images repose généralement sur deux étapes principales : l'extraction de signatures et la classification supervisée. Nous nous intéressons à la partie classification supervisée de ce processus. Parmi les nombreuses approches de la littérature, les approches symboliques offrent de la lisibilité et présentent l'avantage d'être intuitives, ce qui permet une meilleure compréhension des données. Nous nous intéressons aux deux méthodes symboliques que sont l'arbre de décision et le treillis de Galois lorsqu'il est utilisé comme classifieur. Le treillis de Galois ou treillis des concepts utilisé depuis une vingtaine d'année en classification supervisée donne des résultats comparables aux méthodes standards. C'est un graphe dont les nœuds sont des concepts. En classification, il existe de nombreuses méthodes utilisant le treillis de Galois, la plupart pour sélectionner des concepts les plus pertinents pour la tâche de classification (généralement menée à l'aide d'un classifieur tel que les K-PPV ou le classifieur Bayésien) [MEP 05, OOS 88, SAH 95]. Nous avons développé la méthode Navigala [GUI 07] qui utilise quant à elle la structure complète du treillis pour reconnaître des images détériorées de symboles par navigation dans son diagramme de Hasse [1] à partir de la racine et de manière similaire à l'arbre de décision. Différemment de l'arbre, le treillis propose plusieurs chemins vers un concept donné, ce qui lui confère une meilleure robustesse vis-à-vis du bruit [GUI 06]. De par sa construction à partir de données continues nécessitant une discrétisation, Navigala manipule des treillis dits *dichotomiques* qui sont structurellement proches des arbres de décision. Tandis que pour l'arbre [BRE 84, QUI 86, RAK 05], la discrétisation des données s'effectue le plus souvent au fur et à mesure de la construction (discrétisation locale), la construction du treillis nécessite généralement une phase préalable de discrétisation (discrétisation globale), qui détermine complètement les concepts et la taille du treillis, lorsque l'algorithme de génération est directement appliqué sur ces données discrétisées. Dans cet article, nous proposons un algorithme de discrétisation locale, menée au fur et à mesure de

1. Le diagramme de Hasse du treillis de Galois est le graphe de sa réduction réflexive et transitive.

la construction du treillis, ce qui pourrait nous permettre de mettre en œuvre une méthode d'élagage en cours de génération et ainsi d'améliorer les performances du treillis et éviter le sur-apprentissage.
Ce papier est organisé comme suit. Dans la partie 2, nous décrivons la méthode Navigala ainsi que les treillis dichotomiques et leurs liens structurels avec les arbres de décision. Dans la partie 3 nous présentons notre algorithme.

2. Contexte

2.1. *La méthode Navigala*

La méthode de classification supervisée Navigala a été développée pour la reconnaissance d'images de symboles issus de documents techniques. Cependant, il s'agit d'une méthode pouvant être utilisée dans un contexte plus large de classification supervisée d'objets décrits par des vecteurs numériques de taille fixe. On y retrouve les trois étapes classiques de *préparation des données, d'apprentissage supervisé et de classement de nouveaux exemples*.
Les signatures extraites des images sont des vecteurs numériques (ie à chaque image correspond un vecteur de caractéristiques numériques) de *taille fixée*. Ces vecteurs sont stockés dans une table de données regroupant les images, leurs caractéristiques et leur classe. La préparation des données consiste en une discrétisation des données continues qui permet de créer des intervalles disjoints de valeurs. La table discrète est ensuite binarisée. Ainsi pour une caractéristique donnée issue de la signature, chaque objet n'est associé qu'à un seul intervalle (appelé attribut) issu de cette caractéristique. Le *critère d'arrêt* de la discrétisation est la séparation des classes, chaque classe se différenciant alors des autres par au moins un attribut (sauf dans le cas où les signatures de deux objets appartenant à des classes différentes sont identiques, cette séparation ne pouvant alors évidemment pas être atteinte). Les caractéristiques non discrétisées au cours du traitement ne sont pas intégrées dans la table binaire, lui conférant ainsi une propriété de réduction encore appelée *sélection de caractéristiques*. Cette discrétisation se faisant en amont de la construction du classifieur, il s'agit d'une *discrétisation globale*. Dans le cas de Navigala, trois *critères de coupe* supervisés ou non ont été testés : la distance maximum, l'entropie, le critère de Hotelling [GUI 07]. Les expérimentations ont montré que le critère de Hotelling était le plus efficace dans notre contexte applicatif où la dispersion à l'intérieur des classes peut être importante (présence de bruit). La table binaire obtenue à l'issue de la phase de discrétisation des données continues est appelée *contexte*. Le contexte se définit par un triplet $(O, I, (f, g))$ où O est un ensemble d'objets, I est un ensemble d'attributs et (f, g) est une correspondance de Galois [2] entre objets et attributs. Le treillis de Galois est constitué d'un ensemble K de *concepts formels* muni d'une *relation d'ordre* [3] $\leq$:

- Un *concept formel* est un sous-ensemble maximal d'objets associés à un même sous-ensemble maximal d'attributs : $\forall A \in O$ et $\forall B \in I$ le couple (A, B) est un concept formel $\Leftrightarrow f(A) = B$ et $g(B) = A$.
- La *relation d'ordre* $\leq$ est définie pour deux concepts (A_1, B_1), (A_2, B_2) par :
  $(A_1, B_1) \leq (A_2, B_2) \Leftrightarrow A_1 \supseteq A_2 \Leftrightarrow B_1 \subseteq B_2$.

Ainsi défini, le *treillis de Galois* $(K, \leq)$ possède la propriété de treillis, c'est à dire que pour deux concepts de K, il existe un unique plus petit successeur commun et un unique plus grand prédécesseur commun. Il possède donc un *concept minimum* noté $\perp = (O, f(O))$ et un *concept maximum* noté $\top = (g(I), I)$. On distingue les *concepts finaux* qui sont les concepts vérifiant un critère de pureté concernant les classes des objets qui les composent et qui sont étiquetés par la classe majoritaire parmi ces objets. Notons que plusieurs concepts finaux peuvent être associés à la même classe.
Le classement d'un nouvel exemple se fait par navigation dans le diagramme de Hasse du treillis, à partir du concept minimum et jusqu'à un concept final par validation d'intervalles, comme lors de la navigation dans un arbre de décision. Le nouvel objet sera donc classé dans la classe-étiquette du concept final ainsi atteint. L'avantage de la structure de treillis par rapport à l'arbre de classification est la multiplicité des chemins menant à un même concept final, ce qui lui confère une meilleure robustesse vis à vis du bruit.

2. f associe à un ensemble d'objets leurs attributs communs, g associe à un ensemble d'attributs les objets qui possèdent ces attributs
3. Une relation d'ordre est une relation transitive, antisymétrique et réflexive

2.2. *Les treillis dichotomiques et les arbres de décision*

Lorsque chaque objet est décrit par un vecteur de caractéristiques, la phase de discrétisation permet d'obtenir une table binaire vérifiant une propriété d'exclusivité mutuelle entre attributs. En particulier, les intervalles créés lors de la discrétisation d'une même caractéristique sont disjoints, donc mutuellement exclusifs. A partir de cette propriété de la table binaire nous définissons le treillis dichotomique associé à cette table.

Définition 1 *Un treillis est dit* dichotomique *lorsqu'il est défini pour une table où il est toujours possible d'associer à un attribut binaire x un ensemble non vide $\overline{X}$ d'attributs binaires (avec $x \notin \overline{X}$) tel que les attributs de $\{x\} \cup \overline{X}$ soient mutuellement exclusifs.*

Comme cela a été démontré dans [GUI 08], les treillis dichotomiques sont sup-*pseudo-complémentés*[4] alors que les treillis sup-pseudo-complémentés ne sont pas toujours dichotomiques. De plus, lorsque chaque objet est associé à un vecteur de caractéristiques de même longueur, comme dans Navigala, les treillis dichotomiques possèdent la propriété de *co-atomisticité*[5]. Les treillis dichotomiques possèdent des liens structurels forts avec les arbres de décision :

1. Tout arbre de décision est inclus dans le treillis dichotomique, lorsque ces deux structures sont construites à partir des mêmes attributs binaires.

2. Tout treillis dichotomique est la fusion de tous les arbres de décision possibles lorsque ces structures sont construites à partir des mêmes attributs binaires.

Ainsi, lorsque les arbres de décision et le treillis de Galois sont définis à partir de la même table discrétisée, ils ont des liens structurels forts. Mais généralement, la discrétisation globale de la table des données n'est nécessaire que pour la construction du treillis de Galois, car la plupart des arbres de décision procèdent à la discrétisation des données au cours de leur construction (*discrétisation locale*). La discrétisation locale consiste à choisir en chaque nœud la segmentation optimale localement, qui permettra de discriminer au mieux les objets du nœud courant selon leur classe. Nous proposons dans la partie 3 un algorithme de construction du treillis à partir de données continues par discrétisation locale, ce qui pourrait nous permettre de mettre en œuvre une méthode d'élagage en cours de génération et ainsi d'améliorer les performances du treillis et éviter le sur-apprentissage.

3. Algorithme de discrétisation locale

Nous proposons de façon similaire à l'arbre de décision l'algorithme 1 de discrétisation locale pour la construction d'un treillis de Galois à la fois dichotomique et co-atomistique (ie. issu de données décrites par des vecteurs numériques de même longueur).
L'étape d'initialisation génère, pour chaque caractéristique de la table, un intervalle (appelé attribut) contenant l'ensemble des valeurs observées. La table discrète ainsi composée d'intervalles est binarisée. On initialise l'ensemble des concepts finaux CF avec le concept minimum $\perp$.
Comme pour la division d'un nœud ne vérifiant pas le critère d'arrêt dans l'arbre de décision, nous sélectionnons parmi les attributs B_i des co-atomes (A_i,B_i) de CF ne vérifiant pas le critère d'arrêt S, l'intervalle I et son point de coupe c_i selon le critère de coupe C (C pouvant être par exemple le critère de Hotelling). Ceci nous permet de segmenter I pour obtenir I_1 et I_2 deux intervalles disjoints, puis de remplacer dans la table de données I par I_1 et I_2. L'ensemble CF des concepts finaux du treillis associé à T est ensuite calculé pour pouvoir réitérer ce processus jusqu'à ce que tous les concepts finaux contenus dans CF vérifient un critère d'arrêt S similaire à celui qui peut être mis en œuvre pour les arbres (généralement mesure de pureté ou nombre minimum d'objets dans chacun des concepts de CF). Le treillis étant co-atomistique, les concepts finaux sont les co-atomes et s'obtiennent en calculant les prédécesseurs immédiats du concept maximum $\top$ avec par exemple une adaptation de la fonction successeurs immédiats de Bordat. Il n'est donc pas nécessaire de calculer le treillis à chaque itération.

4. Un treillis est sup-pseudo-complémentés lorsque pour tout concept (A, B), il existe toujours un *concept complémentaire* (A', B') tel que : $(A, B) \vee (A', B') = \top = (\emptyset, I)$

5. les concepts finaux sont des co-atomes, les co-atomes d'un treillis sont les concepts dont le plus petit successeur commun est l'élément maximum $\top$

Algorithme 1 : Construction d'un treillis par discrétisation locale

Entrées :
- Ensemble de données $(O_i, V_{ij})_{i\in\{1...n\},j\in\{1...p\}}$, chacun des n objets O_i étant décrit par un vecteur de p caractéristiques $(V_{ij})_{j\in\{1...p\}}$
- $\mathcal{S}$: critère d'arrêt et $\mathcal{C}$: critère de coupe

Sorties : TG : treillis de Galois de la table discrétisée

Initialiser une table binaire T avec :
- sur chaque ligne : un objet O_i
- sur chaque colonne : l'intervalle I_j contenant toutes les valeurs observées d'une caractéristique V_j. Chacun des intervalles I_j devient alors un attribut binaire, partagé par tous les objets O_i

Initialiser CF avec $\perp$;

tant que $\exists(A, B) \in CF$ *tel que* $!S((A, B))$ **faire**

Sélectionner selon $\mathcal{C}$ le point de coupe optimum c_{I^*} associé à un intervalle optimum I^* parmi les attributs B_k des concepts $(A_k, B_k) \in CF$ tel que $!S((A_k, B_k))$;
Découper I^* en deux intervalles disjoints I_1 et I_2 selon c_{I^*} ;
Remplacer I^* par I_1 et I_2 dans T, puis mettre à jour T en conséquence ;
CF = co-atomes du treillis associé à T = prédécesseurs immédiat du concept maximum ;

Calculer le treillis TG de T ; **retourner** TG ;

4. Conclusion et Perspectives

Après avoir introduit le contexte de cette étude et décrit la méthode Navigala que nous avons développée, cet article présente les treillis dichotomiques, leurs propriétés et leurs liens structurels forts avec les arbres de décision. Puis un algorithme de discrétisation locale pour la construction du treillis de Galois est proposé. Cet algorithme est inspiré du traitement des données continues lors de la construction de la plupart des arbres de décision. Avec cet algorithme nous construisons un treillis de Galois complet avec une discrétisation locale. Nous sommes en train de mettre en œuvre un protocole expérimental associé à cet algorithme, et espérons avoir rapidement des résultats expérimentaux plus poussés. Une première perspective serait de procéder à l'élagage en cours de génération du treillis ainsi construit pour en améliorer les performances. Une deuxième perspective consisterait en une génération incrémentale de l'ensemble des co-atomes.

5. Bibliographie

[BRE 84] BREIMAN L., FRIEDMAN J., OLSHEN R., STONE C., *Classification and regression trees*, Wadsworth Inc., 1984.

[GUI 06] GUILLAS S., BERTET K., OGIER J.-M., A Generic Description of the Concept Lattices' Classifier : Application to Symbol Recognition, vol. 3926, 2006, p. 47-60, Lecture Notes in Computer Science, Revised and extended version of paper first presented at Sixth IAPR International Workshop on Graphics Recognition (GREC'05).

[GUI 07] GUILLAS S., Reconnaissance d'objets graphiques détériorés : approche fondée sur un treillis de Galois, Thèse de doctorat, Université de La Rochelle, 2007.

[GUI 08] GUILLAS S., BERTET K., VISANI M., OGIER J.-M., GIRARD N., Some Links Between Decision Tree and Dichotomic Lattice, *Proceedings of the Sixth International Conference on Concept Lattices and Their Applications*, CLA 2008, October 2008, p. 193-205.

[MEP 05] MEPHU-NGUIFO E., NJIWOUA P., Treillis des concepts et classification supervisée, *Technique et Science Informatiques, RSTI*, vol. 24, n^o 4, 2005, p. 449-488, Hermès - Lavoisier, Paris, France.

[OOS 88] OOSTHUIZEN G., The use of a Lattice in Knowledge Processing, PhD thesis, University of Strathclyde, Glasgow, 1988.

[QUI 86] QUINLAN J., Induction of Decision Trees, *Machine Learning*, vol. 1, 1986.

[RAK 05] RAKOTOMALALA R., Arbres de Décision, *Revue MODULAD*, vol. 33, 2005.

[SAH 95] SAHAMI M., Learning Classification Rules Using Lattices, LAVRAC N., WROBEL S., Eds., *Proceedings of European Conference on Machine Learning, ECML'95*, Heraclion, Crete, Greece, April 1995, p. 343-346.

Combiner treillis de Galois et analyse factorielle multiple pour l'analyse de traits biologiques

Aurélie Bertaux[1,2], **Florence Le Ber**[1,3], **Pulu Li**[1], **Michèle Trémolières**[1]

1. LHyGeS UMR 7517, ENGEES, 1 quai Koch BP 61039 F 67070 Strasbourg cedex
2. LSIIT UMR 7005, Bd Sébastien Brant BP 10413 F 67412 Illkirch cedex
3. LORIA UMR 7503, BP 35 F 54506 Vandœuvre-lès-Nancy cedex
{aurelie.bertaux, florence.leber}@engees.unistra.fr, tremolie@unistra.fr

RÉSUMÉ. *Dans le cadre de travaux sur la qualification de l'état écologique des cours d'eau, nous nous intéressons aux informations portées implicitement par les caractéristiques (ou traits) biologiques des macrophytes qui y vivent. Classiquement les hydrobiologistes employent des méthodes telles que l'analyse factorielle et la classification hiérarchique qui permettent d'établir des groupes d'espèces et de les caractériser par leurs principaux traits. Des résultats similaires peuvent être obtenus grâce aux treillis de Galois ou treillis de concepts formels, qui permettent de structurer des ensembles d'objets et leurs propriétés. Nous nous intéressons donc à étudier les combinaisons de ces deux approches et leurs apports réciproques dans le cadre de l'analyse des caractéristiques biologiques des macrophytes.*

MOTS-CLÉS : *Treillis de Galois, Analyse Factorielle Multiple, Hydrobiologie.*

1. Introduction

Afin d'établir un cadre de qualification des états écologiques des cours d'eau, les hydrobiologistes s'intéressent aux espèces vivant dans ces milieux et à leurs caractéristiques (ou traits) biologiques et écologiques. L'objectif est de déterminer des groupes d'espèces et de traits biologiques qui soient caractéristiques d'un milieu. Pour déterminer de tels groupes, les biologistes ont classiquement recours à des méthodes statistiques telles que la classification hiérarchique ou l'analyse factorielle [HÉR 07]. Cependant d'autres techniques existent telles que les treillis de Galois qui permettent de créer des concepts qui mettent en relation des ensembles d'espèces et les traits qu'ils partagent. Alors que les méthodes statistiques dégagent des tendances principales, les treillis de Galois fournissent des résultats exacts mais parfois difficiles à exploiter. Nous nous intéressons donc à combiner ces approches.

Dans cet article nous présentons et évaluons brièvement une approche combinant l'analyse factorielle multiple et les treillis de Galois. Après cette première partie introductive, nous présentons les données sur lesquelles nous travaillons qui concernent ici les hydrophytes ou macrophytes (plantes macroscopiques). Les troisième et quatrième parties décrivent les deux étapes de l'approche, la première utilisant l'analyse factorielle multiple, la seconde les treillis de Galois. Enfin nous concluons et présentons les perspectives découlant de cette approche.

2. Données

Les données auxquelles nous nous intéressons concernent 46 macrophytes vivant dans les eaux de la plaine d'Alsace. Ces plantes sont décrites par 10 *traits* biologiques tels que leur taille ou leur forme de croissance. Chacun de ces traits est divisé en plusieurs *modalités*. Par exemple la taille est décrite par 4 modalités : inférieure à 0,08 mètre, entre 0,08 et 0,3 mètre, entre 0,3 et 1 mètre, et entre 1 et 5 mètres. Chaque plante a une *affinité* pour chacune

de ces modalités, affinité exprimée par une valeur indiquant (tableau 1) : qu'il n'y a pas (R) d'individus de l'espèce concernée qui possède cette modalité ; qu'il y en a quelques uns (U) ; qu'il y en a beaucoup (B).

Pour être manipulées *via* les algorithmes classiques que nous utilisons ici, ces données sont préalablement converties en un tableau disjonctif total. Ainsi chaque plante possède ou non un triplet (trait, modalité, affinité). Nous explorons par ailleurs d'autres approches permettant d'éviter la discrétisation [BER 09].

TAB. 1. *Données sur les traits biologiques des macrophytes et conversion en tableau disjonctif total.*

TRAITS	Taille potentielle				Taille potentielle					
MODALITÉS	<0.08m	0.08-0.3 m	0.3 -1m	1-5 m						
	1	2	3	4	t1-U	t2-U	t2-B	t3-B	t4-U	t4-B
ALIP	U	U	B	U	1	1	0	1	1	0
BERE	U	B	B	U	1	0	1	1	1	0
CALO	U	U	B	B	1	1	0	1	0	1
CALP	U	B	B	U	1	0	1	1	1	0
CARA	U	U	B	U	1	1	0	1	1	0
CERD	U	U	B	B	1	1	0	1	0	1

3. Approche statistique

Notre objectif est de déterminer des propriétés environnementales caractéristiques de groupes de plantes et de leurs traits biologiques. Pour cela nous nous intéressons à établir ces groupes en répartissant les espèces de macrophytes selon leurs affinités aux différentes modalités des traits biologiques. Nous présentons ici la partie statistique de notre approche nous permettant de déterminer ces groupes. Nous utilisons pour cela l'analyse factorielle multiple (AFM) [ESC 88]. En particulier, cette méthode permet de comparer k groupes de variables définies sur le même ensemble d'individus.

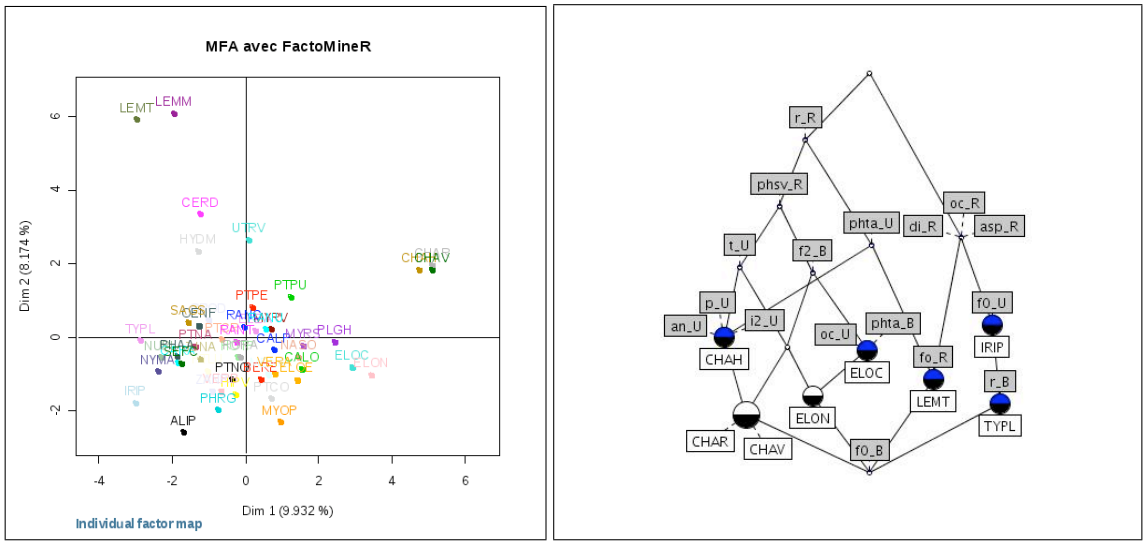
Dans notre cas, les variables sont les 10 traits biologiques et les individus sont les macrophytes. La figure 1(a) présente les deux premiers axes obtenus par l'AFM. Nous constatons que sur l'axe 1 sont représentés principalement les plantes CHAR, CHAV, CHAH (characées), ELOC et ELON (élodées) en positif et TYPL (massette), IRIP (iris), LEMT (lentille d'eau) en négatif. Elles sont associées aux traits p_U (quelques individus de l'espèce ont des organes pérennes), t_U (quelques individus ont une reproduction végétative par turions, bulbilles ou apex), an_U (quelques individus sont annuels), etc., en positif. En négatif, sont associés les traits $f0\_U$, $f0\_R$ (quelques individus – ou pas du tout – ont une faible flexibilité), etc. Par la suite on travaille avec les 8 macrophytes et les 17 traits qui contribuent le plus à ce premier axe de l'AFM.

4. Treillis de concepts formels

Les treillis de concepts formels ou treillis de Galois [BAR 70, GAN 97] sont des structures arborescentes permettant de hiérarchiser des concepts selon une relation de subsomption. Un concept est un couple constitué d'une extension, i.e. un ensemble O d'objets, ici des macrophytes, et d'une intension, i.e. un ensemble P de propriétés, ici les triplets (trait biologique, modalité, affinité). Il s'agit d'ensembles fermés maximaux au sens où tous les objets de O possèdent en commun toutes les propriétés de P et seulement celles-là. Le treillis contient tous les fermés maximaux et permet d'explorer les implications et associations entre propriétés [GUI 86].

À partir des données dont nous disposons (tableau 1), nous pouvons constituer directement un treillis. Cependant un treillis portant sur la totalité des données contient 1401 concepts, ce qui est difficilement exploitable [BER 09]. Ce problème est souvent soulevé, notamment dans [HER 00] où est proposé un calcul de χ^2 pour choisir les couples d'attributs multivalués sur lesquels construire le treillis. Afin de ne pas nous limiter à deux variables tout en réduisant la dimension des données, nous utilisons les résultats obtenus par l'AFM pour sélectionner des

FIG. 1. *Résultats de l'AFM et treillis associé sur les données du tableau 1.*



(a) Les deux premiers axes de l'AFM

(b) Le treillis associé au premier axe

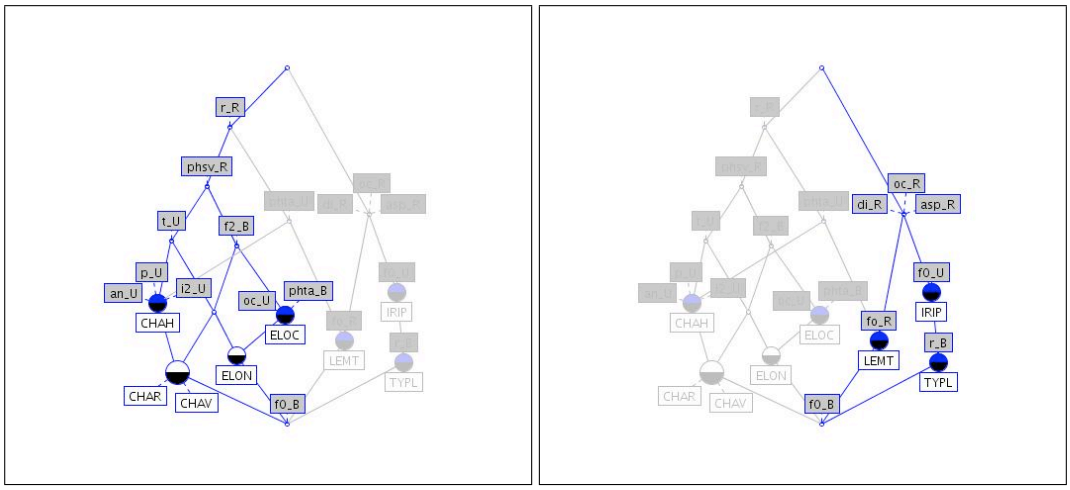
sous-ensembles de plantes ou traits *a priori* intéressants. En effet, chaque axe de l'AFM fournit des groupes de plantes associées à des ensembles de traits particulièrement corrélés. Ainsi le treillis présenté sur la figure 1(b) a été obtenu à partir des 8 plantes (CHAR, CHAV, TYPL, IRIP, etc. : étiquettes claires) et des 17 triplets (trait, modalité, affinité) ($p\_U, an\_U, f0\_U, f0\_R$, etc. : étiquettes sombres) distingués par le premier axe de l'AFM. Ici on ne prend donc pas toutes les propriétés des plantes considérées, ni toutes les plantes possédant les triplets considérés. Un concept (représenté par un nœud du graphe de la figure 1(b)) regroupe un ensemble de plantes « localement semblables » pour les traits biologiques qu'elles partagent.

Tout d'abord, on remarque que le treillis *sépare* les plantes à coordonnée négative et les plantes à coordonnée positive sur l'axe 1 de l'AFM. Plus précisément on peut visualiser les propriétés partagées ou non par les plantes positives ou négatives[1] : par exemple, côté négatif, les propriétés di_R, asp_R et oc_R (pas de dispersion intermédiaire, pas de reproduction d'août à octobre) sont partagées par les trois plantes TYPL, LEMT, IRIP, tandis que r_B (beaucoup de reproduction par rhizomes ou stolons) est spécifique à TYPL (figure 2(b)) ; côté positif, on voit que les trois plantes CHAH, CHAR et CHAV ont plusieurs propriétés communes et se distinguent des deux plantes ELOC et ELON (figure 2(a)). On remarque également que certaines propriétés comme r_R (pas de reproduction par rhizomes ou stolons) et $phta\_U$ (quelques individus sont présents sur site toute l'année) sont possédées à la fois par des plantes à coordonnée positive et des plantes à coordonnée négative. À l'inverse la propriété $f0\_B$ (beaucoup d'individus peu flexibles) n'est associée à aucune des huit plantes retenues : il faudrait donc étendre cet ensemble.

Considérons maintenant les concepts individuellement. Par exemple, on voit que les propriétés $f0\_U$ et $f0\_R$, concernant la fléxibilité des plantes, sont présentes dans deux concepts distincts, respectivement ($\{f0\_U\}$,{IRIP, TYPL}) et ($\{f0\_U\}$,{LEMT}). Le treillis permet donc de distinguer les individus et propriétés associés sur l'axe 1 de l'AFM. Considérons également le concept étiqueté CHAH. Son extension contient les plantes CHAH, CHAV et CHAR et son intension les propriétés $an\_U, p\_U, i2\_U, t\_U, phsv\_R$ et r_R (figure 2(a)). Ceci traduit en particulier que ces espèces comptent des individus annuels et des individus avec organes pérennes et qui se reproduisent par turions, bulbilles ou apex. Une majorité de ces propriétés ne sont pas possédées par les plantes ELOC et ELON comme on peut le voir aussi en examinant l'axe 2 de l'AFM.

1. Les concepts héritent des propriétés (intension) de leurs pères et des plantes (extension) de leurs fils.

FIG. 2. *Treillis de Galois associé à l'axe 1 de l'AFM : visualisation des plantes positives et négatives et de leurs propriétés.*



(a) Plantes ayant une coordonnée positive sur l'axe 1 (b) Plantes ayant une coordonnée négative sur l'axe 1

5. Conclusion et travaux futurs

Le travail engagé s'attache à combiner les approches statistiques et les treillis de Galois afin d'aider à la déterminination de groupes de plantes et de traits biologiques caractéristiques de leur milieu. Il s'agit d'un travail de fouille de données où les apports de chaque méthode sont examinés par rapport à différents aspects du problème. Nous avons présenté ici quelques éléments concernant la combinaison des treillis de Galois et de l'AFM. Nous étudions sur les mêmes données la combinaison de la classification ascendante hiérarchique et des treillis.

De cette première expérience se dégage un double intérêt : d'une part, grâce aux treillis, une meilleure visualisation et exploration des résultats de l'AF ; d'autre part, grâce à l'AFM (ou d'autres méthodes statistiques), une segmentation des données qui permet de construire des treillis plus petits et mieux focalisés. Différents niveaux de complexité peuvent être explorés, en fonction du nombre d'individus ou propriétés retenus sur les axes de l'AFM ou en fusionnant les informations de plusieurs axes.

Cette approche est actuellement en cours d'application et nous comptons approfondir la comparaison des approches statistiques et des treillis d'un point de vue plus théorique.

6. Bibliographie

[BAR 70] BARBUT M., MONJARDET B., *Ordre et classification – Algèbre et combinatoire*, Hachette, Paris, 1970.

[BER 09] BERTAUX A., LE BER F., BRAUD A., TRÉMOLIÈRES M., Mining Complex Hydrobiological Data with Galois Lattices, *International Journal of Computing and Information Sciences*, 2009, à paraître.

[ESC 88] ESCOFIER B., PAGÈS J., *Analyses factorielles simples et multiples*, Dunod, Paris, 1988.

[GAN 97] GANTER B., WILLE R., *Formal Concept Analysis : Mathematical Foundations*, Springer Verlag, 1997.

[GUI 86] GUIGUES J.-L., DUQUENNE V., Familles minimales d'implications informatives résultant d'un tableau de données binaires, *Mathématiques et Sciences Humaines*, vol. 95, 1986, p. 5–18.

[HER 00] HERETH J., STUMME G., WILLE R., PRISS U., Conceptual Knowledge Discovery and Data Analysis, *ICCS '00 : Proceedings of the Linguistic on Conceptual Structures*, London, UK, 2000, Springer-Verlag, p. 421–437.

[HÉR 07] HÉRAULT B., HONNAY O., Using life-history traits to achieve a functional classification of habitats, *Applied Vegetation Science*, vol. 10, 2007, p. 73–80.

An approach based on Formal Concept Analysis for mining numerical data

Zainab Assaghir, Mehdi Kaytoue, Nizar Messai, and Amedeo Napoli

LORIA, équipe Orpailleur, bâtiment B, B.P. 70239, F-54506 Vandoeuvre les Nancy
email : {assaghiz, kaytouem, messai, napoli}@loria.fr

RÉSUMÉ. *In this paper, we present a method based on Formal Concept Analysis (FCA) for mining numerical data. An extension of standard FCA Galois connection allows takes into account many-valued (MV) numerical attributes and "similarity" between numerical values. This leads to the definition of MV formal concepts and MV concept lattices. Depending on a similarity threshold, MV concept lattices have different levels of precision. Accordingly, multi-level classification tasks can be achieved such as knowledge discovery and knowledge representation, navigation, information retrieval, and data mining.*

MOTS-CLÉS : *Formal Concept Analysis, classification, many-valued concept, similarity*

1. Introduction

Formal concept analysis (FCA) [GAN 99] allows to analyze data by building a *concept lattice* from a *formal context*, i.e. a binary table whose rows correspond to objects and columns correspond to attributes. Relying on attribute sharing between objects, data are classified into *formal concepts* which are partially ordered within a concept lattice. The concept lattice provides a support for organization of data and navigation within a data set. FCA can be considered as a full classification process. Is is used in a wide range of applications, e.g. knowledge representation and knowledge discovery (data mining), information retrieval, and software engineering.

Whereas the standard FCA process requires a binary representation of data, i.e. a formal context, real-world data sets are often complex and heterogeneous. Their representation in terms of a binary table does not produce a formal context, but instead a *many-valued* (MV) context where a table entry (i) is empty when there is no relation between the corresponding object and attribute, (ii) contain an arbitrary value taken by the attribute for the object. Such a value is sometimes considered as a truth degree depending on the link between the object and the attribute. Following the FCA formalism, any operation on data requires the transformation of an MV context into a binary context by applying an appropriate *conceptual scaling* [GAN 99]. Scaling is most of the time arbitrary and may lead to errors or loss of information. By contrast, this paper presents an approach for analyzing a data set represented as an MV context, where the values of attributes are numerical and where no scaling is necessary. Firstly, a Galois connection adapted to an MV context and based on similarity between attribute values is defined for computing MV concepts and MV concept lattices. Then, the principles of the classification process for building an MV concept lattice are detailed : basically, they extend the standard process to deal with MV attributes in a straightforward way. The resulting lattice can be used to represent and organizes the data as they are initially given, without any transformation process. This is the strength and originality of the paper.

The paper is organized as follows. Section 2 recalls basics of FCA and introduces MV contexts. Section 3 proposes a formalization of attribute sharing in an MV context, and describes an MV Galois connection and its derived structures, i.e. MV concepts and MV concept lattices. Finally, a discussion and perspectives conclude the paper.

Attr. / Obj.	Diameter (10^3km)			Mass (10^{21}t)			Distance to Sun (10^6km)			Satellite
	D<10	10<D<20	D>20	M<1	4<M<5	M>5	DS<100	100<DS<200	DS>200	
Mercury			x	x			x			
Venus		x			x			x		
Earth		x				x		x		x
Mars	x			x					x	x

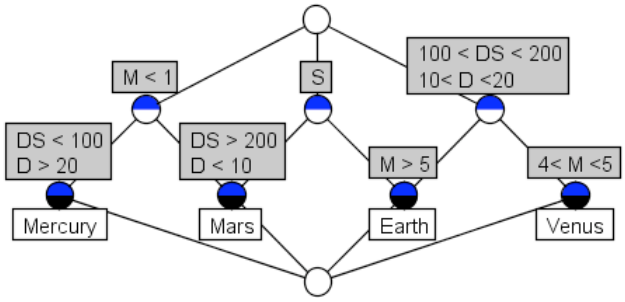


FIG. 1. *A formal context and the associated concept lattice.*

2. Formal Concept Analysis

2.1. *Basics of FCA*

A formal context is denoted by $\mathbb{K} = (G, M, I)$ where G is a set of objects, M is a set of attributes, and I is a binary relation between G and M ($I \subseteq G \times M$) [GAN 99]. $(g, m) \in I$ denotes the fact that an object $g \in G$ is in relation through I with an attribute $m \in M$ (also read as g *has* m). A concept is represented by a pair (A, B) such that $A \subseteq G$, $B \subseteq M$, $A' = B$, and $B' = A$ where $A' = \{m \in M \mid (g, m) \in I \; \forall g \in A\}$, i.e. the set of attributes common to all objects in A, and $B' = \{g \in G \mid (g, m) \in I \; \forall m \in B\}$, i.e. the set of objects which have all attributes in B. A and B are respectively called the $extent$ and the $intent$ of the concept (A, B). The derivation operators $' : \mathfrak{P}(G) \rightarrow \mathfrak{P}(M)$ and $' : \mathfrak{P}(M) \rightarrow \mathfrak{P}(G)$ form a Galois connection between the powerset lattices $\mathfrak{P}(G)$ and $\mathfrak{P}(M)$. The set of all concepts in a formal context is denoted by $\mathfrak{B}(G, M, I)$. A concept $(A1, B1)$ is a sub-concept of $(A2, B2)$ when $A1 \subseteq A2$ (or equivalently $B2 \subseteq B1$) and we write $(A1, B1) \leq (A2, B2)$. The set of concepts in $\mathfrak{B}(G, M, I)$ ordered by the partial order "$\leq$" forms the concept lattice of the context (G, M, I) denoted by $\underline{\mathfrak{B}}(G, M, I)$. An example of formal context and the associated concept lattice is given in Figure 1.

2.2. *Many-valued contexts*

In real-world data sets, objects are described by *many-valued* (MV) attributes, i.e. attributes taking more than one value. Examples of MV attributes are *color* (nominal), *weight* (numerical), and *height* (numerical), etc. The tabular representation of such data sets results in a *many-valued context* (MV context) denoted by (G, M, W, I), where G is a set of objects, M is a set of attributes, W is a set of attribute values, and I is a ternary relation between G, M and W (i.e., $I \subseteq G \times M \times W$). $(g, m, w) \in I$ denotes the fact that "*the attribute* m *takes the value* w *for the object* g" (also denoted by $m(g) = w$).

There are two main approaches for dealing with MV contexts following FCA. The first one consists in transforming an MV context into an ordinary or *scaled* formal context [GAN 99]. The second approach introduces a particular form of MV context, namely *fuzzy context* i.e. an MV context where the attribute values stands for *truth values* of the statement "the object g has the attribute m" [BEL 02]. In the following, we propose an original approach to process an MV context based on an adapted Galois connection taking into account similarities between attribute values. However, a restriction is made to attribute values supposed to be *totally ordered*. Attribute can be either numerical, where values are numbers, interval data, or nominal, where sets of values are totally ordered, e.g. *size* with the set of values {*small, medium, large*}. In this way, this work can be related to Symbolic Data Analysis (SDA) where descriptions of complex symbolic objects are classified [BOC 00].

In this paper, we only consider numerical attributes. The manipulation of other forms of MV attributes is discussed in [MES 09].

3. MV Galois connection, MV concepts, and MV concept lattices

In standard FCA, given a formal context (G, M, I), an attribute m is shared by a set A of objects, i.e. $m \in A'$, if and only if each object g in A has the attribute m, i.e. $(g, m) \in I$. This statement is adapted for taking into account (numerical) MV attributes and the similarity existing between different attribute values :

Definition 1 *Given an MV context (G, M, W, I) and a threshold $\theta \in [0, 1]$:*

- *Two values w_i and w_j of a (numerical) attribute m are* similar *if and only if $|w_i - w_j| \leq \theta$.*
- *Two objects g_i and g_j in G* share *an attribute m in M if and only if $m(g_i) = w_i$ and $m(g_j) = w_j$ are similar i.e. $|w_i - w_j| \leq \theta$. Assuming that $w_i \leq w_j$, g_i and g_j share $m_{[w_i,w_j]}$ and the interval $[w_i, w_j]$ is the similarity interval of m for g_i and g_j.*
- *All objects in a set $A \subseteq G$* share *an attribute m whenever any two objects in A share m. The similarity interval of m for A is $[min_{g\in A}(m(g)), max_{g\in A}(m(g))]$ and the attribute m shared by objects in A is denoted by $m_{[min_{g\in A}(m(g)),max_{g\in A}(m(g))]}$.*

For illustration, consider the MV context given in Figure 2 and the threshold $\theta = 0.2$. The objects *Mercury* and *Venus* share $DS_{[0.3,0.5]}$, i.e. planets *Mercury* and *Venus* have similar distances to the Sun (DS), and more precisely, these distances vary between the two values $0.3 * 22810^6$km and $0.5 * 22810^6$km. The threshold θ defines the maximal difference allowed between two attribute values. The choice of θ depends on the data sets and on domain and expert knowledge. It can be noticed that the choice of θ can be likened to the choice of a frequency threshold for itemsets search in a data mining process.

In an MV context, the first derivation operator associates to a set of objects the set of their common attributes satisfying the appropriate similarity interval. Dually, the second derivation operator associates to a set of attributes satisfying the appropriate similarity interval the set of all objects sharing these attributes.

Definition 2 *Given an MV context (G, M, W, I) and a threshold $\theta \in [0, 1]$:*

- *Given a set of objects $A \subseteq G$: $A'_\theta = \{m_{[\alpha,\beta]} \in M \times \mathfrak{I}_\theta$ such that $m(g) \neq \emptyset$ and $\forall g_i, g_j \in A$, $|m(g_i) - m(g_j)| \leq \theta$, $\alpha = min_{g\in A}(m(g))$, and $\beta = max_{g\in A}(m(g))\}$ (the set of attributes common to objects in A).*
- *Dually, for a set $B \subseteq M \times \mathfrak{I}_\theta$ of attributes with similarity intervals : $B'_\theta = \{g \in G$ such that $\forall m_{[\alpha,\beta]} \in B, m(g) \in [\alpha, \beta]\}$ (the set of objects sharing attributes in B), where $\mathfrak{I}_\theta$ denotes the set of all possible intervals $[\alpha, \beta]$ such that $\beta - \alpha \leq \theta$ and $[\alpha, \beta] \subseteq [0, 1]$.*

A Galois connection between object sets in G and attribute sets in $M \times \mathfrak{I}_\theta$ has to be defined between ordered sets. The first ordered set is given by $(\mathfrak{P}(G),\subseteq)$ for sets of objects in G. The second ordered set is given by $\mathfrak{P}(M \times \mathfrak{I}_\theta)$ with the following partial ordering, combining inclusion of attribute subsets and inclusion of similarity intervals : given two sets B_1 and B_2 in $\mathfrak{P}(M \times \mathfrak{I}_\theta)$: $B_1 \subseteq_\theta B_2$ if and only if $\forall m_{[\alpha_1,\beta_1]} \in B_1, \exists m_{[\alpha_2,\beta_2]} \in B_2$ such that $[\alpha_2, \beta_2] \subseteq [\alpha_1, \beta_1]$. Then, $(\mathfrak{P}(M \times \mathfrak{I}_\theta), \subseteq_\theta)$ is a partially ordered set. The two here-above derivation operators form an MV Galois connection between $(\mathfrak{P}(G), \subseteq)$ and $(\mathfrak{P}(M \times \mathfrak{I}_\theta), \subseteq_\theta)$ [MES 09].

Definition 3
- *A* many-valued concept *is a pair (A, B) where $A \subseteq G$ and $B \subseteq M \times \mathfrak{I}_\theta$ such that $A'_\theta = B$ and $B'_\theta = A$. A and B are respectively the* extent *and the* intent *of (A, B).*
- *If (A_1, B_1) and (A_2, B_2) are MV concepts, (A_1, B_1) is a* sub-concept *of (A_2, B_2) when $A_1 \subseteq A_2$ (which is equivalent to $B_2 \subseteq_\theta B_1$).*
- *The set of all MV concepts of (G, M, W, I) ordered in this way is denoted by $\underline{\mathfrak{B}}_\theta(G, M, W, I)$ and called the* many-valued concept lattice *of (G, M, W, I).*

Obj.\Attr.	D (*48.8)	M (*5.97)	DS (*2281^6)	S (*2)
Mercury	1.	0.1	0.3	
Venus	0.3	0.8	0.5	
Earth	0.3	1.	0.7	0.5
Mars	0.1	0.1	1.	1.

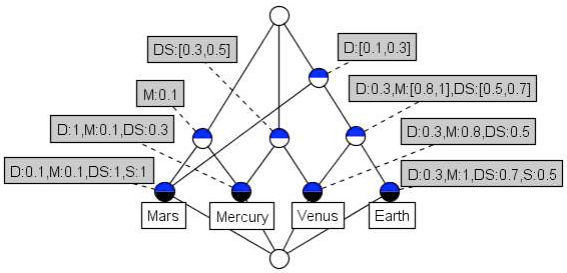


FIG. 2. *An MV context and the associated concept lattice* $\underline{\mathfrak{B}}_{0.2}(G, M, W, I)$ *for* $\theta = 0.2$.

Considering the MV context in Figure 2 and the threshold $\theta = 0.2$, $\{$*Mercury*,*Venus*$\}'_\theta = \{DS_{[0.3,0.5]}\}$ and $\{DS_{[0.3,0.5]}\}'_\theta = \{$*Mercury*,*Venus*$\}$. Then examples of MV concepts are : ($\{$*Mars*,*Venus*,*Earth*$\}$,$\{D_{[0.1,0.3]}\}$) and ($\{$*Venus*,*Earth*$\}$,$\{D_{[0.3,0.3]}$,$M_{[0.8,1]}$,$DS_{[0.5,0.7]}\}$). The MV concept lattice is given on the same Figure on right side.

An MV concept lattice provides an "exhaustive and compact" representation of data in an MV context : attribute values and similarity intervals in the intents of an MV concept represent all possible combinations of MV attributes shared by objects given a similarity threshold θ.

4. Conclusion and future work

In this paper, we introduced an extension of FCA to deal with complex numerical data represented as multi-valued contexts. We defined an MV Galois connection based on similarity between attribute values. The basic idea is that two objects share an attribute whenever the values taken by this attribute for these objects are similar (i.e. their difference is less than a threshold). This Galois connection is the basis of the computation of MV concepts and MV concept lattices. Depending on the similarity threshold, MV concept lattices can have different levels of precision, making them good candidates for multi-level classification (this is particularly useful for large contexts).

In the future, the MV Galois connection will be extended to deal with general MV contexts (including symbolic MV attributes). Similarity between attribute values can be taken from domain ontologies and thesaurus. Resulting lattices can be used in different application domains including classification, knowledge discovery and knowledge representation, information retrieval, and data mining.

5. Bibliographie

[BEL 02] BELOHLAVEK R., *Fuzzy Relational Systems : Foundations and Principles*, Series on Systems Science and Engineering (Vol. 20), Kluwer Academic/Plenum Press, New York, 2002.

[BEL 05] BELOHLAVEK R., VYCHODIL V., *Fuzzy Equational Logic*, Studies in Fuzziness and Soft Computing (Vol. 186), Springer, Berlin, 2005.

[BOC 00] BOCK H., DIDAY E., Eds., *Analysis of Symbolic Data : Exploratory Methods for Extracting Statistical Information from Complex Data*, vol. 15 de *Studies in Classification, Data Analysis, and Knowledge Organization*, Springer, February 2000.

[GAN 99] GANTER B., WILLE R., *Formal Concept Analysis*, Springer, mathematical foundations édition, 1999.

[MES 09] MESSAI N., Analyse de concepts formels guidée par des connaissances de domaine. Application à la découverte de ressources génomiques sur le Web, Thèse d'informatique, Université Henri Poincaré (Nancy 1), 2009.

Tatouages et motivations pour se faire détatouer : une étude prospective sur 151 patients vivant dans le sud de la France

Julie Latreille[1], Jean-Luc Lévy[2], Christiane Guinot[1,3]

[1] *CE.R.I.E.S, Unité de Biométrie et Epidémiologie,*
20, rue Victor Noir,
92521 Neuilly sur Seine, France
{julie.latreille, christiane.guinot@ceries-lab.com}
[2] *Centre Laser Dermatologique,*
3, boulevard Lord Duveen,
13008 Marseille, France
{dr.levy@centrelaser.fr}
[3] *Laboratoire d'Informatique,*
Université François Rabelais de Tours,
64, avenue Jean Portalis, 37200 Tours, France

RÉSUMÉ. Le nombre de personnes portant des tatouages a fortement augmenté en occident, ainsi que les demandes pour les enlever. Cette recherche a été menée pour identifier des groupes de patients présentant des caractéristiques similaires et étudier leurs motivations pour se faire détatouer. Cent cinquante et un patients ont été inclus prospectivement, chacun devant décrire les motivations de sa demande au cours d'un entretien médical. Les réponses ont été enregistrées sous forme de données textuelles dans un questionnaire informatisé qui comportait également des informations sur le patient et le tatouage. Les liens entre les caractéristiques des patients et de leur tatouage ont été analysés, puis une typologie des patients a été recherchée. Par ailleurs, les motivations ont été étudiées à l'aide d'une méthode d'analyse textuelle. Quatre types de patients ont été identifiés. Les motivations pour se faire détatouer ont été trouvées associées avec le degré de satisfaction après la réalisation du tatouage, mais aussi avec le sexe et l'âge du patient. Notre étude a montré que le tatouage pouvait devenir un problème social et être responsable de souffrances psychologiques. Il semble donc important d'avoir des campagnes d'information concernant le tatouage afin d'éviter les tatouages non désirés.

MOTS-CLÉS : algorithme K-means, analyse factorielle des correspondances, analyse textuelle, classification ascendante hiérarchique.

1. Introduction

Ces dernières années, le nombre d'individus portant des tatouages a fortement augmenté en occident et, en parallèle, le nombre de demandes pour enlever ces tatouages [STI 06 et LAU 06]. L'objectif de cette étude était d'étudier les caractéristiques et les motivations d'un large échantillon de patients voulant se faire détatouer.

2. Matériel

Cent cinquante et un patients ont été inclus prospectivement dans cette étude de novembre 2006 à juillet 2007 au Centre Laser Dermatologie de Marseille. Chaque patient devait décrire pendant un entretien médical les motivations de sa demande de détatouage. Les réponses libres ont été enregistrées sous forme de données textuelles dans un

questionnaire informatisé. Ce questionnaire comportait également des informations sur le patient (sexe, âge, statut marital…) et sur le tatouage (taille, position, couleur…).

3. Méthodes

Les liens entre les caractéristiques des patients, celles de leur tatouage ont été étudiés à l'aide d'une analyse des correspondances multiples (ACM) [JOB 92]. La formule de Benzécri a été utilisée pour calculer le % de variance expliquée par chaque composante [BEN 79]. Les composantes retenues ont été ensuite utilisées pour rechercher une typologie des patients à l'aide d'une méthode de classification ascendante hiérarchique [EVE 93]. Par ailleurs, les motivations concernant le tatouage et celles concernant le détatouage ont été résumées à l'aide d'une méthode d'analyse textuelle [LEB 88]. La procédure MOTS a permis de créer le vocabulaire initial de « mots ». L'outil interactif de modification du vocabulaire a ensuite été utilisé afin de supprimer, corriger et mettre en équivalence certains « mots ». Enfin, deux tableaux de contingence ont été créés (procédure TEXNU) avec en ligne l'identifiant « patients » (151 lignes) et en colonne les « mots » (10 motivations pour le tatouage, 12 pour le détatouage). Finalement, les groupes de patients ont été décrits avec l'ensemble des informations disponibles. Les analyses statistiques ont été réalisées à l'aide des logiciels SAS® version 9.1.3 (SAS Institute Inc., SAS Campus Drive, Cary NC 27513, USA), et SPAD® version 5.6 (Coheris.SPAD-La Défense, Courbevoie, France).

4. Résultats

La population était composée de 65 femmes et 86 hommes, âgés entre 17 et 60 ans. Les femmes étaient en moyenne plus jeunes que les hommes, 51% avait moins de 30 ans versus 29% pour les hommes. Soixante quatre patients (42%) ont déclaré avoir un seul tatouage et six patients (7%) au moins dix. La majorité des tatouages que les patients voulaient enlever (78%) étaient situés sur une zone visible du corps : bras/avant-bras (43%), épaule/dos (26%) ou main (9%). Le tatouage concerné pouvait être soit récent soit très ancien : 11% des patients avaient le tatouage depuis moins d'un an et 24% depuis plus de 20 ans. Cinquante-huit pourcent des tatouages étaient de petite taille (définie comme inférieure à 30 cm²), 77% étaient monochromes, 72% étaient personnalisés et 39% avaient été réalisés par des non-professionnels. Les tatouages présentaient différent types de motif : animal (25%), fleur (15%), cœur (15%), initiales /prénom (12%), motif « tribal » (9%)... Soixante seize pourcent des patients ont déclaré qu'ils avaient été satisfaits ou très satisfaits par le tatouage après sa réalisation et seulement 5% ont déclaré avoir été non satisfaits.

Les raisons mentionnées pour la réalisation du tatouage concerné sont : entrainement/influence des autres (24%), esthétique (21%), affection/sentiment (20%), envie/plaisir (13%), affirmation de soi (13%), erreur de jeunesse (5%), symbolisme (4%), recouvrement d'un ancien tatouage (3%), masquage d'un problème de peau (3%) et 2% tatouage réalisé sous contrainte.

Les raisons mentionnées pour faire enlever le tatouage sont : esthétique (41%), regret/changement de vie (26%), discrédit social (19% : gène/honte, inadéquation par rapport à l'âge), professionnelles (17%), pression familiale ou du partenaire (11%), raisons personnelles non précisée (11%), et seulement 3% pour permettre la réalisation d'un nouveau tatouage. Comme attendu, les motivations sont clairement associées avec le degré de satisfaction après la réalisation du tatouage, mais aussi avec le sexe et l'âge du patient. En effet, les femmes de moins de 30 ans mentionnent plus fréquemment des pressions familiales ou du partenaire, et les hommes de quarante ans et plus évoquent plus fréquemment un discrédit social et des raisons professionnelles.

Les deux premières composantes de l'ACM, expliquant respectivement 87% et 8% de la variance totale selon la formule de calcul de Benzécri, ont été conservées pour la recherche de typologie (figure 1). La première composante est caractérisée par les conditions techniques de réalisation du tatouage (professionnel ou non, 25%), le délai de réalisation (13%), la localisation (visible ou non, 12%) et le type de tatouage (d'après modèle ou personnalisé, 12%), ainsi que l'âge du patient (13%). La seconde composante est caractérisée par la taille du tatouage (41%) et sa couleur (11%), le sexe du patient (18%) et son niveau d'éducation (16%).

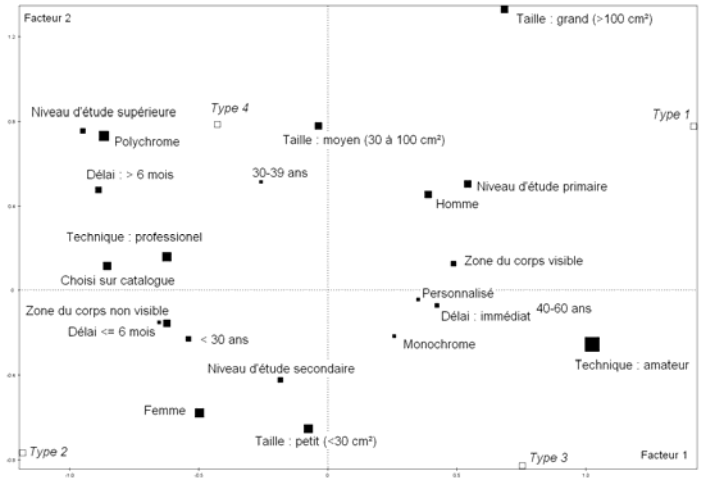


FIGURE 1 – *Carte des associations entre les caractéristiques des patients et celles de leurs tatouages (■), la taille des symboles indiquant leur contribution respective. Les deux premières composantes principales de l'ACM ont été utilisées comme système d'axes. La première composante explique 87% de la variance totale, et la seconde composante 8%. La typologie des patients (□) a été ajoutée sur la figure en variable supplémentaire.*

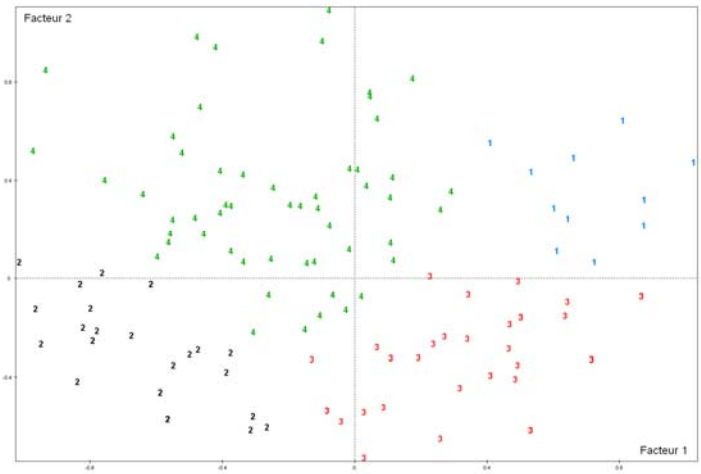


FIGURE 2 – *Carte des liens entre les patients. Les deux premières composantes principales de l'ACM ont été utilisées comme système d'axes. Chaque patient est indiqué par un chiffre correspondant à son type (type 1 = 1, type 2 = **2**, type 3 = **3** et type 4 = 4), certains patients étant masqués par d'autres.*

La classification ascendante hiérarchique a permis, à partir de ces deux composantes principales, d'identifier quatre types de patients (figures 1 et 2) :

- Le type 1 (n=19) est principalement composé d'hommes de quarante et plus, ayant un niveau d'étude primaire portant un tatouage monochrome et personnalisé, de taille moyenne voire grande (> 30cm²), réalisé par un non-professionnel sur une zone cutanée visible. De plus, le tatouage avait été généralement réalisé immédiatement après que la décision de se faire tatouer ait été prise.
- Le type 2 (n=30) est majoritairement composé de femmes âgées de moins de 30 ans, ayant un niveau d'étude secondaire, portant un tatouage de petite taille (< 30cm²), le plus souvent polychrome et réalisé

d'après modèle par un professionnel sur une zone cutanée non visible. Chez ces patients, le tatouage avait été réalisé généralement avec un délai de réflexion après la décision d'acquérir un tatouage.

- Le type 3 (n=43) est composé de patients des deux sexes, de quarante ans et plus avec un niveau d'étude primaire ou secondaire. Ces patients portent un tatouage monochrome et personnalisé de petite taille (< 30cm²) réalisé par un non-professionnel sur une zone cutanée visible. Le tatouage avait généralement été réalisé immédiatement après que la décision de se faire tatouer ait été prise.
- Le type 4 (n=56) est, quant à lui, principalement constitué d'hommes, leurs tatouages sont souvent polychromes de taille moyenne à grande (> 30cm²) et ont été réalisés par des professionnels.

Concernant les motivations pour la réalisation du tatouage, les patients de type 1 et 3 ont plus souvent évoqué un effet « d'entrainement/influence des autres » (58% et 37%, respectivement) que les patients de type 2 et 4 (7% et 12%, respectivement). Ces derniers ont, quant à eux, plus fréquemment évoqué une motivation "esthétique" (40% et 30%, respectivement). À propos des motivations pour l'enlèvement du tatouage, le fait de « ne pas aimer le tatouage » a été plus fréquemment mentionné par les patients de type 4 (34%), un sentiment de « gène /honte » par les patients de types 1 et 3 (37% et 33%, respectivement), et une motivation dite « professionnelle » par les patients de type 3 (30%).

5. Discussion

Quelles sont les raisons qui conduisent certains individus à modifier leur apparence par des tatouages : affirmation de soi, estime de soi, ou effet mode lié à la popularité croissante des tatouages qui sont à présent répandus dans toutes les couches de la population [VAR 99] ? Bien que les tatouages soient de plus en plus populaires en occident, de nombreuses personnes sont tôt ou tard amenées à regretter leurs tatouages. Notre étude montre que le tatouage peut même devenir un problème social et être responsable de souffrances psychologiques. Il semble donc important d'avoir des campagnes d'information concernant le tatouage afin d'éviter la réalisation de tatouages non désirés. Parmi les différentes techniques permettant de les enlever, les lasers pigmentaires sont actuellement le traitement de référence. Toutefois, ils ne permettent pas toujours d'effacer totalement un tatouage car le résultat dépend, entre autres, de la profondeur à laquelle les pigments ont été déposés dans le derme, de la quantité, et de la nature chimique des pigments utilisés (http://www.sfdermato.org/pdf/tatouageLPigm.pdf). Par ailleurs, il faut savoir que l'effacement d'un tatouage par traitement laser nécessite plusieurs séances à deux mois d'intervalle pour une même zone cutanée traitée, que cela peut être douloureux, et que c'est un acte médical non remboursé.

6. Bibliographie

[BEN 79] BENZECRI J.P., Sur le calcul des taux d'inertie dans l'analyse d'un questionnaire. *Cahiers de l'analyse des données* 1979, 4, 377-378.

[EVE 93] EVERITT B.S., *Cluster analysis,* Arnold, 1993.

[JOB 92] JOBSON J. D., Principal components, factors and correspondence analysis. Dans : *Applied Multivariate Data Analysis. Volume II: Categorical and Multivariate Methods* (Fienberg S, Olkin I, éditeurs). Dunod, 1992.

[LAU 06] LAUMANN A.E., DERICK A.J., Tattoos and body piercings in the United States: A national data set. *Journal of the American Academy of Dermatology* 2006, 55, 413-421.

[LEB 88] LEBART L., SALEM A., *Analyse statistique des données textuelles*, Dunod, 1988.

[STI 06] STIRN A., HINZ A., BRÄHLER E., Prevalence of tattooing and body piercing in Germany and perception of health, mental disorders, and sensation seeking among tattooed and body-pierced individuals. *Journal of Psychometrics Research* 2006, 60, 531-534.

[VAR 99] VARMA. S., LANIGAN S.W., Reasons for requesting laser removal of unwanted tattoos. *British Journal of Dermatology* 1999, 140, 483-485.

Approche pour le suivi des changements sur des données évolutives : application aux données du *marketing*

Alzennyr Da Silva*, Yves Lechevallier* et Francisco De Carvalho**

**Projet AxIS, INRIA Paris-Rocquencourt*
Domaine de Voluceau, Rocquencourt, B.P. 105,78153 Le Chesnay – France
{Alzennyr.Da_Silva, Yves.Lechevallier}@inria.fr

***CIn/UFPE, Caixa Postal 7851, CEP 50732-970, Recife (PE) – Brésil*
fatc@cin.ufpe.br

RÉSUMÉ. *Dans la fouille des données d'usage du Web, la dimension temporelle joue un rôle très important car les comportements des internautes peuvent changer au cours du temps. Dans cet article, nous présentons une approche de classification automatique basée sur des fenêtres sautantes pour la détection de changements sur les données d'usage. Cette approche combine les cartes auto organisatrices de Kohonen, la classification ascendante hiérarchique avec le critère de Ward et la coupure du dendrogramme selon le critère du gain d'inertie pour la découverte automatique du nombre de classes. Son efficacité a été analysée sur un jeu de données contenant les achats d'un panel de consommateurs.*

MOTS-CLÉS : *Classification automatique, Données évolutives, Fouille d'usage du Web*

1. Introduction

La fouille de données d'usage du Web (*Web Usage Mining*, en anglais) désigne l'ensemble de techniques basées sur la fouille de données pour analyser le comportement des utilisateurs d'un site Web [COO 99, SPI 99]. Dans ce cadre, cet article propose une approche pour la détection automatique et le suivi des changements d'usage au cours du temps (cf. section 2). Nous validons l'efficacité de notre approche sur des données évolutives réelles issues du *marketing* (cf. section 3). La conclusion et les perspectives sont présentées dans la section 4.

2. Approche de classification automatique pour la détection et le suivi des changements des données évolutives

Notre approche pour la détection et le suivi des changements sur des données évolutives consiste dans un premier temps à partitionner les données en fenêtres sautantes de taille fixe. Soit W_1 la première fenêtre du flux de données. L'idée est d'appliquer une méthode de classification non supervisée quelconque sur des données de W_1. Soit $G = (g_1, \ldots, g_c, \ldots, g_k)$ l'ensemble de prototypes des clusters découverts par cette classification. Ensuite, on repère la fenêtre suivante dans le temps, nommée W_2. Cette fenêtre ne se chevauche pas avec la première, d'où la désignation *sautante*. Par la suite, on affecte les données de W_2 aux prototypes de G, ce qui nous définit une première partition P_1. Puis, on applique la même méthode de classification non supervisé sur les données de W_2, ce qui défini une nouvelle partition P_2. La détection de changement entre les deux fenêtres sera donc mesurée par la comparaison des deux partitions P_1 et P_2 à l'aide des critères d'évaluation. Pour la comparaison des deux partitions nous appliquons deux indices de validation basés sur l'extension (individus) : la F-mesure [RIJ 79] et l'indice corrigé de Rand [HUB 85]. Le premier effectue une analyse cluster par cluster, alors que le deuxième fournit une mesure globale basée sur tout l'ensemble de clusters dans les deux partitions. La F-mesure assume des

valeurs contenues dans l'intervalle $[0, +1]$ et l'indice corrigé de Rand assume des valeurs entre $[-1, +1]$. Dans les deux cas, les valeurs proches de 1 correspondent à des partitions semblables, alors que les valeurs proches de 0 correspondent à des partitions très différentes.

Comme méthode de classification, nous utilisons les cartes auto organisatrices de Kohonen [KOH 95] initialisées à partir d'une analyse en composantes principales [ELE 99]. La couche de compétition est initialisée avec une centaine de neurones disposés sur une grille rectangulaire. Après la convergence, une classification ascendante hiérarchique utilisant le critère de Ward est appliqué sur les prototypes correspondant aux neurones de la grille. La coupure du dendrogramme résultant est effectuée selon le critère du gain d'inertie intra-classe. Les neurones dans un même groupe sont donc fusionnés. De cette manière, le nombre de clusters est automatiquement découvert. Il est important de remarquer que notre approche est totalement indépendante de la méthode de classification non supervisée appliquée, la seule restriction est que celle-ci doit fournir un prototype pour chaque cluster découvert.

3. Analyse des résultats sur des données réelles issues du *marketing*

Comme étude de cas, nous avons analysé un jeu de données diffusé dans le cadre du concours jeunes chercheurs au SLDS2009 [1]. Ce jeu de données concerne le suivi des achats de 10 068 clients pendant 14 mois (du 09 juillet 2007 jusqu'au 08 septembre 2008) sur 2 marchés de biens de consommation. Chaque marché commercialise 3 marques de produits. Pour l'application de notre méthode, nous avons défini un tableau croisé *client x marque achetée* ordonné selon la date de l'achat. Ce tableau contient un total de 262 215 lignes.

Dans notre approche, nous avons utilisé une fenêtre de temps de taille égale à une semaine. C'est-à-dire, notre méthode de classification non supervisée est appliquée sur l'ensemble d'individus ayant réalisé des achats dans une même semaine, ceci afin de segmenter les clients en fonction de leurs habitudes d'achat dans le temps. Le nombre total de clusters de préférences d'achat découverts par notre méthode varie entre 2 et 8 (cf. figure 1). L'axe des abscisses du graphique dans la figure 1 répresente la date de début de la semaine analysée. Le nombre de clusters le plus stable étant 7 pour la majorité des semaines. Un cluster représente un ensemble de clients ayant des préférences d'achat similaires pour une même semaine. Remarquons que pendant les trois premières semaines du mois de novembre de 2007, le nombre total de clusters de préférences d'achat a été stabilisé et égal à 2. Ceci peut être une réponse à une stratégie de vente mise en ligne pendant cette période de temps.

Les valeurs obtenues par les deux indices de comparaison de partition cités dans la section 2 sont montrés dans la figure 2. La F-mesure effectue une analyse cluster par cluster en cherchant la meilleure représentation (match) d'un cluster dans la première partition par un cluster correspondant dans la deuxième partition. La F-mesure obtient donc autant de valeurs qu'il y a de clusters dans la première partition. On trace une *boxplot* à partir de ces valeurs pour chaque semaine analysée. Les périodes les plus stables sont celles qui présentent les valeurs les plus élevées de la F-mesure. Sur le jeu de données analysé, ces périodes correspondent aux semaines débutées le *20 août 2007*, *17 septembre 2007*, *10 décembre 2007*, *25 février 2008* et *02 juin 2008* (cf. côté gauche de la figure 2). Ces mêmes périodes ont été également repérées par l'indice de Rand corrigé (cf. côté droit de la figure 2). Cet index fournit une mesure globale basée sur tout l'ensemble de clusters dans les deux partitions. Ceci montre l'accord entre ces deux indices et leur aptitude à résoudre ce type de problème. Pour les autres semaines analysées, ce jeu de données reste assez instable. Les clusters de comportement subissant un nombre important de changement au cours du temps.

Afin de mieux connaître les profils d'achat de ces clients, nous avons tracé sur la figure 3 la distribution des pourcentages d'achat par marque pour les 7 clusters détectés par la méthode pendant les périodes les plus stables, ces clusters représentent les préférences d'achat les plus typiques. Ces profils d'achat peuvent être décrits comme suit :

- **Cluster 1** : clients ayant une préférence d'achat pour la marque B en premier plan et la marque A en deuxième plan.
- **Cluster 2** : profil mixte de clients ayant des fortes préférences pour les marques A et D.
- **Cluster 3** : clients ayant une préférence d'achat majoritaire pour la marque A.
- **Cluster 4** : clients ayant une préférence d'achat pour la marque F en premier plan et les marques A et D en deuxième plan.
- **Cluster 5** : clients ayant une préférence d'achat pour la marque C en premier plan et les marques E et A en deuxième plan.
- **Cluster 6** : clients ayant une préférence d'achat majoritaire pour la marque D.

1. Symposium Apprentissage et Science des Données 2009, `www.ceremade.dauphine.fr/SLDS2009`

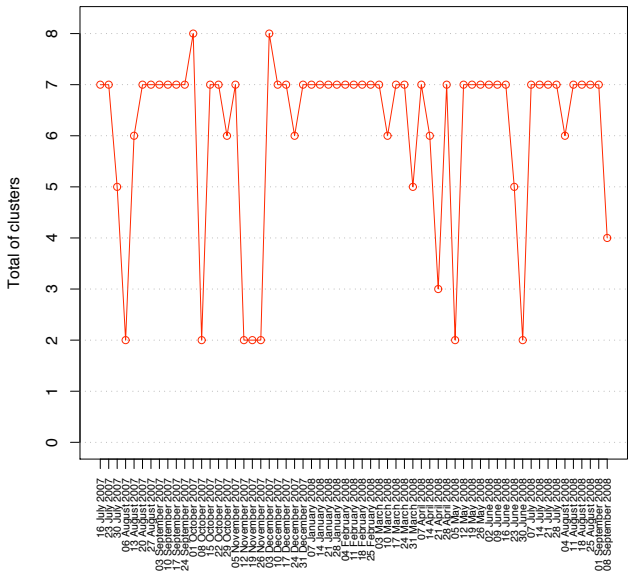


FIGURE 1. *Nombre de clusters de comportement par fênetre de temps (semaine).*

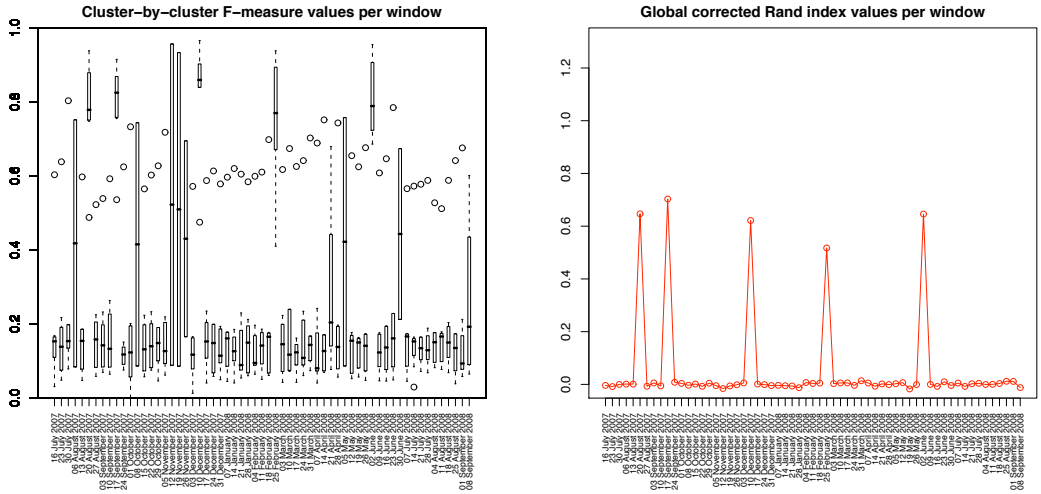


FIGURE 2. *Valeurs obtenus par la F-mesure (à gauche) et par l'indice de Rand corrigé (à droite) pour chaque semaine analysée.*

– **Cluster 7** : clients ayant une préférence d'achat pour la marque E en premier plan et les marques A et D en deuxième plan.

La marque A a une forte préférence parmi tous les profils d'achat. Les autres marques sont présentes dans des profils ponctuels. En consultant les résultats obtenus, il est possible d'extraire la liste de clients appartenant à un cluster spécifique.

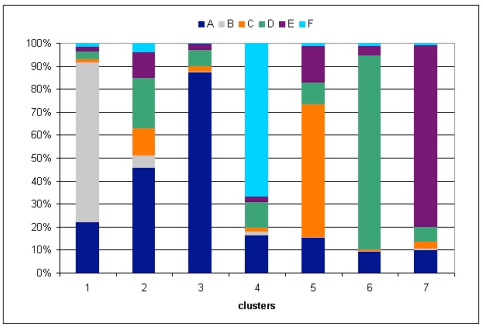


FIGURE 3. *Clusters de préférences d'achat les plus typiques.*

Pour suivre le comportement d'achat d'un client spécifique, il suffit de vérifier si le client en question change de cluster au cours du temps. Si le client reste toujours dans un même cluster au cours du temps, celui si peut être considéré 'fidèle' aux marques concernées par ce cluster. D'un autre côté, si le client change considérablement de cluster au cours du temps, celui si peut être classé tel un 'zappeur"' entre les marques de produit.

4. Conclusion et perspectives

Dans cet article nous avons décrit les résultats obtenus par de notre méthode de classification non supervisée pour la détection et suivi des clusters de comportement dans le temps appliquée sur un jeu de données du *marketing*. Nous avons pu suivre l'évolution de clusters de préférences d'achat des clients sur 6 marques de produits de deux marchés différents au long de 14 mois. Notre méthode a permis la traçabilité des changements subis par les clusters de préférences d'achat des clients. Tous les résultats obtenus par l'application de notre méthode sur le jeu de données en question ont été enregistrés dans une base de données MySQL et peuvent de ce fait faire l'objet a posteriori d'un rapport technique. Il est ainsi possible de fournir à des questions spécifiques des réponses plus détaillées. Par exemple, la segmentation des clients pour une semaine spécifique peut être facilement repérée par une simple requête à la base de données. Il est également possible de mesurer la popularité de certaines marques de produit en fonction du suivi des effectifs d'un même groupe de comportement au cours du temps.

Comme travaux futurs, nous envisageons d'intégrer dans cette approche une analyse sémantique capable d'interpréter la nature des changements (apparition/ disparition des groupes de comportement, migration des individus de et vers un groupe de comportement, etc.).

5. Bibliographie

[COO 99] COOLEY R., MOBASHER B., SRIVASTAVA J., Data Preparation for Mining World Wide Web Browsing Patterns, *Journal of Knowledge and Information Systems*, vol. 1, n^o 1, 1999, p. 5-32.

[ELE 99] ELEMENTO O., Apport de l'analyse en composantes principales pour l'initialisation et la validation de cartes topologiques de Kohonen, *Actes des 7èmes journées de la Société Francophone de Classification (SFC'99)*, Nancy, France, 1999.

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

[KOH 95] KOHONEN T., *Self-Organizing Maps*, vol. 30 de *Springer Series in Information Sciences*, Springer, third édition, 1995, Last edition published in 2001.

[RIJ 79] VAN RIJSBERGEN C. J., *Information Retrieval*, Butterworths, London, second édition, 1979.

[SPI 99] SPILIOPOULOU M., Data Mining for the Web, *Workshop on Machine Learning in User Modelling of the ACAI99*, , 1999, p. 588-589.

Classification des émotions dans un espace à deux dimensions

Maher CHEMSEDDINE, Monique NOIRHOMME

FUNDP
Faculté d'Informatique, Namur,
21 rue grandgagnage
5000 Namur, Belgique
mch@info.fundp.ac.be , monique.noirhomme@info.findp.ac.be

RÉSUMÉ. Dans cet article, nous présentons une classification des émotions liées à la musique dans un espace à deux dimensions. Nous avons choisi 18 termes (émotions) de différentes langues (Arabe, Anglais et Français). La classification est effectuée à partir d'une évaluation suivant deux axes orthogonaux : la « valence » (émotion positive ou négative) et l'excitation (passive ou active). Nous avons développé un questionnaire en ligne destiné à proposer une classification des émotions. Au total 97 participants de différentes cultures ont répondu à ce questionnaire. Nous avons considéré les émotions comme des données symboliques. Les analyses statistiques montrent une répartition des émotions en 8 classes principales distinctes dans cet espace, qui sont concordantes avec les résultats antérieurs mais plus faciles à interpréter.

MOTS-CLÉS : classification, objet symbolique, émotions, musique.

1. Introduction

Notre travail s'inscrit en fait dans un contexte plus large. Afin de créer du son à partir de données, nous avons pris le parti d'associer les données à des émotions, sachant que de nombreux travaux existent liant son et émotion. Pour pouvoir lier données et émotion, nous nous sommes dès lors intéressés à la classification des émotions dans un contexte musical. C'est cette partie de notre travail qui est présentée ici.

Beaucoup d'études psychologiques, philosophiques et sociologiques ont essayé de faire une classification des émotions. Différentes méthodes ont été proposées pour la classification des émotions [STR 03]. Il existe principalement deux grands systèmes de classification : la classifications par catégories et la classification dimensionnelle. La classification par catégories suppose que les émotions ont différentes significations comme heureux et triste qui sont distincts et indépendants. En revanche, la classification dimensionnelle présente les émotions comme liées au sein d'un espace sémantique à n-dimensions.

Au 19ème siècle, Wilhelm Wundt [PRL 06], un psychologue allemand, suggère que l'expérience émotionnelle peut être décrite en termes de trois dimensions: la « valence », l'excitation, et la puissance. La « valence » est la dimension la plus dominante. Elle représente les valeurs positif (heureux) ou négatif (triste) de l'émotion. L'excitation (ou l'activité) représente le degré de l'engagement psychologique d'une personne - le degré de sa vigilance. Elle représente des valeurs entre passif (valeur négative) et actif (valeur positive). La puissance se réfère aux sentiments de l'individu par rapport à une émotion. Cependant la « valence » et l'excitation ont été identifiées comme les dimensions majeures de l'espace émotionnel [LAR 92]. Leur relation à la puissance a également été un sujet de débat.

2. Travaux annexes

Dans la littérature, nous retenons le « circumplex » [RUS 89], parmi les modèles à deux dimensions. Dans une telle présentation, les émotions sont réparties approximativement le long d'un cercle centré sur le plan cartésien

(Figure 1). Selon ce modèle les émotions qui ont presque le même sens sont géométriquement proches comme par exemple les émotions ravi et heureux.

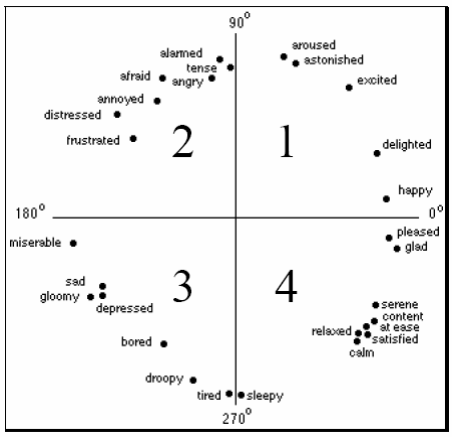


FIGURE 1 – LE MODEL DE « CIRCUMPLEX », SOURCE : RUSSELL 1989

Nous distinguons encore les « Adjectives » de Hevner [HEV 36]. Ces « Adjectives » consistent en une liste de 67 termes arrangés en huit classes. Chaque classe a été organisée de telle sorte que les termes qui le constituent ont une forte relation du point de vue du sens. Les classes sont réparties dans l'espace à deux dimensions de façon circulaire de telle manière qu'ils forment une sorte de continuum (Figure 2).

A partir d'une expérimentation sur 28 étudiants en musique Schubert [SCH 99] à transformé numériquement les termes de Hevner afin de les projeter dans l'espace [valence, arousal (*excitation*)]. Il a ensuite représenté les 8 groupes de Hevner par des ellipses empiriques. (Figure 3)

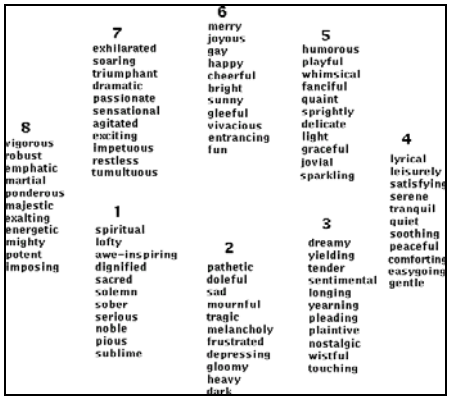


FIGURE 2 – LES « ADJECTIVE » DE HEVNER, SOURCE : HENVER 1936

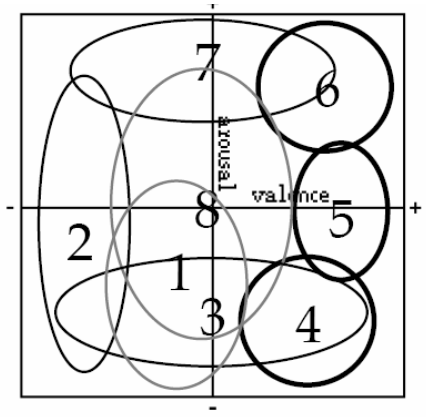


FIGURE 3 – LA CLASSIFICATION DE SCHUBERT EN 8 ELLIPSES, SOURCE SCHUBERT 1999

Nous remarquons que les différentes tentatives de classifications restent significativement pauvres. Par exemple, le « circumplex » suggère que les émotions peuvent être décrites en deux dimensions avec une forme circulaire, en outre il spécifie l'ordre dans lequel les émotions sont reparties radialement le long du périmètre d'un cercle. La présentation des classes sous forme d'ellipses ne montre pas exactement la répartition des émotions dans l'espace, elle indique uniquement le groupement des émotions proches ou similaires dans une même classe. Dans ce papier nous présentons une méthode de classification en utilisant l'analyse des données symboliques. Cette méthode nous permet d'interpréter les classes obtenues en fonction des axes de départ.

3. Méthode

Conformément à la procédure adoptée par Russell [RUS 89], nous avons conçu et développé une application interactive en ligne appelée "EmoQuery"[1], qui est un outil d'autoévaluation de l'expérience émotionnelle. L'espace émotionnel de notre outil se compose de deux axes orthogonaux qui se croisent à l'origine du plan cartésien: la dimension valence est représentée par l'axe des abscisses avec une valeur maximale de 100% (émotion heureux) et une minimale égale à -100% (émotion triste). La dimension d'excitation est représentée par l'axe des ordonnées avec une valeur maximale 100% (active) et une minimale -100% (passive). L'emplacement de chaque émotion dans l'espace cartésien est le résultat de la combinaison des valeurs de ces deux axes. 18 termes (émotion) ont été sélectionnés dans trois langues (Arabe, Anglais et Français), les émotions sélectionnées ont été validées entre autres par Schubert [SCH 99]. Des 24 émotions de Schubert nous en avons écarté 6 à cause d'une traduction difficile de l'anglais vers le français. En effet le français ne permettait pas de distinguer entre ces termes (exemple agité et excité, en anglais « agitated » et « excited »).

Nous avons sélectionné 97 participants de différents pays, d'âge compris entre 20 et 50 ans. Nous avons demandé de positionner les 18 émotions dans ce plan à deux dimensions.

4. Résultats et discussion

Pour chaque émotion nous avons calculé la valeur minimale et la valeur maximale sur l'ensemble des individus. Puis nous avons écarté 10% des résultat les plus éloignés du centre afin d'obtenir des objets plus denses [YVE 08].

Chaque émotion peut être représentée sur les deux axes par un intervalle. Nous pouvons dès lors considérer ces données comme des données symboliques [BOC 00]. Chaque émotion est représentée dans le plan [valence, excitation] par un rectangle (Figure 4). Nous pouvons appliquer à l'ensemble des rectangles une méthode de classification hiérarchique. Nous avons choisi la distribution euclidienne et la classification sur les centres (average linking) pour classer les données [BRI 00] Nous avons également utilisé une méthode de classification dynamique (SCLUST) [FRA 08]. En demandant une classification en 8 classes, nous obtenons le même résultat que pour la classification hiérarchique. La table 1 donne le résultat de la classification. La Figure 5 fournit les 8 classes dans l'espace à deux dimensions.

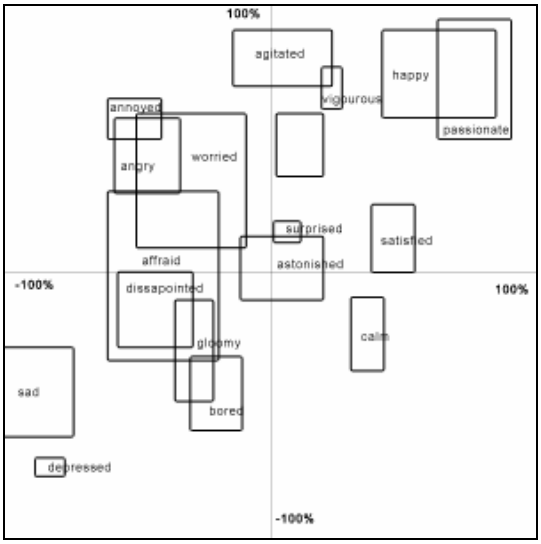


FIGURE 4 – LA REPARTITION DES EMOTIONS DANS L'ESPACE A DEUX DIMENSIONS

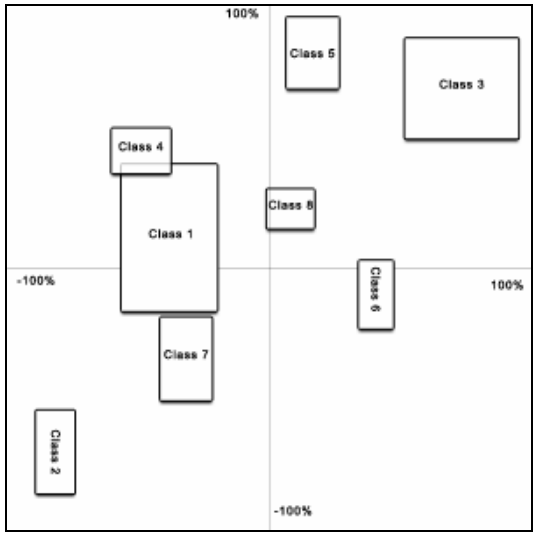


FIGURE 5 – REPARTITION DES CLASSES PAR LA CLASSIFICATION DYNMAIQUE

1" Emoquery": site web : *www.media19.be/emotion/*

TABLE 1 – Répartition des émotions dans les 8 classes.

Classe 1	Classe 2	Classe 3	Classe 4	Classe 5	Classe 6	Classe 7	class8
apeuré inquiet	déprimé, triste	heureux, passionné	en colère, agacé	agité, vigoureux	calme, satisfait	ennuyé, déçu, morose	étonné, sérieux, surpris

A noter que la classification obtenue est très proche de celle de la Figure 2 mais nous voyons mieux les limites des classes. Par exemples pour la classe 1 (apeuré, inquiet) la valence est compris entre -19.5% et -55.5%, l'excitation entre 44.5% et -11.5%.

Nous avons aussi essayé de voir s'il y avait une différence culturelle ou linguistique entre les répartitions des termes, mais malheureusement nous n'avons pas obtenu de résultats significatifs. Un problème vient du fait que plusieurs participants n'ont pas répondu au questionnaire dans leur langue maternelle (ou de leur pays) donc il était impossible de mettre en relation l'aspect culturel et les termes utilisés.

5. Conclusion

Nous avons sélectionné 18 termes décrivant des émotions en relation avec la musique et validés par des travaux antérieurs grâce au questionnaires en ligne. Nous avons demandé à un échantillon de 97 personnes de positionner ces émotions dans un espace cartésien à deux dimensions. Chaque émotion, représentée par 97 points, a été interprétée comme objet symbolique. Nous avons ensuite classé ces émotions en 8 classes par une méthode d'analyse symbolique. Le résultat obtenu est cohérent avec les recherches antérieures mais plus riches pour l'interprétation. Nous allons utiliser cette classification pour mettre en relation des données avec des émotions et ensuite sonoriser ces émotions. De point de vue de la classification, nous pensons comparer le résultat obtenu avec une méthode basée sur les rangs des observations plutôt que sur les valeurs exactes.

6. Bibliographie

[STR 03] STRONGMAN K.T., *The psychology of emotion (5th edition)*, 2003 p. 180-182, Wiley.

[PRL 06] POPULATION RESEARCH LABORATORY, *Valence and Arousal as Fundamental Dimensions of Emotional Experience*, 2006 Site web: http://www.uofaweb.ualberta.ca/prl/

[LAR 92] LARSEN, R., DIENER, E., *Promises and problems with the circumplex model of emotion*, 1992 dans M. S. Clark (Ed.). Emotion , p. 25-59 Newbury Park, CA: Sage Publications.

[RUS 89] RUSSELL, J.A., *Measures of emotion*, 1989 dans R. Plutchik & H. kellerman edition, Emotion: theory research and experience, Vol 4, p. 81 – 111, New york : Academic Press.

[HEV 36] HEVNER, K., 1936, Adapter de Schubert E. 2003, *Update of the hevner adjective checklist*, vol. 96 (2), n°3, p. 1117-1122, Perceptual and motor skills.

[BRI 00] BRITO, P., *Hiearchical and pyramidal clustring with complete symbolic objects, Analysis of symbolic data*, 2000, Berlin-Heidelberg p. 312 – 341, Springer.

[BOC 00] BOCK, H.H. ET DIDAY, E., *Symbolic objects, Analysis of symbolic data*, 2000, Berlin-Heidelberg, p 54 – 77, Springer.

[SCH 99] SCHUBERT, E., *Measurement and time series analysis of emotion in music*, 1999, thèse de doctorat, p. 155-199, University of New South Wales.

[YVE 08] LECHEVALIER Y., L. EL GOLLI A., HEBRAIL G., *Improved generation of symbolic objects from relation database*, 2008 dans symbolic data analysis and the SODAS software, p. 45-60 Wiley.

[FRA 08] FRANCISCO A.T, LECHEVALIER Y., VERDE R., *Clustering methods in symbolic data analysis*, 2008 dans symbolic data analysis and the SODAS software, p. 181-204 Wiley.

Utilisation de RandomForest pour la détection d'une signature protéomique du cancer du poumon.

Cécile Amblard[1], Sylvie Michelland[2],Florence de Fraipont[2] , Denis Moro-Sibilot[3], Francois Godard[2], Marie Christine Favrot[2], Michel Seve[2].

1 : TIMC-IMAG, Faculté de Médecine, 38730 La Tronche,
2 : Centre d'Innovation en Biologie, Pav B, CHU Grenoble BP217, 38043 Grenoble,
3 : PMAC Clinique de Pneumologie, UF Oncologie thoracique, CHU Grenoble.

RÉSUMÉ. *Ce travail est consacré à la recherche d'une signature protéomique du cancer du poumon. A partir de spectres acquis avec un spectromètre de masses SELDI, nous recherchons un ensemble de protéines permettant de discriminer les patients atteints d'un cancer du poumon des sujets sains. Le problème est difficile, en particulier parce qu'il y a plus de protéines que de patients et parce qu'il y a une grande variabilité dans les spectres d'une même population. Nous cherchons aussi à nous affranchir des effets de l'âge et du sexe. Nous avons choisi d'utiliser Random Forest. Cette méthode de classification naturellement multivariée, permettant d'identifier les variables discriminantes et étant capable de travailler avec plus de variables que d'échantillons nous semble bien adaptée au problème posé. Nous avons ainsi trouvé un ensemble de protéines discriminantes dont certaines ont pu être identifiées.*

MOTS-CLÉS : *Analyse des données, Protéomique, RandomForest*

1. Introduction

Le cancer du poumon est un cancer particulièrement agressif et est une cause majeure de mortalité dans le monde. Le diagnostique se fait par scanner-outil ne permettant pas de déceler un cancer précoce ou par prélèvement-méthode invasive et onéreuse, ne pouvant être utilisée dans le cadre d'un dépistage systématique. La mise en évidence d'anomalies biologiques, détectables dans le sang, des mois avant l'apparition de signes cliniques de ce cancer permettrait d'effectuer une campagne de dépistage, de diagnostiquer des cancers en phase précoce et donc d'améliorer le taux de survie des sujets atteints. La présence d'une tumeur dans un tissu se traduit par une modification significative du profil protéomique du patient : présence ou absence anormale de certaines protéines. Ce travail est consacré à la recherche d'une signature protéomique du cancer du poumon, c'est à dire à l'identification de différences significatives entre les profits protéomiques de patients atteints du cancer et ceux de sujets sains. Les profils protéomiques sont acquis avec un spectromètre de masses à temps de vol SELDI, déjà utilisé en oncologie [WRI 02], [WIE 04]. La recherche de protéines permettant de discriminer les sujets sains et les patients est un problème difficile : Il y a une grande variation entre les profils protéomiques de sujets appartenant à une même population. Le nombre de protéines ou de peptides est généralement bien supérieur au nombre de sujets. De plus, les protéines sécrétées par les tissus sont bien moins présentes que d'autres propres au serum très abondantes comme l'albumine, l'immunoglobuline, la transférine ou la fibrinogène, ce qui rend difficile leur identification. Enfin d'autres facteurs que la maladie peuvent avoir un effet significatif sur le profil protéomique d'un sujet, par exemple le sexe, l'âge...Nous avons choisi d'utiliser Random Forest [BRE 01] pour réaliser l'influence du sexe et de l'âge chez les sujets sains puis pour discriminer les profils protéomiques des sujets sains et des patients et identifier une signature protéomique du cancer du poumon en prenant en compte les effets d'âge et de sexe. Cette méthode de classification naturellement multivariée, capable de travailler dans un contexte où il y a plus de variables que d'échantillons et enfin permettant d'identifier les variables les plus discriminantes nous a semblé très

appropriée. Dans une deuxième partie, nous présentons les données. Dans une troisième partie, nous expliquons comment nous avons identifié l'effet du sexe, de l'âge et de la maladie sur les profils protéomiques à l'aide de Random Forest. Enfin nous présenterons les résultats obtenus.

2. Description des données

L'étude porte sur un ensemble de 83 échantillons de plasma issus de 42 sujets identifiés comme cliniquement sains et de 41 patients atteints du cancer du poumon, dont 14 sont atteints d'un adénocarcinome, 26 d'un SCC (squamous cell carcinome) et 1 d'un autre type de cancer. Le profil protéomique de ces échantillons a été réalisé a l'aide de la plateforme ProtéinChip Biomarker System Cyphergen Biosystems Inc, Fremont CA). Un spectre est une collection ordonnée de couples (masse, quantité de particules ionisées détectées).
Un spectromètre de masses à temps de vol se compose schématiquement d'une entrée où l'on dépose une plaque de métal contenant une solution ionisée à analyser, d'une portion cylindrique que l'on peut soumettre à un champ electrique, d'une portion cylindrique sous vide qui aboutit à un détecteur de ions et un laser dont on peut régler l'intensité qui permet de decrocher de leur support les particules ionisées. Une quantité de plasma mélangée à une solution ionisante : la matrice, est déposée sur un support métallique. Ce support est placé à l'entrée du spectromètre. Les particules ionisées sont détachées du support à l'aide du laser. Ces particules sont accélérées dans la partie soumise au champ électrique puis continuent à "voler" dans la partie sous vide jusqu'a atteindre le détecteur. Un spectre "brut" est donc un ensemble de couples où la première coordonnée est le temps (l'instant d'acquisition) et la seconde coordonnée est le nombre de ions d'etectés à cet instant. Le temps de vol étant dépendant de la masse de la particule, le temps est classiquement transformé par une équation quadratique en ensemble de masses.
Avant de véritablement commencer l'analyse statistique des spectres, il est nécessaire d'effectuer un certain nombre de prétraitements sur ces données tels que le débruitage, la normalisation, l'alignement de spectres. Ces prétraitrements ont été réalisés à l'aide du logiciel Cyphergen ProteinChip Software 3.1. Le bruit additif, dû à la solution chimique ionisante se traduit par la présence d'un grand nombre de protéines en début de spectre. On s'affranchit de ce bruit en effectuant une régression sur les minima locaux et en soustrayant au signal cette tendance lisse obtenue. On s'affranchit du bruit de comptage en ne conservant sur une fenêtre de taille fixe que les pics pour lesquels le rapport signal sur bruit est supérieur à 5. L'alignement des pics se fait manuellement en alignant les centres de chaque pic diffus. Enfin, les spectres sont normalisés, en supposant que la charge totale de chaque spectre est identique. Après toutes ces étapes, nous disposons de 83 spectres constitués de 221 couples (masses, intensités) que nous appelons pics.

3. Analyse statistique

Le but de notre travail est de détecter un ensemble de protéines-signature du cancer du poumon afin d'être capable d'identifier un patient atteint de cette pathologie. Cette signature ne doit pas être biaisée par des facteurs exterieurs tels que le sexe ou l'âge. La recherche d'une signature protéomique du cancer du poumon (resp. d'une histologie particulière du cancer du poumon est un problème d'analyse discriminante classique. La méthode utilisée doit identifier parmi tous les pics la combinaison qui discrimine au mieux les sujets sains et malades et doit nous permettre de détecter si un individu est sain ou malade avec un faible taux d'erreur. La méthode doit ètre capable de travailler avec un nombre de variables plus grand que le nombre de sujets. Random Forest (RF) permet justement de réaliser ces 2 objectifs. C'est une méthode naturellement multivariée, capable de travailler lorsque il y a plus de variables que d'individus. De plus le pouvoir discriminant de chaque variable est réellement basé sur l'aptitude d'une variable à partitionner une population en sous groupes. Notons que cette méthode a déjà été utilisée dans le cadre de traitement de données génomiques (par exemple [LEC 07] ou protéomiques [IZM 04]. Nous avons utilisé la version libre sous R de Random Forest disponible à l'adresse http ://cran.r-project.org. Nous avons utilisé RF sur l'ensemble des patients sains/malsains (resp. sains/atteints d'une histologie particulière du cancer du poumon). Afin de prendre en compte le fait que les classes sont mal équilibrées, nous avons utilisé comme paramètres un vecteur de poids égal à la proportion des individus dans chaque classe. Les pics liés à l'âge ou au sexe sont retirés. Après avoir utilisé RF, nous conservons les 11 pics identifiés comme étant les plus discriminants par RF, rangés

par pouvoir discriminant décroissant. Toutes les combinaisons, contenant le premier ou le second pic, de 2 à 9 pics, sont alors considérées. Pour chaque combinaison, RF est alors utilisé sur l'ensemble des spectres réduits à cet ensemble de variables. L'évaluation des résultats se fait alors à l'aide du taux d'erreur estimé à partir des éléments Out of Bag (OOB) [BRE 96]. Ce taux compte pour chaque arbre de la forêt, le nombre d'élements mal classés parmi les individus non répliqués dans l'échantillon bootstrap utilisé pour la construction de cet arbre. Les meilleures combinaisons sont donc celles ayant obtenu le meilleur taux d'erreur.
Afin de déterminer si le sexe d'un sujet a une influence sur le profil protéomique, nous avons utilisé RF comme suit. Nous nous sommes limité à l'ensemble des individus sains, constitué de 27 femmes et de 15 hommes.

1. Après avoir utilisé RF avec un vecteur de poids égal aux proportions de chaque classe, sur cette population, nous conservons les 11 variables les plus discriminantes, rangées par ordre de pouvoir discriminant decroissant.

2. Nous considérons toutes les combinaisons de 2 à 9 variables contenant la première variable. Si cette combinaison permet au moyen de RF de classer les individus de manière assez satisfaisante (sensibilité et spécificité évaluées avec les éléments OOB supérieures à 70%), la variable la plus discriminante est enlevée. Elle est identifiée comme étant liée au sexe. On recommence alors en 1 avec les spectres précédents privés de cette variable.

3. L'algorithme s'arrête lorsqu'aucune combinaison restante ne permet d'obtenir une spécificté et une sensibilité supérieures à 70%. Les variables liées au sexe sont celle qui ont été ôtées.

Pour déterminer les effets eventuels de l'âge sur le profil protéomique d'un sujet, nous avons utilisé RF comme méthode de regression. La méthodologie est similaire à celle utilisée pour étudier l'effet du sexe ; une combinaison "gagnante" à laquelle nous ôtons la première variable est une combinaison permettant d'expliquer 40% des variations de l'âge.

4. Résultats

Les peptides liées au sexe sont celles situées aux masses 49201, 28115, 6633, 28327, 50167, 12850, 13793, 4588, 6436, 4452, 13289, 7443, 9350, 11728, 13051 et 9254. Les peptides liées à l'âge ont pour masses 13387, 9367, 13758 et 3102. Tous ces pics sont dorénavent ôtés des spectres. La meilleure combinaison de masses permettant de discriminer les patients sains et atteints d'un cancer du poumon est 11761, 6167, 8734, 8559, 8144, 6812 et 9160. Cette combinaison permet d'obtenir un taux d'erreur (OOB) de 91,5% avec une sensibilité de 90,2% et une spécificité de 92,8 %. 37 patients sur 41 et 39 sujets sains sur 42 sont bien classés. Il est intéressant de noter que 3 parmi les 4 patients mal classés sont atteints d'un adénocarcinome. Enfin, notons que la combinaison des trois premiers marqueurs permet d'obtenir un taux d'erreur de 87%. 6 combinaisons de peptides permettent de classer 96,1 % des patients atteints d'un SCC et 95,2 % des sujets sains. Ces combinaisons contiennent toutes les masses 11761, 6167, 14114 et 15287. Toutes contiennent de plus les masses 6609 ou 6812. On peut remarquer aussi que le marqueur 8734 précédement trouvé comme discriminant des sains/cancéreux n'apparait plus. Afin de compléter l'interprétation de ces résultats, nous avons fait une étude descriptive des différents marqueurs. Il apparait que les masses 6812, 6609 et 6418 sont sous-exprimées chez les patients atteints de SCC et très fortement positivement corrélées. Ces 3 masses contiennent probablement la même information. La masse 6167 ne fait pas parti de ce groupe. Nous n'avons pas fait l'étude des sujets sains/patients atteints d'un adénocarcinome parce qu'il y a trop peu d'adénocarcinomes (14).
Parmi toutes les protéines trouvées, trois ont pu être identifiées : la masse 11761 est la SAA : d'autres publications ont mentionné que cette protéine était élevée chez les patients atteints de cancer de la prostate ou du larynx. Les masses 13758 et 13793 sont respectivement la transtyretine et la cystatine.

5. Conclusion

Nous avons étudié les profils protéomiques de sujets sains et de sujets atteints d'un cancer du poumon. En utilisant RandomForest, nous avons trouvé que le sexe et l'âge d'un patient peuvent avoir un effet sur le profil protéomique. Nous avons aussi trouvé un ensemble de 7 masses capables de discriminer les patients sains de ceux

atteints de cancer avec un taux d'erreur (estimé à l'aide des éléments "out of bag") de 91,5%. Malheureuement l'identifcation protéomique de ces pics n'a pas pu être menée à bout.

6. Bibliographie

[BRE 96] BREIMAN L., Out of bag estimate, *Curr Pharm Biotechnol*, , 1996.

[BRE 01] BREIMAN L., Random Forest, *Machine Learning*, vol. 45, 2001, p. 5–22.

[IZM 04] IZMIRLIAN G., Application to the random forest classification algorithm to SELDI TOF proteomic study in the settinfg of a cancer prevention trial, *Ann NY Accad Sci*, vol. 1020, 2004, p. 154–174.

[LEC 07] LE CAO K., GONÇALVES O., BESSE P., GADAT S., Selection of Biologically Relevant Genes with a Wrapper Stochastic Algorithm, *Statistical Applications in Genetics and Molecular Biology*, vol. 6 : Iss. 1, 2007, p. art 29–.

[WIE 04] WIESNER A., Detection of tumor markers with Protenichips technology, *Curr Pharm Biotechnol*, vol. 5, 2004, p. 45–67.

[WRI 02] WRIGHT G., SELDI Protenichips MS : a platform for biomarker discovery and cancer diagnosis, *Expert Rev Mol Diagn*, vol. 2, 2002, p. 549–563.

Une méthode de combinaison de résultats de classifications : application à la fusion d'hypnogrammes

Teh Amouh, Monique Noirhomme-Fraiture, Benoît Macq

Faculté d'informatique – FUNDP
Rue Grandgagnage, 21
5000 Namur, Belgique
{tam,mno}@info.fundp.ac.be

Laboratoire de Télécommunications et Télédétection – UCL
place du Levant, 2
B-1348 Louvain-la-Neuve, Belgique
benoit.macq@uclouvain.be

RÉSUMÉ. *Les méthodes d'ensemble en classification sont des techniques d'apprentissage dont le but est de mettre à profit la complémentarité entre plusieurs classifieurs en fusionnant leurs différentes réponses à un même problème de classification. Les algorithmes de fusion généralement utilisés peuvent être regroupés en trois grandes familles : le vote majoritaire, les formulations bayésiennes et les fonctions de croyance. Nous proposons un cadre de combinaison de classifieurs, appelé aggrégation stratifiée, qui est applicable quels que soient les algorithmes de fusion utilisés. Nous montrons quelques résultats numériques obtenus dans le domaine de la construction automatique d'hypnogrammes.*

MOTS-CLÉS : *Méthode d'ensemble, Fusion de classifieurs, Classification supervisée, Approximation de distributions discrètes.*

1. Introduction

Plusieurs classifieurs différents (obtenus par la manipulation des donnés d'apprentissage, l'utilisation de différents vecteurs de caractéristiques ou de différents paramétrages des algorithmes d'apprentissage, etc.) peuvent être mis en œuvre pour aborder un même problème de *reconnaissance de pattern*. Il est admis que des méthodes d'ensemble (méthodes de fusion des différentes réponses fournies par différents classifieurs) peuvent permettre d'améliorer les performances de la classification. Certaines méthodes d'ensemble (*bagging*, *boosting*, etc.) incluent la génération et l'apprentissage des classifieurs individuels et utilisent des techniques de vote pour la fusion de leurs réponses [DIE 00]. D'autres supposent que les classifieurs à fusionner sont donnés et fixes (préalablement appris) et les fusionnent soit par le vote [LAM 97], soit par des techniques bayésiennes [KAN 99, KIT 97], ou encore par des approches basées sur les fonctions de croyance [APP 02]. Dans tous les cas, les méthodes d'ensemble fusionnent systématiquement l'ensemble des réponses de tous les classifieurs. Ce type de fusion sera qualifié de "fusion complète" dans la suite de cet article. La fusion complète ne conduit pas nécessairement aux meilleures performances possibles puisque l'on a pu observer expérimentalement [KAN 99] que le sous-ensemble de classifieurs dont la fusion serait optimale n'est pas nécessairement l'ensemble complet de tous les classifieurs. Dans cet article, nous proposons de regrouper d'abord les classifieurs en *strates*, une strate étant simplement un sous-ensemble de l'ensemble des classifieurs, ensuite d'opérer une fusion complète à l'intérieur de chaque strate pour obtenir la réponse de chaque strate, puis enfin de fusionner l'ensemble des réponses des strates par le vote majoritaire relatif pondéré.

2. Formulation du problème de fusion de classifieurs à sortie nominale

Le problème de fusion considéré dans cet article porte sur des classifieurs qui, à un pattern en entrée, associent chacun une catégorie unique qui est soit le rejet, soit l'une des classes de l'application. Par "données d'apprentissage" nous entendons un ensemble de patterns étiquettés servant à entraîner non pas les classifieurs à fusionner, mais bien la méthode de fusion puisque chacun de classifieurs est supposé déjà entraîné. En notant R la variable dont la valeur correspond à la classification correcte, $R(p)$ désigne la vraie classe à laquelle appartient le pattern p. Soient K le nombre de classifieurs et N le nombre de classes. On note e_k ($k = 1, \cdots, K$) le classifieur k, et $\Lambda = \{1, \cdots, N\}$ l'ensemble des classes. Etant donné un pattern p pris dans l'ensemble d'apprentissage, la paire $(R(p) = j, e_k(p) = j_k)$ signifie que e_k attribue la classe $j_k \in \Lambda \cup \{N+1\}$ au pattern p dont la vraie classe est $j \in \Lambda$. Lorsque $j_k \in \Lambda$, on dit que e_k accepte le pattern p (reconnaissance correcte si $j = j_k$, ou bien erreur si $j \neq j_k$) ; sinon e_k rejette p. Le problème de la fusion des classifieurs se pose alors en ces termes : "Etant donné un pattern inconnu x (c'est-à-dire un pattern dont la vraie classe $R(x)$ n'est pas connue), comment fusionner les différentes décisions $e_1(x), \cdots, e_K(x)$ pour finalement décider de la classe à attribuer à x" ? Le problème de fusion ainsi posé peut être vu comme un problème de discrimination classique où les classifieurs appris anterieurement joueraient le role de variables nominales à valeurs dans l'ensemble des classes possibles, avec la possibilité d'absence de valeur pour ces variables (cas de rejet). Les solutions les plus généralement adoptées relèvent du cadre bayésien. En notant $\mathbf{D} = (j_1, \cdots, j_K)$ le vecteur des K décisions, $X = (e_1(x), \cdots, e_K(x))$ le vecteur des K variables aléatoires constituées par les classifieurs, et $X \equiv \mathbf{D}$ le fait que les classifieurs $e_1, \cdots, e_K$ prennent respectivement les décisions $j_1, \cdots, j_K$ au sujet d'un pattern inconnu x, l'approche de solution bayésienne définit les fonctions de fusion F et de décision E suivantes :

$$\begin{array}{rcl} F(\mathbf{D}, j) & = & P(R(x) = j \mid X \equiv \mathbf{D}) \\ E(\mathbf{D}) & = & \left\{ \begin{array}{ll} c & \text{si} \quad F(\mathbf{D}, c) > F(\mathbf{D}, j) \quad \forall j \neq c \\ N+1 & \text{sinon} \end{array} \right. \end{array} \qquad [1]$$

$F(\mathbf{D}, j)$ évalue la probabilité à postériori de la classe j étant donné le vecteur de décision $\mathbf{D}$. La règle de décision E correspond à l'idée que la classe c à associer finalement au pattern x est celle qui donne le maximum de probabilité à postériori. La fonction F se réécrit :

$$F(\mathbf{D}, j) = \frac{P(R(x) = j, e_1(x) = j_1, \cdots, e_K(x) = j_K)}{P(e_1(x) = j_1, \cdots, e_K(x) = j_K)}$$

Sous l'hyposthèse d'indépendance entre les classifieurs conditionnellement à la variable R, on peut écrire :

$$\begin{array}{rcl} F(\mathbf{D}, j) & = & P(R(x) = j) \times \prod_{k=1}^{K} P(e_k(x) = j_k \mid R(x) = j) \times \frac{1}{P(e_1(x) = j_1, \cdots, e_K(x) = j_K)} \\ & = & P(R(x) = j) \times \prod_{k=1}^{K} P(e_k(x) = j_k \mid R(x) = j) \times \eta \end{array}$$

où la valeur de η ne dépend pas de j et donc n'influencera pas la décision E. On peut donc écrire :

$$F(\mathbf{D}, j) = P(R(x) = j) \times \prod_{k=1}^{K} P(e_k(x) = j_k \mid R(x) = j) \qquad [2]$$

L'équation [2] correspond à une fusion complète par la *règle bayesienne naïve* (ou règle du produit) dans laquelle les facteurs $P(e_k(x) = j_k \mid R(x) = j)$ et $P(R(x) = j)$ sont estimés à partir des données d'apprentissage pour tout j.

3. Aggrégation stratifiée

La notion de strate de classifieurs nous permet de prendre en compte dans la fusion tous les $2^K - 1$ sous-ensembles de classifieurs. Chaque sous-ensemble définit une strate dont la réponse se présente sous la forme d'un vecteur de probabilités à N composantes. Ces $2^K - 1$ vecteurs de probabilités en sortie des strates sont ensuite fusionnés grâce à une règle de fusion expliquée ci-dessous (voir équation [4]).

Avant d'expliquer la manière dont est calculé le vecteur de probabilités en sortie d'une strate ainsi que la façon dont est fusionné l'ensemble des $2^K - 1$ vecteurs, définissons quelques notations. La strate i ($1 \leq i \leq 2^K - 1$) est notée X_i. Dans la représentation binaire de la valeur décimale de l'indice i, les bits positionnés à "1" indiquent les classifieurs qui sont membres de X_i. Par exemple si les classifieurs membres de X_i sont les quatre classifieurs numéros 2, 3, 6 et 8, alors la représentation binaire de i est 10100110, où le poids des bits décroît lorsqu'on va de gauche à droite (le bit de poids 7 correspond au classifieur numéro 8, le bit de poids 6 correspond au classifieur numéro 7 et ainsi de suite jusqu'au bit de poids 0 qui correspond au classifieur numéro 1). Autrement dit $X_{166} = (e_2, e_3, e_6, e_8)$, le nombre 166 étant la valeur décimale de 10100110. Si $\mathbf{D} = (j_1, \cdots, j_K)$ désigne un vecteur de décisions –de dimension K– au sujet d'un pattern inconnu x, l'expression $\mathbf{D}[166]$ par exemple désigne le sous-vecteur de décisions –de dimension 4– dont les composantes sont les décisions respectives des classifieurs e_2, e_3, e_6, et e_8 à l'intérieur de $\mathbf{D}$. On remarque aisément que $\mathbf{D}[2^K - 1] = \mathbf{D}$ puisque X_{2^K-1} désigne la strate constituée de l'ensemble des K classifieurs.

Dans cet article, nous utilisons une approche bayésienne pour fusionner les classifieurs à l'intérieur d'une strate. Soit une strate X_i. Nous considérons la distribution jointe estimée $P(R, X_i)$ comme étant une fonction traduisant le comportement de la strate X_i face au problème de classification bien connu de la variable R. Nous entendons par "apprentissage" de la fusion stratifiée, l'estimation –sur base des données d'apprentissage– des distributions jointes discrètes $P(R, X_i)$ $\forall i$. Plusieurs méthodes existent pour estimer ces distributions. Nous avons implémenté celle proposée dans [CHO 68] qui est basée sur un calcul de l'information mutuelle entre les classifieurs. Lorsque, pour un pattern inconnu x, les K classifieurs produisent un certain vecteur de décisions $\mathbf{D}$, le vecteur de probabilités qui indique l'opinion de la strate X_i au sujet du pattern x est donné par la fonction $V_i(\mathbf{D}[i], j)$:

$$V_i(\mathbf{D}[i], j) = \frac{P(R(x) = j, X_i \equiv \mathbf{D}[i])}{\sum_{n=1}^{N} P(R(x) = n, X_i \equiv \mathbf{D}[i])} \quad \forall j \in \Lambda \qquad [3]$$

Chaque vecteur V_i ainsi défini est de dimension N (le nombre total de classes). La règle de fusion des $2^K - 1$ vecteurs V_i s'énonce alors :

$$F(\mathbf{D}, j) = \sum_{i=1}^{2^K-1} \delta_{ij} \times V_i(\mathbf{D}[i], j) \qquad [4]$$

où

$$\delta_{ic} = \begin{cases} 1 & \text{si} \quad V_i(\mathbf{D}[i], c) > V_i(\mathbf{D}[i], j) \quad \forall j \neq c \\ 0 & \text{sinon} \end{cases}$$

La règle de décision E, bien que restant mathématiquement identique à celle dans l'équation [1] correspond ici au vote à la majorité relative (majorité simple), pondéré par les poids fournis par les vecteurs V_i.

4. Application à la classification automatique en stades de sommeil

La cotation en stades de sommeil consiste à attribuer un stade de sommeil à chacune des pages successives d'un tracé polysomnographique. On distingue cinq stades de sommeil : l'éveil, le REM (*Rapid Eye Mouvement*), le stade 1, le stade 2 et le stade 3. Nous avons utilisé 6 algorithmes de classification, tous basés sur les réseaux de neurones et les forêts d'arbres aléatoires. Ils diffèrent entre eux par les caractéristiques extraites des pages et le soin apporté à la qualité de la reconnaissance de tel ou tel autre stade de sommeil. En ce qui concerne les données d'apprentissage, autrement dit celles à utiliser pour estimer les distributions jointes intra-strates, nous avons disposé de 18 tracés polysomnographiques cotés en stades, donnant un total de 18068 patterns d'apprentissage. Nous avons effectué des tests avec 8 autres tracés, référencés en colonne dans le TAB.1. On peut lire dans ce tableau, outre les résultats obtenus par une fusion complète (règle dans l'équation [2]) des 6 classifieurs, les résultats obtenus par une fusion stratifiée (règle dans l'équation [4]) des 63 strates formées à partir des 6 classifieurs. Pour chacun des 8 tracés testés, nous faisons figurer également dans le TAB.1 la plus mauvaise performance individuelle observée (deuxième ligne) ainsi que la meilleure performance individuelle observée (troisième ligne). Il est important de noter que le plus mauvais et le meilleur classifieurs individuels sont variables d'un tracé à l'autre et sont inconnus

	Tracé 1	*Tracé 2*	*Tracé 3*	*Tracé 4*	*Tracé 5*	*Tracé 6*	*Tracé 7*	*Tracé 8*
Min perfor. individ.	72,53	77,46	79,76	73,97	56,49	73,27	88,41	84,41
Max perfor. individ.	73,88	79,70	87,30	77,12	76,84	81,51	90,78	88,51
Fusion complète ([2])	74,51	86,09	86,71	86,90	61,06	75,46	88,41	85,54
Fusion stratifées ([4])	76,59	80,41	87,40	80,00	80,79	77,74	90,69	89,54

TAB. 1. *Tableau comparatif des performances (en % de reconnaissance) des classifieurs, de leur fusion directe (règle [2]), et de leur fusion par la méthode d'aggrégation stratifiée (règle [4])*

d'avance. Ils ne sont identifiables qu'à postériori. Dans ce contexte, l'objectif de la fusion stratifiée est de fournir des performances au moins égales à celles du classifieur qui se révélerait le plus performant étant donné un tracé.

On constate dans le TAB.1 que dans les cas où la fusion stratifiée ne fournit pas des performances au delà de celles du meilleur classifieur individuel (tracés 6 et 7), les performances de la fusion stratifiée restent néanmoins plus proches des meilleures performances individuelles que des plus mauvaises performances individuelles. La fusion stratifiée affiche ainsi un comportement plus stable et plus proche de l'objectif que la fusion complète par la règle bayesienne naïve. Cette dernière règle peut passer de très bonnes performances (tracés 2 et 4 pour lesquels la fusion complète donne des performances supérieures à celles de la fusion stratifiée) à de très mauvaises (tracés 5, 6, 7 et 8 pour lesquels la fusion complète donne des performances plus proches de celles du plus mauvais classifieur individuel).

5. Conclusion

Etant donnée un problème de classification et un ensemble de classifieurs qui y répondent, la combinaison stratifiée consiste en l'aggrégation des classifieurs en strates, puis en la fusion –par le vote majoritaire relatif pondéré– des réponses fournies par les strates. Une strate de classifieurs est un sous ensemble de l'ensemble des classifieurs à fusionner. Nous entendons par "réponse fournie par une strate" le résultat de la fusion (par le vote, les fonctions de croyance, des approches bayésiennes, etc.) des classifieurs membres de cette strate. Dans cet article, nous avons expérimenté la combinaison stratifiée de classifieurs pour un problème de cotation automatique en stades de sommeil. Nous nous sommes basés sur l'estimation des distributions jointes pour la fusion des classifieurs à l'intérieur de chaque strate. Les résultats empiriques obtenus montrent que la fusion stratifiée donne souvent des performances au delà de celles du meilleur classifieur individuel identifié à postériori. Lorsque les performances de la fusion stratifiée ne dépassent pas celles du meilleur classifieur individuel, elles restent cependant proches.

6. Bibliographie

[APP 02] APPRIOU A., Discrimination Multisignal par la Théorie de l'Evidence, LAVOISIER, Ed., *Décision et Reconnaissance de Formes en Signal*, p. 219-257, 11, rue Lavoisier, 75008 Paris, 2002.

[CHO 68] CHOW C., LUI C., Approximating Discrete Probability Distributions with Dependence Trees, *IEEE Transaction on Information Theory*, vol. 14, n^o 3, 1968, p. 462-467.

[DIE 00] DIETTERICH T. G., Ensemble Methods in Machine Learning, *MCS '00 : Proceedings of the First International Workshop on Multiple Classifier Systems*, London, UK, 2000, Springer-Verlag, p. 1–15.

[KAN 99] KANG H.-J., LEE S.-W., Combining Classifiers Based on Minimization of Bayes Error Rate, *Fifth International Conference on Document Analysis and Recognition*, 1999, p. 398-401.

[KIT 97] KITTLER J., HATEF M., DUIN R., MATAS J., On Combining Classifiers, *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 20, n^o 3, 1997, p. 226-239.

[LAM 97] LAM L., SUEN C., Application of Majority Voting to Pattern Recognition : an Analysis of its Behavior and Performance, *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 27, n^o 5, 1997, p. 553-568.

Consensus de partitions : une approche empirique

Jean-Baptiste Angelelli, Alain Guénoche

IML 163 Av. de Luminy, 13288 Marseille cedex 9
{angele,guenoche}@iml.univ-mrs.fr

RÉSUMÉ. *Etant donné un profil* Π *de partitions sur un même ensemble* X*, nous cherchons à construire une partition consensus qui contienne un nombre maximum de paires réunies et de paires séparées dans le profil. Pour ce faire, nous définissons une fonction de score* S_Π *associée à toute partition sur* X *et les partitions consensus pour ce profil sont celles qui maximisent cette fonction. Nous présentons deux algorithmes efficaces pour construire une partition sous-optimale.*

MOTS-CLÉS : *Partitions, Consensus*

1. Introduction

Nombreuses sont les situations où l'on dispose d'un profil de partitions sur un même ensemble X : par exemple quand on applique plusieurs méthodes de partitionnement aux mêmes données, quand on part de plusieurs initialisations du même algorithme (K-means) ou quand les éléments de X sont décrits par des variables nominales. C'est ce dernier problème qui est à l'origine des travaux sur le consensus de partitions avec l'article de S. Régnier (1965). Ce problème s'est réactualisé avec le *bootstrap clustering* (Chu et al. 1998) qui consiste à générer, par re-échantillonage ou autre, plusieurs jeux de données à partir des données initiales. A chacun correspond une partition et leur consensus est supposé plus robuste que la partition d'un seul jeu.

Si nous parlons d'une approche *empirique*, c'est que nous n'abordons la question ni sur un plan axiomatique, ni en cherchant à déterminer un élément particulier dans le treillis des partitions (pour une telle approche, voir Barthélemy & Leclerc, 1995). Nous reprenons la démarche usuelle du consensus de X-arbres, à savoir construire un arbre qui possède un nombre maximum de bipartitions du profil (consensus majoritaire étendu). Dans le cas des partitions, on cherche à construire une partition qui possède un nombre maximum de paires réunies et de paires séparées dans le profil. Ceci est équivalent au problème de la *partition médiane*, celle dont la somme des distances aux partitions du profil est minimum. De plus, nous nous attachons à construire une telle partition en présentant deux algorithmes efficaces pour établir une partition sous-optimale.

2. Formalisation du consensus

Soit X un ensemble fini de n éléments, $\mathbf{P}$ l'ensemble de toutes les partitions de X et Π un *profil* de m partitions en classes disjointes, non vides, et dont l'union est égale à X et qui ne sont pas nécessairement distinctes. Pour une partition donnée $\pi \in \mathbf{P}$, tout élément $i \in X$ appartient à la classe notée $\pi(i)$. Dans ce qui suit, δ designe le symbole de Kronecker habituel ; on a donc $\delta_{\pi(i)\pi(j)} = 1$ si i et j sont réunis dans π, $\delta_{\pi(i)\pi(j)} = 0$ sinon.

Etant donné un ensemble X fini et un profil Π, le problème de la *partition consensus* consiste à déterminer une partition $\pi^* \in \mathbf{P}$ qui résume au mieux le profil au sens d'un certain critère. Etant donné deux partitions π et θ de X, nous établissons une mesure de *l'accord* entre π et θ. Sa valeur est d'autant plus élevée qu'un grand nombre de paires d'éléments réunis (resp. séparés) dans π le sont également dans θ. Les partitions sur X étant des relations d'équivalence sur les paires d'éléments de X, il est naturel de mesurer l'écart entre deux partitions par la

distance de la différence symétrique entre ces relations, notée Δ. Nous utiliserons par la suite la notion équivalente de similitude, $S(\pi,\theta) = \frac{n(n-1)}{2} - \Delta(\pi,\theta)$ ou encore :

$$S(\pi,\theta) = \sum_{i<j} \Big(\delta_{\pi(i)\pi(j)}\delta_{\theta(i)\theta(j)} + (1-\delta_{\pi(i)\pi(j)})(1-\delta_{\theta(i)\theta(j)}) \Big) \tag{1}$$

Le score d'une partition P relativement à un profil de partition $\Pi = (\pi_1, \ldots, \pi_m)$ est défini par la somme des similitudes de P relativement à chacune des partitions du profil :

$$S_\Pi(P) = \sum_{k=1}^{m} S(P,\pi_k). \tag{2}$$

Etant donné un profil $\Pi = (\pi_1, \ldots, \pi_m)$, on note T_{ij} le nombre de partitions dans lesquelles deux éléments donnés i et j sont réunis. Dans ces conditions, le score d'une partition P relativement au profil Π peut s'écrire :

$$\begin{aligned} S_\Pi(P) &= \sum_{i<j} \Big(\delta_{P(i)P(j)}T_{ij} + (1-\delta_{P(i)P(j)})(m - T_{ij}) \Big) \\ &= 2\sum_{i<j} \delta_{P(i)P(j)}T_{ij} + \sum_{i<j} m - \sum_{i<j} \delta_{P(i)P(j)}m - \sum_{i<j} T_{ij} + \sum_{i<j} \delta_{P(i)P(j)}T_{ij} \end{aligned}$$

Les quantités $\sum_{i<j} m$ et $\sum_{i<j} T_{ij}$ sont propres au profil Π et ne dépendent pas de P. Ainsi, en éliminant ces deux termes et en multipliant par un facteur $1/2$, maximiser $S_\Pi(P)$ est équivalent à maximiser la quantité :

$$\sum_{i<j} \delta_{P(i)P(j)}T_{ij} - \frac{1}{2}\sum_{i<j} \delta_{P(i)P(j)}m.$$

Si R_P désigne l'ensemble des paires réunies dans P, on obtient un critère équivalent à $S_\Pi(P)$:

$$S'_\Pi(P) = \sum_{(i,j)\in R_P} \Big(T_{ij} - \frac{m}{2} \Big). \tag{3}$$

Le critère S'_Π peut s'interprêter de façon très intuitive. Il signifie que, dans une partition P, une paire éléments de X réunis a une contribution positive (resp. négative) au critère quand ces deux éléments sont réunis dans plus (resp. moins) de la moitié des partitions de Π.

Soit K_n le graphe complet sur X dont les arêtes sont pondérées par $w(i,j) = T_{ij} - m/2$, que nous appellerons *graphe de Régnier*, en hommage à l'auteur, bien qu'il n'ait jamais fait référence aux graphes dans son article. Soit P une partition en p classes $P = (P_1, \ldots, P_p)$ et $Q = (Q_1, \ldots, Q_p)$ l'ensemble des cliques dans K_n correspondant aux classes de P. La quantité $W(P_k) = \sum_{i,j\in P_k} w(i,j)$ est le poids de la clique correspondant à P_k et on a alors

$$S'_\Pi(P) = \sum_{k\in[1..p]} W(P_k). \tag{4}$$

3. Problème d'optimisation

Maximiser S'_Π revient à construire une partition de poids maximum pour W ou encore un ensemble de cliques disjointes dans le graphe de Régnier, qui soit de poids maximum. C'est une extension aux graphes pondérés du problème de Zahn (1971), bien connu pour être NP-difficile, et donc on ne connaît pas d'algorithmes polynomiaux qui donnent une solution optimale.

On a la propriété évidente $\max_{P\in\mathbf{P}} S'_\Pi(P) \geq 0$. En effet, en notant P_0 la partition atomique de X dans laquelle tous les éléments sont séparés, on a $R_{P_0} = \emptyset$ et donc $S'_\Pi(P_0) = 0$. Ainsi, quel que soit le profil Π, il existe toujours

une partition de score nul. Le cas où le maximum de la fonction S'_Π sur l'ensemble $\mathbf{P}$ est atteint pour la partition atomique s'interprête de façon intuitive : les partitions qui composent le profil sont en trop grand *désaccord* pour qu'un consensus non trivial émerge ; la partition consensus est alors la partition atomique.

Plus généralement, soit $E = \{(i,j)$ tels que $w(i,j) \geq 0\}$ et le graphe pondéré $G_+ = (X, E, w)$. Si $E = \emptyset$, la partition atomique est une partition consensus ; elle n'est pas forcément unique, s'il y a des arêtes de poids nul dans E. Par contre, si $E \neq \emptyset$ on est sûr qu'il existe une partition consensus non triviale, qui possède au moins une paire *majoritaire*, dont les éléments sont réunies dans la moitié des partitions du profil. Il existe une autre situation dans laquelle on connait la partition consensus ; c'est la cas où chaque composante connexe de G_+ est une clique. Cette partition est néssairement optimale, puisque toute paire interclasse a une valeur négative.

Comme il est bien signalé dans Régnier (1965), le problème de la partition consensus est un problème d'optimisation discrète que l'on peut résoudre par programmation linéaire en nombres entiers.

3.1. *Méthode exacte : programmation linéaire en nombres entiers*

Etant donné une partition P, en posant $\alpha_{ij} = \delta_{P(i)P(j)}$, le critère S'_Π se réécrit

$$S'_\Pi(\alpha) = \sum_{i<j} \alpha_{ij} w(i,j) \tag{5}$$

avec la contrainte que la relation P est une relation d'équivalence sur X. Le problème d'optimisation revient donc à trouver la matrice α maximisant S'_Π sous les contraintes :

$$\begin{cases} \forall (i,j), \alpha_{ij} \in \{0,1\} \\ \forall (i,j), \alpha_{ij} = \alpha_{ji} \text{ et } \alpha_{ii} = 1 \\ \forall (i,j,k), \alpha_{ij} = \alpha_{jk} = 1 \Rightarrow \alpha_{ik} = 1 \end{cases}$$

Il s'agit d'un problème NP de programmation linéaire en nombres entiers pour lequel il existe des méthodes de resolution exacte qui permettent de trouver α, donc la partition P, réalisant le maximum *global* de la fonction S'_Π sur $\mathbf{P}$. Cette méthode n'est pas polynomiale et donc utilisable dans certaines limites.

3.2. *Méthodes heuristiques*

De nombreuses méthodes heuristiques ont été envisagées, à commencer par une *méthode de transfert* déjà proposée dans Régnier (1965). Elle consiste, en partant d'une partition quelconque, non précisée dans l'article originel, à affecter un élément à une autre classe tant que le critère à optimiser croît. C'est une simple méthode de descente ; de nos jours on lui préfèrerait des méthodes d'optimisation stochastique, comme les méthodes *tabou*. La fonction w pouvant s'interpréter comme une *similarité* sur X, on peut également envisager des méthodes de partitionnement de G_+ (Guénoche, 2008). Dans ce qui suit, nous exposons deux méthodes heuristiques d'optmisation de S'_Π qui donnent d'excellents résultats.

3.2.1. *Méthode ascendante-descendante (AD)*

Dans la première partie, on part de la partition atomique P_0 et, à chaque étape, on réunit les deux classes qui maximisent la valeur de la partition résultante. Le processus s'arrête quand aucune réunion ne permet plus d'accroître le critère.

Dans la seconde partie, on transfère les éléments dont la contribution au critère est négative, et on les affecte soit à une autre classe à laquelle ils contribuent positivement (s'il en est), soit à une nouvelle classe supplémentaire.

3.2.2. *Méthode de fusion-fission (FF)*

Il s'agit d'une méthode de partitionnement de graphes, dont le principe est exposé dans Angelelli & Reboul (2008). C'est une méthode ascendante sur des classes qui peuvent être chevauchantes. On choisit comme classes

initiales les arêtes de G_+, c'est-à-dire la famille $E = \{\{i,j\} \in X$ tels que $w(i,j) \geq 0\}$ qui ne forme pas une partition. Cependant, E réalise le maximum global de S'_Π sur l'ensemble des familles de classes *chevauchantes* de X. A chaque itération, on choisit deux classes de E, que l'on réunit (fusion) ou que l'on sépare pour qu'elles deviennent disjointes (fission) jusqu'à ce qu'une partition stricte soit atteinte. A chaque étape on réalise l'opération pour laquelle la valeur resultante de la fonction S'_Π est maximale.

3.3. *Un protocole de simulation*

Afin de tester et de comparer nos heuristiques, nous avons fait des simulations : on part d'une partition de X à n éléments en p classes équilibrées, c'est la partition initiale du profil. Ensuite, on génère $m-1$ partitions en appliquant à la partition initiale, t transferts aléatoires, dans lesquelles un élément pris au hasard est affecté à une classe de la partition en cours, ou à une nouvelle classe. Pour le premier transfert, on tire une classe au hasard entre 1 et $p+1$ et, si une nouvelle classe a été ajoutée, entre 1 et $p+2$ pour le second tranfert, .. etc. Ainsi les partitions obtenues n'ont pas nécessairement le même nombre de classes.

A valeurs fixées pour n et m, suivant la valeur de t on obtient des profils homogènes dans lesquels la partition initiale est la partition consensus ($t < n/3$) ou des profils très dispersés pour lesquels la partition atomique est le plus souvent la partition consensus ($t > n/3$). de même une forte valeur de m par rapport à n tend à homogénéiser les profils. Afin de conserver des problèmes difficiles, aux solutions non triviales, nous avons pris $m = n$, $p = n/10$ et $t = n/2$.

Sur 500 profils tirés au hasard suivant ces pramètres, nous mesurons la moyenne :
- du score de la meilleure partition du profil ($S_{Best\Pi}$)
- du score de la partition optimale (S_{Opt}),
- du score de la partition calculée par AD (S_{AD}), et
- le pourcentage de problèmes pour lesquels AD a trouvé l'optimum ($\%_{Opt}$).

$n = m$	$S_{Best\Pi}$	S_{Opt}	S_{AD}	$\%_{Opt}$
20	118	183	182	.93
30	-65	204	204	.83
50	-1063	172	171	.79

Pour conclure, soulignons les performances de l'algorithme AD qui, s'il ne trouve pas systématiquement la partition optimale, en construit une dont le score est pratiquement identique.

Bibliographie

Angelelli J.B., Reboul L. (2008) Network modularity optimization by a fusion-fission process and application to a protein-protein interaction networks, *Actes de JOBIM'2008*, 105-110.

Barthélemy J.P., Leclerc, B. (1995) The median procedure for partitions, *DIMACs series in Discrete Mathematics and Thoretical Computer Science*, 19, 3-34

Chu S., DeRisi J. ; Eisen, M., Mulholland J., Botstein D., Brown P.P. (1998) The transcriptional program of sporulation in budding yeats, *Science* 282, 699-705.

Guénoche A. (2008) Comparison of algorithms in graph partitioning, *RAIRO*, 42, 469-484.

Régnier S. (1965) Sur quelques aspects mathématiques des problèmes de classification automatique, *Mathématiques et Sciences humaines*, 82, 1983, 13-29, reprint of *I.C.C. bulletin*, 4, 1965, 175-191.

Zahn C.T. (1971) Graph-theoretical methods for detecting and describing gelstalt clusters, *IEEE Trans. on Computers*, 20.

Analyse en Composantes Principales de Variables symboliques de types histogrammes. (SFC 2009)

Sun Makosso Kallyth, Edwin Diday

CEREMADE
Université Paris Dauphine, Paris,
Pl. du Ml de L. de Tassigny 75775
PARIS Cedex 16 France, France
{makosso, diday}@ceremade.fr

RÉSUMÉ Les histogrammes permettent de résumer l'information d'unités statistiques. Cet article propose d'étendre l'analyse en composantes principales ACP aux tableaux de données pour lesquels la nature des variables est symbolique de type histogrammes. L'extension de l'ACP à de telles variables s'avère très utile et peut servir de cadre pour la réduction de dimension. Dans cet article nous proposons deux nouvelles approches. La première enrichie l'approche de [NAG.et KUM. 2007] en utilisant des outils numériques à l'instar des méthodes proscrutéennes et graphiques. La seconde approche utilise une algèbre d'histogrammes spécifique.

MOTS-CLÉS : variable histogramme, composante principale, produit d'histogramme.

1. Introduction

L'Analyse en Composantes Principales ACP est l'une des plus anciennes technique d'analyse des des données multidimensionnelles. Elle a été introduite indépendamment par Pearson puis Hotelling au début de vingtième siècle. Ses deux principaux objectifs sont de réduire la dimension d'une table de données nxp contenant des variables quantitatives et de créer de nouvelles variables synthétiques. Toutefois, dans la pratique on est de plus en plus confronté à l'étude de tableaux plus complexes. Quand les tableaux à analyser sont munis de règles de taxonomies ou contiennent par exemples, des variables de type intervalles, des variables à valeurs multiples ou des variables à valeur histogramme comme en analyse des données symboliques [DID. et NOIR. 2008], leur analyse requiert l'usage de techniques liées à leur nature. C'est dans cette optique que nait le formalisme des objets symboliques [DID. 1996]. [ROD. et DID. 2000], [NAG.et KUM. 2007], [ICH. 2008] ont proposé des approches permettant d'effectuer une ACP de variables de type histogramme. Cet article propose premièrement une approche intitulée HPCA1 qui complète l'approche de [NAG.et KUM. 2007] basée sur des opérateurs de concaténation d'histogrammes. Ensuite, elle propose une nouvelle méthodologie que nous intitulons HPCA2 qui effectue une ACP d'histogramme basée sur une arithmétique d'histogrammes différente.

2. Notations

On suppose que le nombre d'individus est n ; le nombre de variables égal à p, que le nombre de modalités par variable est égal à m_j. Pour j allant de 1 à p.

$$H= (H_{ij})_{i=1,\ldots,n;\ j=1,\ldots,p} = \begin{pmatrix} H_{11} & \ldots & H_{1p} \\ H_{n1} & \ldots & H_{np} \end{pmatrix}$$

Une variable de type histogramme Y_j est telle que $Y_j =[H_{1j},\ldots,H_{nj}]^t$. Le ième individu est représenté par noté $\omega_i =[H_{i1},\ldots,H_{ip}]$. On suppose que les histogrammes ont le même nombre de modalités i.e que $m_j=m$ quelque soit. En d'autres termes, on suppose que $Hij=[H_{ij}^{(1)},\ldots,Hij^{(m)}]$. Dans le cas de tables usuelles, on généralement une matrice nxp $X=(X_{ij})$ i=1,…,n; j=1…,p et chaque élément de cette matrice contient des valeurs ponctuelles. Soit $M=I_p$ (la matrice identité d'orde p) une métrique d'individu.

3. **Algèbre d'histogrammes.**

On assimile un histogramme à m modalités à un vecteur appartenant à R^m. Soit H le m sous-espace de R^m d'histogrammes. R^m est muni d'une loi additive, d'une soustraction, d'une addition et soustraction de matrices d'histogrammes. En ce qui concerne la définition d'un produit, les possibilités suivantes sont envisageables :

- La Concatenation $H_{ij} \oplus H'_{ij}=(H_{ij}^{(1)},\ldots, H_{ij}^{(m)}) \oplus (H'_{ij}^{(1)},\ldots, H'_{ij}^{(m)})=(H_{ij}^{(1)} xH'_{ij}^{(1)},\ldots., H_{ij}^{(m)} xH'_{ij}^{(m)})$ (1)
- Le produit scalaire:
  $<H_{ij} ,H'_{ij}>=<(H_{ij}^{(1)},\ldots, H_{ij}^{(m)}), (H'_{ij}^{(1)},\ldots, H'_{ij}^{(m)})>= H_{ij}^{(1)} xH'_{ij}^{(1)}+\ldots+ H_{ij}^{(m)} xH'_{ij}^{(m)}= \sum_{k=1,m} H_{ij}^{(k)} xH'_{ij}^{(k)}$ (2)

4. De l'ACP classique à l'ACP d'histogramme.

4.1 Méthodologie de l'ACP classique

On considère la matrice X. de taille nxp. On se donne également une variable $x_{.j}=[x_{1j},\ldots,x_{nj}]^t$, un individu $ei^t=[x_{i1}, \ldots, x_{ip}]$ et une métriqueD=(1/n)I_n (I_n est la matrice identité d'ordre n). Le barycentre est tel que g est tel que is $g= [\mu_1,\ldots, \mu_p]^t =X^tD1_n$ où $1_n=[1,\ldots,1]$ appartient à R^n et R^n and $\mu_j=\sum^n{}_{i=1}$ pi x_{ij} . Soit Y le tableau centré $Y=X-1_ng^t$. La matrice de variance covariance est dans ce cas $V=X^tDX-gg^t=Y^tDY$. Le but de l'ACP ordinaire est maximiser la Trace(VMP) où P est un projecteur orthogonal. La jème composante principale C_j est telle que $C_j=Xu_j$ où u_j représente le jème vecteur propre de VM.

4.2 Méthodologie de [NAG.et KUM. 2007].

Pour étendre ce formalisme à des variable de type histogramme, [NAG.et KUM. 2007] utilisent ce formalisme avec cette fois $W=(1/n)H^t \oplus H-g \oplus g^t=(1/n)Y^t \oplus Y$ où cette fois Y est une matrice d'histogrammes. Le produit de matrice d'histogramme $Y^t \oplus Y$ est spécifié par la concaténation. Dès lors, la matrice de covariance obtenue est une matrice dont les éléments appartiennent non pas à l'espace réel R mais à l'espace R^m. Pour diagonaliser W, [NAG.et KUM. 2007] érigent *m* matrices de covariance ordinaire à partir des modalités respectives des vecteurs. En d'autres termes, W est une sorte de concaténation des matrice de variance covariance $V^{(1)},\ldots, V^{(m)}$ obtenues à partir des premières, deuxième,…, mième modalités. Si à partir de H on érige $H_1,H_2,\ldots,H_m$ les tabeaux classiques nxp contenant les premières, secondes,…, m ième modalités, l'ACP de [NAG.et KUM. 2007] consiste effectue m ACP classiques des tableaux $H_1,H_2,\ldots,H_m$. La jème composante principale généralisée est $F_j=(F_j^{(1)},\ldots, F_j^{(m)})$ où sont les jème composantes principales des tableaux $H_1,H_2,\ldots,H_m$.

4.3 HPCA1 : Extension de la méthodologie de [NAG.et KUM. 2007].

4.3.1 Etude de la dépendance des composantes principale et des variables intiales.

Pour appréhender la dépendance des composantes principales F_l et les variables Y_j on propose le coefficient :

$$Dep(F_l,Y_j)= (Cor(F_l^{(1)},Y_j^{(1)}) +\ldots + Cor(F_l^{(m)},Y_j^{(m)}))/m \quad (3)$$

Où $Cor(F_l^{(k)},Y_j^{(k)})$ représente la corrélation entre la lème composante principale ordinaire du kème tableau H_k et la jème variable de ce même tableau. Il est également possible d'utiliser le coefficient de corrélation vectorielle RV de Escouffier entre F_l et Y_j.

4.3.2 Outil graphique pour la visualisation des composantes.

Pour la visualisation des sorties nous proposons deux graphiques possibles.
A. Premièrement on suggère de représenter dans un tableau les nuages de points obtenus lorsqu'on effectue une association du type $F1^{(k1)}xF2^{(k2)}$ pour k1 et k2 allant de 1à m.

B. D'autre part, la masse d'information obtenue par les m analyse en composantes principales est si importante que un outil de comparaison est indispensable. Cette comparaison peut se faire à partir d'un système d'axes communs aux différents nuages de points associés aux nuages des ACP respectives.Pour ce faire nous proposons deux approches possibles :

1. La première possibilité consiste à déterminer un système d'axes compromis à partir des axes principaux respectifs de $H_1,H_2,\ldots,H_m$. Pour ce faire on détermine le tableau compromis H_{moy}

$$H_{moy}=\mu_1H_1 + \mu_1H_1+\ldots+ \mu_mH_m$$

où μ_k pour k=1,…,m représentent le rapport de l'inertie du tableau H_k divisé par la somme des inerties des tableaux $H_1,H_2,\ldots,H_m$. On détermine ensuite les axes principaux du compromis.

2. La seconde possibilité consiste à utiliser l'analyse proscrutéenne orthogonale et sa généralisation dans le cas multiple (cf. [GOW. 1975]).

4.4 HPCA2: Analyse en composantes principales d'histogrammes via le le produit scalaire d'histogramme.

Dans l'approche de [NAG.et KUM. 2007], on procède par des séries d'ACP au point que la recherche d'outils qui permettent d'établir une connexion entre ces ACP s'impose. Dans cette section, nous proposons une autre approche possible pour construire les composantes principales. Elle s'appuie sur le produit scalaire (relation (2)). Si H est une matrice d'histogrammes centrée (le centrage comme dans le cas usuelle mais modalité par

modalité), la matrice de variance covariance empirique $V=\frac{1}{n}\,H\,{}^{t}H$ est cette fois une matrice pxp ordinaire. Soient u1,…,up les vecteurs propres de V. Soit $Z_\alpha=Hu_\alpha$.le produit d'une matrice d'histogramme et un vecteur appartenant à R^p. On a

$$Z_\alpha=\begin{pmatrix} H_{11} & \dots & H_{1p} \\ H_{n1} & \dots & H_{np} \end{pmatrix}\begin{pmatrix} u_{\alpha 1} \\ u_{\alpha p} \end{pmatrix}=\begin{pmatrix} u_{\alpha 1}H_{11}+\dots & +u_{\alpha p}H_{1p} \\ u_{\alpha 1}H_{n1}+\dots & +u_{\alpha p}H_{np} \end{pmatrix}$$

Les Z_α pour $_{\alpha=1,\dots,p}$ sont des composantes principalesOutil graphique pour la visualisation des composantes. Pour la visualisation des composantes principales, comme précédemment, deux possibilités sont envisageables. Premièrement, on peut représenter dans un tableau les nuages de points obtenus lorsqu'on effectue une association du type $Z_1^{(k1)}xZ_2^{(k2)}$ pour k1 et k2 allant de 1à *m*. D'autre part, pour les composantes généralisées dans un système d'axes communs, l'on n'a guère besoin de construire un système d'axes compromis ou de recourir aux méthodes proscrutéennes. Les outils numériques sont les mêmes que ceux précédemment utilisés.

5. Application.

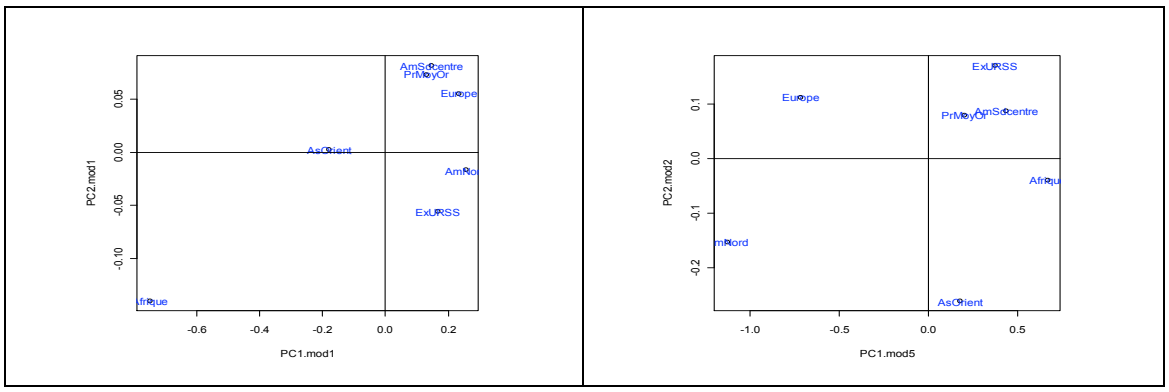


Figure1 : Visualisation de zones selon les axes PC1.mod1xPC2.mod1 et PC1.mod5xPC2.

Dans cette partie nous appliquons les méthodes proposées à un jeu de données réelles publiées par la banque mondiale. Ces données portent sur n=7 parties du monde : l'Afrique, l'Amérique du sud et du centre, l'Amérique du nord, Asie orientale, états de l'ex URSS, l'Europe et Proche et Moyen Orient ; ces régions sont décrites par n=6 variables de type histogramme ayant toute m=5 modalités. Il s'agit des variables Populations, Dépense en matière de Santé, Consommation d'électricité, Consommation des ménages, IDH (Indice de Développement Humain) et PIB. Nous appliquons sur ces données les deux méthodes proposées. La méthode HPCA1 consiste à d'abord appliquée l'approche de [NAG.et KUM. 2007] i.e appliquer *m*=5 ACP classiques. La visualisation d'un extrait de la grille des composantes 1 et 2 donne la figure1. Dans la figure1 quand on considère PC1.mod1 (première modalité de la composante 1) et PC2.mod1 (première modalité de la composante 2) par exemple, on constate que l'Asie orientale et surtout l'Afrique sont les deux zones qui se démarquent des autres. Pour trouver une explication à cette constatation, il est nécessaire d'étudier la figure 2. Cela permet de conclure que l'Afrique et l'Asie par opposition à l'Europe et l'Amérique du Nord sont les régions où on a le plus grand nombre de pays à faible IDH, PIB,…En outre, nous superposons sur le même espace les représentations obtenues à partir des *m* ACP classiques. La figure3 nous donne les superpositions des nuages des points par la méthode proscrutéenne et sur les axes principaux des compromis des différentes ACP effectuées dans la méthode des concaténations. La méthode proscrutéenne fournit une meilleure visualisation. On distingue clairement trois groupes : l'Afrique et l'Asie orientale qui correspond aux pays où les premières modalités sont très fortes. Ce groupe est opposé à l'Amérique du Nord et l'Europe qui ont des modalités 5 très fortes. Enfin, on a troisième groupe, une sorte de groupe intermédiaire qui correspond à l'Amérique du Sud et du Centre, au pays de l'ex URSS et au Proche et Moyen Orient. La méthode HPCA2 basée sur le produit scalaire d'histogramme ressort les mêmes tendances.

6. Conclusion.

Ce travail propose deux méthodologies susceptibles d'étendre l'ACP aux variables de type histogrammes. La première méthodologie complète celle de [NAG.et KUM. 2007] en proposant des outils graphiques et numériques pour traiter la masse d'information que l'on obtient à partir des différentes ACP. Nous proposons également une nouvelle approche basée sur le produit scalaire qui permet d'obtenir des résultats semblables. L'avantage de cette approche est qu'elle permet d'obtenir directement un système d'axes communs.

Figure2 : Dépendance entre composantes principales généralisées et variables histogramme

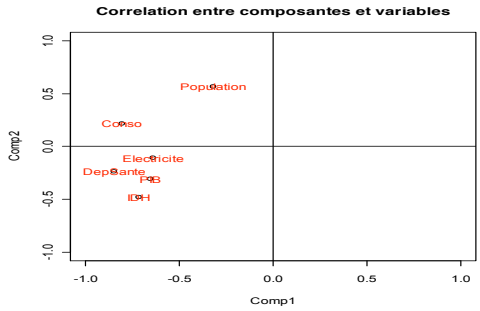


Figure3 : Superposition (resp) par la méthode proscrutéenne et l'axe compromis

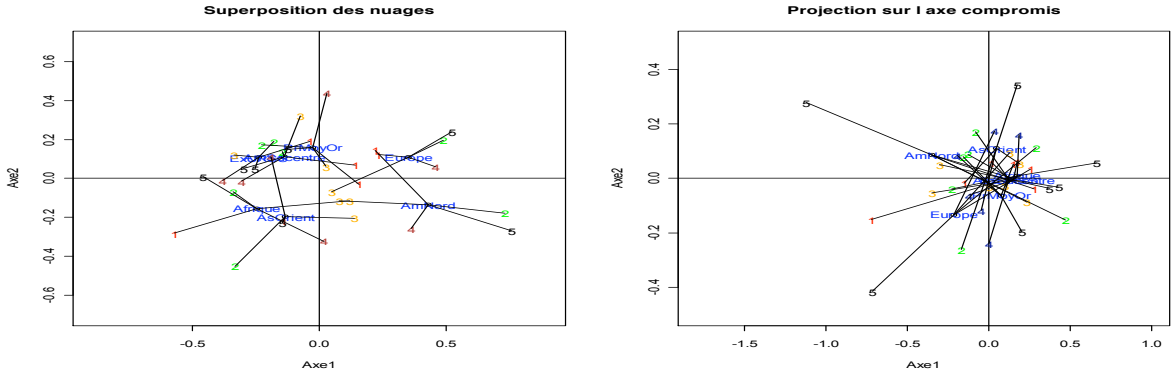
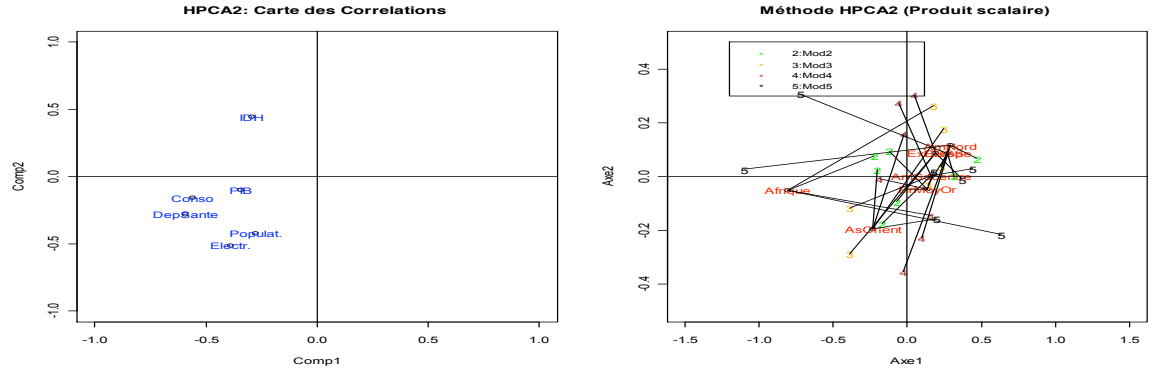


Figure4 : Carte des corrélations et représentation des individus par la méthode HPCA2.



7. Bibliographie

1.Billard, L. (2004): *Dependencies in bivariate interval-valued data*. In Classification, Clustering and New Data Problems (eds. D. Banks,L. House, F.R. McMorris, P. Arabie, and W. Gaul).

2. Cazes P., Chouakria A., Diday E. et Schektman Y. (1997*): Extension de l'analyse en composantes principales a des données de type intervalle"*, Rev. Statistique Appliquee, Vol. XLV Num. 3 pag. 5-24, France.

3. Escouffier, Y. (1973) *Le traitement des variables vectorielles*. Biometrics 29 751–760.

4. E. Diday, M. Noirhomme (eds and co-authors) (2008): *Symbolic Data Analysis and the SODAS software*. 457 pages. Wiley. ISBN 978-0-470-01883-5

5. Diday, E.: (1996) : *Une introduction à l'analyse des données symboliques*, *SFC*,Vannes, France.

6. Ichino M. (2008): *Symbolic PCA for histogram-valued data Proceedings* IASC December 5-8, Yokohama, Japan .

7. Gower (J.C) - *Generalized Procrustes analysis.* Psychometrica, 1975, vol. 40, pp. 33-51. Penin

8. Nagabhsushan P. , Kumar P.(2007): *Principal Component Analysis of histogram Data*. Springer-Verglag Berlin Heidelberg. EdsISNN Part II LNCS 4492, 1012-1021

9. Rodriguez, O., Diday E., Winsberg S. (2001): *Generalization of the Principal Component Analysis to Histogram Data*. Workshop on Symbolic Data Analysis of the 4th European Conference on Principles and Practice of Knowledge Discovery in Data Bases, Septembre 12-16, 2000, Lyon,

Une méthode d'ACP de données en ligne

Jean-Marie Monnez

Institut Elie Cartan, UMR 7502,
Nancy-Université, CNRS, INRIA
BP 239
54506 VANDOEUVRE lès NANCY Cedex, France
Jean-Marie.Monnez@iecn.u-nancy.fr

RÉSUMÉ. Des vecteurs de données arrivant en ligne sont considérés comme des réalisations indépendantes d'un vecteur aléatoire. On établit dans ce cadre un résultat de convergence presque sûre d'un processus d'approximation stochastique des facteurs de l'ACP de ce vecteur aléatoire. On peut l'appliquer par exemple à l'analyse factorielle multiple. On étudie ensuite le cas où l'espérance mathématique du vecteur aléatoire varie dans le temps selon un modèle linéaire.

MOTS-CLÉS : analyse de données en ligne, approximation stochastique, analyse en composantes principales, analyse factorielle multiple.

1. Introduction

On observe p caractères quantitatifs sur n individus : on obtient des vecteurs de données $z_1,\ldots,z_n$ dans R^p. On peut effectuer une ACP du tableau de données. La métrique utilisée, qui dépend des données, est a priori quelconque : on peut souhaiter effectuer par exemple une ACP normée ou une analyse factorielle multiple (AFM) ou une analyse canonique généralisée (ACG).

On considère ici le cas où les vecteurs de données arrivent séquentiellement dans le temps : on observe z_n au temps n. On a une suite de vecteurs de données $z_1,\ldots,z_n,\ldots$.

Supposons dans un premier temps que $z_1,\ldots,z_n,\ldots$ constituent un échantillon i.i.d. d'un vecteur aléatoire Z défini sur un espace probabilisé $(\Omega,\mathcal{A},P)$. Ω représente une population d'où on a extrait un échantillon. On peut définir une ACP de ce vecteur aléatoire (ACPVA), présentée dans le paragraphe 2, qui représente l'ACP effectuée sur la population, dont on va chercher à estimer au temps n les résultats à partir de l'échantillon dont on dispose à ce temps. Soit θ un résultat de l'ACPVA, par exemple une valeur propre, un facteur (on considère ici le cas d'un facteur). On peut effectuer une estimation récursive de θ : disposant d'une estimation θ_n de θ obtenue à partir des observations $z_1,\ldots,z_{n-1}$, on introduit l'observation z_n et on définit à partir de θ_n et z_n une nouvelle estimation θ_{n+1} de θ. On utilise pour cela un processus d'approximation stochastique défini dans le paragraphe 4, dont on établit la convergence. Ce processus est une version stochastique d'une méthode itérative de gradient définie dans le paragraphe 3. On présente des variantes de ce processus dans le paragraphe 5.

Considérons dans un deuxième temps le cas où la loi de Z évolue dans le temps. On étudie dans le paragraphe 6 le cas où l'espérance mathématique de Z varie dans le temps selon un modèle linéaire. On estime simultanément les paramètres du modèle linéaire et le résultat de l'ACPVA par des processus d'approximation stochastique.

2. ACP d'un vecteur aléatoire

Soit un vecteur aléatoire Z dans R^p. R^p est muni d'une métrique M. L'ACP du vecteur aléatoire Z consiste à : 1) rechercher une combinaison linéaire des composantes centrées de Z, $f^1{}'(Z-E(Z))$, f^1 appartenant au dual R^{p*} de R^p, de variance maximale sous la contrainte de normalisation $f^1{}'M^1f^1=1$; 2) rechercher une deuxième

combinaison linéaire des composantes de Z, $f^{2'}(Z-E(Z))$, non corrélée à la première, de variance maximale sous la contrainte $f^{2'}M^{-1}f^{2}=1$; 3) et ainsi de suite jusqu'à un rang r au plus égal à p.

La $i^{ème}$ combinaison linéaire est appelée le $i^{ème}$ facteur ; on appelle également $i^{ème}$ facteur le vecteur f^{i}. Soit

$$C = E\big((Z - E(Z))(Z - E(Z))'\big) = E(ZZ') - E(Z)E(Z')$$

la matrice de covariance de Z. f^{i} est vecteur propre M^{-1} unitaire de MC associé à la $i^{ème}$ plus grande valeur propre.

Si Z a un ensemble fini de N réalisations, l'ACP de Z équivaut à l'ACP usuelle du tableau (N,p) des réalisations, le poids de chaque réalisation étant défini par sa probabilité.

3. Une méthode itérative de détermination des facteurs

On suppose dans ce paragraphe que la matrice de covariance C et la métrique M sont connues.

La fonction $F(x) = \dfrac{< MCx, x >_{M^{-1}}}{< x, x >_{M^{-1}}}$ est maximale pour $x = f^{1}$ et minimale pour $x = f^{p}$, de gradient

$$G(x) = \frac{2M^{-1}}{x'M^{-1}x}\big(MC - F(x)I\big)x.$$

Pour déterminer f^{1}, on peut utiliser un processus de gradient (X_n) défini récursivement par

$$X_{n+1} = X_n + a_n\big(MC - F\big(X_n\big)I\big)X_n\,.$$

Pour déterminer les r premiers facteurs, on peut utiliser le processus suivant :

$$X^{i}_{n+1} = orth_{M^{-1}}(X^{i}_{n} + a_n(MC - F(X^{i}_{n})I)X^{i}_{n}),\ i = 1,...,r.$$

$X^{i}_{n+1} = orth(Y^{i}_{n+1})$ signifie que $(X^{1}_{n+1},...,X^{i}_{n+1})$ est obtenu à partir de $(Y^{1}_{n+1},...,Y^{i}_{n+1})$ par une orthogonalisation de Gram-Schmidt au sens de M^{-1}. En supposant les r plus grandes valeurs propres de MC distinctes, alors, pour $i=1,\ldots, r$, le processus $\left(\dfrac{X^{i}_{n}}{\left\|X^{i}_{n}\right\|_{M^{-1}}}\right)$ converge vers f^{i}, en prenant la suite (a_n) telle que

$$a_n > 0,\quad \sum_{1}^{\infty} a_n = \infty,\quad \sum_{1}^{\infty} \frac{a_n}{\sqrt{n}} < \infty,\quad \sum_{1}^{\infty} a_n^{2} < \infty.$$

4. Approximation stochastique des facteurs

On suppose maintenant que $E(Z)$, C et M sont inconnus et que l'on dispose d'une suite d'observations $(Z_1,\ldots, Z_n,\ldots)$ arrivant dans le temps et constituant un échantillon i.i.d. de Z.

Soit, au temps n, M_n un estimateur de M et Θ_n un estimateur de $E(Z)$ fonctions de $Z_1,\ldots, Z_{n-1}$. Soit

$$B_n = M_n(Z_n Z_n' - \Theta_n \Theta_n'), \quad F_n(X_n^i) = \frac{< B_n X_n^i, X_n^i >_{M_n^{-1}}}{< X_n^i, X_n^i >_{M_n^{-1}}}.$$

On définit le processus d'approximation stochastique :

$$X_{n+1}^i = orth_{M_n^{-1}}(X_n^i + a_n(B_n - F_n(X_n^i)I)X_n^i), i = 1,...,r.$$

Sous les hypothèses précédentes sur la suite (a_n) et les hypothèses complémentaires

$$M_n \xrightarrow{p.s.} M, \quad \sum_1^\infty a_n \|M_n - M\| < \infty \ p.s.,$$

$$\Theta_n - E(Z) \xrightarrow{p.s.} 0, \quad \sum_1^\infty a_n \|\Theta_n - E(Z)\| < \infty \ p.s.,$$

on établit à partir d'un théorème démontré dans [BOU 98] la convergence presque sûre du processus $(\frac{X_n^i}{\|X_n^i\|_{M_n^{-1}}})$ vers f^i pour $i=1,\ldots, r$.

Par exemple, dans le cas de l'analyse factorielle multiple de Z, qui est une ACP de Z avec un choix particulier de métrique M, on peut définir un processus d'approximation stochastique (M_n) convergeant presque sûrement vers M et établir alors la convergence presque sûre en direction du processus $(X_n^1,...,X_n^r)$ vers les r premiers facteurs [MON 06].

5. Variantes

1) Au pas n, on peut utiliser plusieurs observations $Z_{n1},...,Z_{nm_n}$ de Z. On définit alors :

$$B_n = M_n(\frac{1}{m_n}\sum_{k=1}^{m_n} Z_{nk} Z_{nk}' - \Theta_n \Theta_n').$$

2) Au pas n, on peut utiliser toutes les observations faites jusqu'à ce pas. On définit alors :

$$B_n = M_n(\frac{1}{n}\sum_{i=1}^{n} Z_i Z_i' - \Theta_n \Theta_n').$$

6. Cas où l'espérance de Z_n est fonction du temps n

On suppose que l'on dispose d'une suite d'observations $(Z_1,\ldots,Z_n,\ldots)$ arrivant dans le temps telles que $E(Z_n) = \theta_n$ dépende du temps n et que les vecteurs $R_n = Z_n - E(Z_n)$ constituent un échantillon i.i.d. d'un vecteur aléatoire R de matrice de covariance C. Les facteurs de l'ACP de R sont vecteurs propres de MC.

Considérons le cas d'un modèle linéaire d'évolution de l'espérance de Z_n défini de la façon suivante.

Soit $\theta_n^1,...,\theta_n^p$ les composantes de l'espérance θ_n de Z_n. On suppose que pour $k=1,\ldots, p$,

$$\theta_n^k =< \beta^k, U_n^k >, \beta^k \in R^{q_k}, U_n^k \in R^{q_k} ;$$

β^k est un vecteur inconnu et U_n^k un vecteur connu au temps n à q_k composantes.

Pour estimer les paramètres β^k , on utilise les processus d'approximation stochastique (B_n^k) tels que

$$B_{n+1}^k = B_n^k - a_n U_n^k (U_n^{k\prime} B_n^k - Z_n^k), k = 1,..., p.$$

Soit $\Theta_n^k =< B_n^k, U_n^k >, \Theta_n = (\Theta_n^1,...,\Theta_n^p)'$, $B_n = M_n(Z_n Z_n' - \Theta_n \Theta_n')$. On définit le processus $(X_n^1,...,X_n^r)$ comme dans le paragraphe 3 ; on en établit la convergence presque sûre vers les facteurs de l'ACP de R en faisant des hypothèses complémentaires portant sur les U_n^k [MON 08b].

7. Conclusion

Dans le cas où la loi de Z n'évolue pas dans le temps, on a défini un processus d'approximation stochastique des facteurs et donné un résultat général de convergence qui a été appliqué à l'ACP, l'AFM et l'ACG.

Ce résultat étend au cas de plusieurs facteurs et au cas où l'espérance de Z et la métrique M sont inconnues un résultat de convergence vers le premier facteur lorsque l'espérance de Z est connue et la métrique M est l'identité que l'on déduit d'un théorème de Krasulina [KRA 70]. Dans le cas où la métrique M est connue, la méthode d'orthogonalisation a été utilisée par Benzécri [BEN 69] dans le cadre d'un autre processus.

Dans le cas où l'espérance mathématique de Z évolue dans le temps selon un modèle linéaire, on a établi un résultat de convergence qui a été appliqué à l'ACP normée [MON 08b]. Dans une autre étude en préparation, on considère l'application à l'ACG ; on traite également le cas de modèles non linéaires.

On peut mettre en œuvre ces processus pour effectuer des ACP en ligne de données arrivant en ligne.

8. Bibliographie

[BEN 69] BENZECRI J.P., "Approximation stochastique dans une algèbre normée non commutative", *Bulletin de la SMF*, vol. 97, 1969, p. 225-241.

[BOU 98] BOUAMAINE A., MONNEZ J.M., "Approximation stochastique de vecteurs et valeurs propres", *Publications de l'ISUP*, vol. 42, n° 2-3, 1998, p. 15-38.

[KRA 70] KRASULINA T.P., "Method of stochastic approximation in the determination of the largest eigenvalue of the mathematical expectation of random matrices", *Automation and Remote Control*, vol. 2, 1970, p. 215-221.

[MON 06] MONNEZ J.M., "Approximation stochastique en analyse factorielle multiple", *Publications de l'ISUP*, vol. 50, n° 3, 2006, p. 27-45.

[MON 08a] MONNEZ J.M., "Stochastic approximation of the factors of a generalized canonical correlation analysis", *Statistics & Probability Letters,* vol. 78, n° 14, 2008, p. 2210-2216.

[MON 08b] MONNEZ J.M., "Analyse en composantes principales d'un flux de données d'espérance variable dans le temps", *Revue des Nouvelles Technologies de l'Information,* Vol C-2, 2008, p. 43-56.

Régression - corrélation : un point de vue francocentrique sur une lecture de Legendre, Cauchy, Bienaymé, et Carvallo

Antoine de Falguerolles

Université de Toulouse (UPS) – Institut de mathématiques (Laboratoire de statistique et probabilités)
118 route de Narbonne
F-31062 Toulouse Cedex 9
Antoine.Falguerolles@math.toulouse.fr

RÉSUMÉ. *Régression et corrélation, telles que conceptualisées de nos jours, sont indéniablement les résultats remarquables obtenus par des statisticiens britanniques et, notamment, par Francis Galton et Karl Pearson. Pourtant, la lecture de publications de statistique de langue française du XIX^e siècle rend compte d'une grande créativité dont certains prolongements sont toujours d'actualité. Dans cet exposé, ce sont des textes d'Adrien-Marie Legendre, Augustin-Louis Cauchy, Irénée-Jules Bienaymé et Emmanuel Carvallo qui servent de prétexte à rappeler des approches maintenant méconnues en régression. Ceci conduit naturellement à s'interroger sur la notion de corrélation entre deux variables. Curieusement, la ligne de réponse choisie permet de justifier une pratique de l'analyse descriptive multivariée consistant à introduire les carrés de coefficients de corrélation linéaire ou les carrés de coefficients de corrélation linéaire partielle plutôt que leurs valeurs.*

MOTS-CLÉS : *Régression, méthode des moindres carré, méthode de Cauchy, indice de corrélation, analyse des données, histoire.*

1. Introduction

La méthode d'ajustement linéaire de Cauchy est évoquée dans des textes de la fin du XIX^e et début du XX^e siècles. Curieusement, elle l'est encore, mais indirectement, dans le récent *Rapport au ministre de l'Éducation nationale – L'enseignement des sciences mathématiques* (Kahane, 2002) qui rappelle un avis officiel ancien en défaveur de la méthode de Legendre :

> « La complexité de ses calculs effraie les praticiens et Le Verrier interdit même l'enseignement de la méthode des moindres carrés à l'École polytechnique dans la réforme de 1850. »

(voir [KAH 02, page 184]). La méthode de Cauchy ? Questionné, Michel Armatte m'envoyait par retour de courrier une photocopie de la partie du manuel de Lucien March (1930, [MAR 30, pages 479-497]) la décrivant . Mais grâce à la numérisation, bien d'autres textes parmi lesquels le texte fondateur de Cauchy sont facilement accessibles. En effet, c'est vers 1835 qu'Augustin-Louis Cauchy propose une méthode d'ajustement (1836, [Cau 36]). Bien que critiquée par Irénée-Jules Bienaymé (1853, [BIE 53]) qui lui préfère les moindres carrés, elle a connu une certaine vogue en France. Ainsi, Vilfredo Pareto (1897, [PAR 97]) et Lucien March (1898, [MAR 98a] et [MAR 98b]), dans des exposés faits à la Société de Statistique de Paris, se doivent de la considérer en même temps que celle de Legendre (1805, [Leg 05]). Pareillement, Emmanuel Carvallo (1890 [CAR 90], 1912 [CAR 12]), dans des présentations très générales des deux méthodes, Legendre et Cauchy, ne se prononce pas sur leurs valeurs respectives.

Cet exposé vise à rendre compte d'une lecture contemporaine des textes de Legendre, Cauchy, Bienaymé et Carvallo. La méthode de Cauchy est présentée dans le cadre actuel de la régression, cas simple puis multiple. Les deux stratégies d'estimation numérique, alors usuelles, sont rappelées. Le recours possible à une large ménagerie d'estimateurs pour l'estimation du coefficient de pente en régression linéaire simple invite à associer à chaque estimateur un indice de corrélation entre deux variables, celui de Legendre n'étant d'ailleurs que le carré du coefficient

de corrélation linéaire de Pearson (1920, [PEA 20]). L'apparition de ce carré vient de façon inattendue cautionner son introduction dans des méthodes exploratoires multivariées (Falguerolles et Jmel, [Fal 92]).

2. Régressions linéaires simples

Considérons le « modèle » de régression linéaire simple $y_i \approx b_0 + b_1 x_i$ ou, après centrage des variables, $y_{(0)i} = y_i - \overline{y} \approx b_1 x_{(0)i} = b_1(x_i - \overline{x})$ et $b_0 = \overline{y} - b_1\overline{x}$. L'histoire des critères ou des heuristiques numériques d'estimation du coefficient de pente (b_1) est assez bien établie, particulièrement pour les plus anciennes. Il suffit pour cela de se reporter aux ouvrages classiques de Farebrother (1999, [FAR 99]), ou de Stigler (1986, [STI 86]) : la méthode des moyennes de Tobias Mayer, les critères variés considérés par Simon Laplace (norme L_1, minimax, moindres carrés ...), la méthode des moindres carrés publiée par Adrien-Marie Legendre en 1805 et par Carl Friedrich Gauss en 1809 ... Il y a d'ailleurs bien d'autres méthodes : moindres carrés élagués, moindre valeur absolue de la médiane, méthode de Cauchy ... Mais, avec quelques bonnes raisons, l'approche de Cauchy reste assez largement ignorée.

On rappelle ci-après trois méthodes d'estimations du coefficient de pente b_1 dans l'ordre historique où elles sont apparues. Il ressort que toutes trois sont de la forme $\frac{<\psi(\underline{x}_{(0)}),\underline{y}_{(0)}>}{<\psi(\underline{x}_{(0)}),\underline{x}_{(0)}>}$, où $\underline{\bullet}$ désigne un vecteur de $\mathbb{R}^n$, ψ une application de $\mathbb{R}^n$ dans $\mathbb{R}^n$ et $<\bullet,\bullet>$ le produit scalaire usuel dans $\mathbb{R}^n$. L'indicatrice d'une condition $\bullet$ est notée $\mathbf{1}(\bullet)$.

Mayer : $\psi_i(\underline{x}_{(0)}) = \frac{1}{\sum_{\ell=1}^n \mathbf{1}(x_{(0)i}>a)}\mathbf{1}(x_{(0)i} > a) - \frac{1}{\sum_{\ell=1}^n \mathbf{1}(x_{(0)i}<a)}\mathbf{1}(x_{(0)i} < a)$
où $\min_{\ell\in\{1,\ldots,n\}}\{x_{(0)\ell}\} < a < \max_{\ell\in\{1,\ldots,n\}}\{x_{(0)\ell}\}$. Dans la pratique : $a = 0$.
Legendre : $\psi_i(\underline{x}_{(0)}) = \underline{x}_{(0)i}$.
Cauchy : $\psi_i(\underline{x}_{(0)}) \in \{-1, 0, +1\}$ tel que $< \psi(\underline{x}_{(0)}), \underline{x}_{(0)} > \neq 0$. Dans la pratique : $\psi_i(\underline{x}_{(0)}) = \text{signe}(\underline{x}_{(0)i}) = \mathbf{1}(x_{(0)i} > 0) - \mathbf{1}(x_{(0)i} < 0)$.

Propriété 1 : Les estimateurs associés aux trois méthodes ci-dessus sont linéaires et sans biais.

Propriété 2 : La méthode de Cauchy, avec $\psi_i(\underline{x}_{(0)}) = \text{signe}(\underline{x}_{(0)i})$, et la méthode de Mayer, avec $a = 0$, sont des variantes proches : $\frac{\sum_{i=1}^n y_{(0)i}\mathbf{1}(x_{(0)i}>0)+\gamma\sum_{i=1}^n y_{(0)i}\mathbf{1}(x_{(0)i}=0)}{\sum_{i=1}^n x_{(0)i}\mathbf{1}(x_{(0)i}>0)}$ avec $\gamma = \frac{1}{2}$ pour Cauchy et $\gamma = \frac{\sum_{i=1}^n \mathbf{1}(x_{(0)i}>0)}{n}$ pour Mayer.

Propriété 3 : Pour un couple (X, Y) de variables aléatoires de loi binormale, $\frac{E[Y|X>E[X]]-E[Y|X<E[X]]}{E[X|X>E[X]]-E[X|X<E[X]]} = \frac{E[(X-E[X])(Y-E[Y])]}{E[|X-E[X]|^2]} = \frac{E[\text{signe}(X-E[X])(Y-E[Y])]}{E[|X-E[X]|]}$.

3. Deux approches numériques pour une même régression multiple

Considérons le « modèle » de régression linéaire multiple $y_i \approx b_0 + b_1 x_i^1 + \ldots + b_p x_i^p$ soit $y_{(0)i} \approx b_1 x_{(0)i} + \ldots + b_p x_{(0)i}^p$ après centrage des variables. Deux approches pour le calcul des estimations d'une régression multiple ont été considérées au XIX$^{\text{e}}$: résolution d'un système d'équation linéaires défini par les pentes des régressions linéaires simples de tous les couples de variables ou calcul de coefficients de pente entre couples de variables transformées pas à pas (voir Aldrich, 1998 [ALD 98]).
Approche 1 : Elle consiste à résoudre un système linéaire qui, pour les moindres carrés, n'est autre que celui déduit des équations normales. Ce système a pour expression :

$$\left[\begin{array}{c|cccccc} 1 & \overline{x}^1 & \ldots & \overline{x}^j & \ldots & \overline{x}^p \\ \hline 0 & 1 & \ldots & b_1^j & \ldots & b_1^p \\ \vdots & & & \vdots & & \\ 0 & b_i^1 & \ldots & b_i^j & \ldots & b_i^p \\ \vdots & & & \vdots & & \\ 0 & b_p^1 & \ldots & b_p^j & \ldots & 1 \end{array}\right] \left[\begin{array}{c} b_0 \\ \hline b_1 \\ \vdots \\ b_i \\ \vdots \\ b_p \end{array}\right] = \left[\begin{array}{c} \overline{y} \\ \hline b_1^0 \\ \vdots \\ b_i^0 \\ \vdots \\ b_p^0 \end{array}\right]$$

où les b_i^j (resp. les b_i^0) désignent les coefficients de pente obtenus par n'importe quelle méthode dans les régressions linéaires simples des variables x^j, $j = 1, \ldots, p$, sur les variables x^i, $i = 1, \ldots, p$ (resp. de la variable réponse y sur les variables explicatives x^i, $i = 1, \ldots, p$).

Approche 2 : Elle permet encore d'utiliser n'importe quelle méthode d'estimation de la pente dans une régression linéaire simple.

Initialisation : Toutes les variables sont centrées : $\underline{y}_{(0)} = \underline{y} - \bar{y}\underline{\mathbf{1}}$, et les $\underline{x}^j_{(0)} = \underline{x}^j - \overline{x^j}\underline{\mathbf{1}}$ ($j = 1, \ldots, p$).

Pour h **variant de** 1 **à** p : Introduction de la variable $\underline{x}^h_{(h-1)}$

1. Calcul du coefficient b_h^0 de la régression de $\underline{y}_{(h-1)}$ sur $\underline{x}^h_{(h-1)}$, et calcul du vecteur résidu $\underline{y}_{(h)} = \underline{y}_{(h-1)} - b_h^h \underline{x}^h_{(h-1)}$.

2. Si $h < p$, calcul des coefficients b_h^j de la régression de variables $\underline{x}^j_{(h-1)}$ sur $\underline{x}^h_{(h-1)}$, et calcul des vecteurs résidus $\underline{x}^j_{(h)} = \underline{x}^j_{(h-1)} - b_h^j \underline{x}^h_{(h-1)}$ (j variant de $h+1$ à p).

On note que les variables demeurent centrées à chaque itération puisqu'initialement centrées.

Si la procédure est stoppée à l'itération h, les coefficients de la régression sont obtenus en résolvant alors le problème inverse : $b_h = b_h^0$, $b_{h-1} = b_{h-1}^0 - b_{h-1}^h b_h$, $b_{h-2} = b_{h-2}^0 - b_{h-2}^{h-1} b_{h-1} - b_{h-2}^h b_h \ldots$ Les estimations ainsi obtenues dépendent en général de l'ordre dans lequel les variables explicatives sont introduites. À l'issue de l'étape h, l'examen du vecteur $\underline{y}_{(h)}$ peut éclairer la décision d'arrêt de la procédure. À chaque étape, les régressions linéaires simples effectuées sont donc **dégagées de l'influence des variables explicatives déjà introduites**.

Propriété 4 : Si l'on utilise les estimateurs usuels des moindres carrés, alors les deux approches sont équivalentes.

4. Indices de corrélation linéaire entre deux variables

Au vu des résultats précédents, pourquoi ne pas définir un indice de corrélation linéaire entre deux variables par le produit des pentes de la régression, après centrage, de chacune des deux variables sur l'autre : $c(X,Y) = b_X^Y b_Y^X$. À la ménagerie des estimateurs de pente correspond alors celle des indices de corrélation linéaire ! Dans le cas des estimateurs linéaires de Mayer-Cauchy et de Legendre, on obtient respectivement :

Mayer-Cauchy : $c(X,Y) = \frac{(E[Y|X>E[X]]-E[Y|X<E[X]])(E[X|Y>E[Y]]-E[X|Y<E[Y]])}{(E[X|X>E[X]]-E[X|X<E[X]])(E[Y|Y>E[Y]]-E[Y|Y<E[Y]])}$
et $c(X,Y) = \frac{(E[signe(X-E[X])(Y-E[Y])]E[signe(Y-E[Y])(X-E[X])]}{(E[|X-E[X]|]E[|Y-E[Y]|])}$ respectivement.

Legendre : $c(X,Y) = \frac{(E[(X-E[X])(Y-E[Y])]E[(Y-E[Y])(X-E[X])]}{(E[|X-E[X]|^2]E[|Y-E[Y]|^2])} = \rho^2_{XY}$.

Propriété 5 : Ces indices vérifient les trois propriétés suivantes : i) $c(X,Y) \in [-1,+1]$ ($c(X,Y) \in [0,+1]$ pour Legendre) ; ii) $X \perp Y \Rightarrow c(X,Y) = 0$; iii) $c(a+bX, c+dY) = c(X,Y) \quad \forall b,d$ tels que $bd \neq 0$.

5. Propriété des carrés de coefficients de corrélation linéaire de Pearson

D'après ce qui précède, le carré du coefficient de corrélation linéaire de Pearson pourrait être l'indice de corrélation linéaire le plus connu à prendre en compte [1]. Serait-ce une pratique recommandable ? En analyse descriptive multivariée des données, la matrice des coefficients de corrélation de Pearson est souvent considérée comme une matrice de ressemblance. Ainsi, à valeur absolue égale du coefficient, deux variables X et Y négativement corrélées seront considérées comme moins ressemblantes que les variables X et $-Y$ positivement corrélées. Pourtant, dans certains problèmes, la similarité ne dépend pas du signe mais seulement de l'importance de la liaison.

1. Dans cet esprit, il convient de rappeler ici l'identité entre le Φ^2 d'indépendance de Cramer d'une table de contingence 2×2 croisant deux variables qualitatives et le carré du coefficient de corrélation linéaire de Pearson calculé en affectant des valeurs numériques arbitraires aux modalités de ces deux variables.

Valeur absolue ou carré, que recommander ? On sait que la matrice des valeurs absolues des coefficients de corrélation linéaire peut ne pas être semi définie positive. À l'opposé, la matrice des carrés des coefficients de corrélation linéaire est toujours semi définie positive. C'est une propriété bien connue selon laquelle le produit de Hadamard de deux matrices semi définies positives est une matrice semi définie positive. La matrice ainsi construite se prête donc bien à des méthodes d'analyse descriptive multivariée telles que le positionnement multimensionnel métrique, la représentation des variables en analyse en composantes principales, la classification de variables.

Et les corrélations conditionnelles ? Soit $R = [r_{ij}]$ une matrice de corrélation supposée inversible. On sait que la corrélation $r_{ij.\overline{ij}}$ entre les variables i et j ($i \neq j$) conditionnellent aux autres variables, notées $\overline{ij}$, vaut $-\frac{r^{ij}}{\sqrt{r^{ii}r^{jj}}}$ où $R^{-1} = [r^{ij}]$. Le problème des signes se pose à nouveau pour cette matrice qui, circonstance aggravante, n'est pas en général semi définie positive. Il n'est pas non plus assuré que la matrice des valeurs absolues de ces coefficients le soit. Toutefois, la matrice des carrés des coefficients de corrélation conditionnelle l'est dans tous les cas ([Fal 92]). Cette matrice se prête alors aux méthodes de l'analyse des données multivariées en permettant maintenant de représenter des proximités entre variables conditionnellement aux autres, hors effet signe.

6. Bibliographie

[ALD 98] ALDRICH J., Doing least-squares : perspectives from Gauss and Yule, *International Statistical Review*, vol. 66, n^o 1, 1998, p. 61-81.

[BIE 53] BIENAYMÉ J., Remarques sur les différences qui distinguent l'interpolation de M. Cauchy de la méthode des moindres carrés, et qui assurent la supériorité de cette méthode, *Comptes rendus de l'Académie des sciences*, vol. tome 37, 1853, p. 5-13, Bachelier.

[CAR 90] CARVALLO E., Mémoire sur l'optique : influence du terme de dispersion de Briot sur les lois de la double réfraction, *Annales scientifiques de l'École Normale Supérieure*, vol. 3^e série, tome 7 , 1890, p. 3-123.

[CAR 12] CARVALLO E., *Le calcul des probabilités et ses applications*, vol. 3^e série, tome 7 , Gauthier-Villars, Paris, 1912.

[Cau 36] CAUCHY A.-L., *Mémoire sur la dispersion de la lumière par M. A. L. Cauchy*, J. G. Calve, Prague, 1836, publié par la Société royale des Sciences de Prague en Bohème.

[Fal 92] DE FALGUEROLLES A., JMEL S., Modèles graphiques gaussiens et analyse en composantes principales. Complémentarité et choix de variables, rapport n^{o}03-92, 1992, Université de Toulouse III (UPS), Publications du Laboratoire de statistique et probabilités.

[FAR 99] FAREBROTHER R. W., *Fitting linear relationships. A history of the calculus of observations 1750-1900*, Springer-Verlag, New York, 1999.

[KAH 02] KAHANE J.-P., *Rapport au ministre de l'Éducation nationale. L'enseignement des sciences mathématiques*, Odile Jacob & CNDP, Paris, 2002.

[Leg 05] LEGENDRE A.-M., *Nouvelles méthodes pour la détermination des orbites des comètes*, Firmin Didot, Paris, 1805, avec un *appendice sur la méthode des moindres quarrés* (72-80).

[MAR 98a] MARCH L., Quelques exemples de distribution des salaires. Contribution à l'étude comparative des méthodes d'ajustement (I), *Journal de la Société de Statistique de Paris*, vol. 1re série, Volume 39 , n^o juin, 1898, p. 193-206.

[MAR 98b] MARCH L., Quelques exemples de distribution des salaires. Contribution à l'étude comparative des méthodes d'ajustement (II), *Journal de la Société de Statistique de Paris*, vol. 1re série, Volume 39 , n^o juillet, 1898, p. 241-248.

[MAR 30] MARCH L., *Les principes de la méthode statistique*, Librairie Félix Alcan, Paris, 1930.

[PAR 97] PARETO V., Quelques exemples d'application des méthodes d'interpolation à la statistique, *Journal de la Société de Statistique de Paris*, vol. 1re série, Volume 38 , n^o novembre, 1897, p. 367-379.

[PEA 20] PEARSON K., Notes on the history of correlation, *Biometrika*, vol. 13, n^o October, 1920, p. 25-45.

[STI 86] STIGLER S. M., *The history of statistics : the measurement of uncertainty before 1900*, Harvard University Press, Cambridge, Mass, 1986.

Classification non-supervisée de données multi-représentées par une approche collaborative

Guillaume Cleuziou, Matthieu Exbrayat, Lionel Martin, Jacques-Henri Sublemontier

LIFO (Laboratoire d'Informatique Fondamentale d'Orléans,
Batiment IIIA, Rue Léonard de Vinci, B.P. 6759 F-45067 ORLEANS Cedex 2
prenom.nom@univ-orleans.fr

RÉSUMÉ. *Nous nous intéressons dans cette étude à la classification non-supervisée de données multi-représentées, i.e. des données décrites par plusieurs sources d'information. Nous proposons un cadre méthodologique permettant la recherche d'une classification réalisant un consensus entre les différentes représentations. Nous nous inspirons d'un travail récent de Bickel et Sheffer visant à étendre les modèles de mélanges au cas des données multi-représentées (Co-EM) et proposons un modèle de classification floue basé sur K-moyennes. Les expérimentations proposées valident l'étude sur un jeu de données adapté.*

MOTS-CLÉS : *Classification automatique, K-moyennes flou, Données multi-représentées.*

1. Introduction

Nous nous intéressons ici à des données multi-représentées, c'est à dire à un même ensemble d'individus décrits par plusieurs représentations. De telles données se rencontrent en Recherche d'Information (contenus textuels, hyper-textuels, images, audio et vidéo pour un même document), en Analyse de gènes (expression, localisation, profil phylogénétique pour un gène), ou encore en analyse de données textuelles (lexique, morphosyntaxe et sémantique).

Dans le cadre de la classification non-supervisée il s'agit d'organiser l'ensemble des individus en classes d'individus similaires de manière à réaliser un consensus sur un ensemble des représentations. Deux stratégies naturelles pour procéder au regroupement sur données multi-représentées consistent à fusionner des informations en amont (fusion *a priori*) ou en aval (fusion *a posteriori*) d'un processus traditionnel de clustering. La première pose les problèmes du mélange des niveaux de description et de la malédiction de dimensionalité. L'enjeu pour la seconde consiste à faire correspondre dans une fusion finale des organisations locales pouvant s'avérer très différentes.

Nous choisissons d'explorer une troisième voie, l'approche collaborative, consistant à réaliser la fusion des informations au cours du processus de classification. Pour cela nous nous inspirons des travaux récents de [PED 02] : Co-FC et [BIC 05] : Co-EM.

2. Le modèle CoFKM : Co-k-moyennes floues

L'approche que nous proposons ici est une extension de la méthode des k-moyennes floues ou FKM ([BEZ 81]) qui vise à obtenir, dans chaque représentation, une organisation spécifique, et qui comme dans Co-EM, favorise des organisations "proches" sur les différentes représentations, par l'introduction d'un terme de désaccord.

2.1. *Le critère à optimiser*

Dans la suite, R dénote l'ensemble des représentations. Dans chaque représentation $r \in R$, les individus sont décrits par un vecteur appartenant à $\mathbb{R}^{N_r}$, où N_r est la dimensionnalité de r.

Nous proposons une approche collaborative qui consiste à minimiser les valeurs des inerties de FKM dans chaque représentation, tout en pénalisant les divergences d'organisation entre chaque couple de représentations. Nous définissons le critère suivant à minimiser :

$$Q_{CoFKM} = \sum_{r \in R} \sum_{x_i \in X} \sum_{k=1}^{K} u_{i,k,r}^{\beta} ||x_{i,r} - c_{k,r}||^2 + \eta \Delta$$

avec $\forall x_i \in X, r \in R, \sum_{k=1}^{K} u_{i,k,r} = 1$, où les variables du problème sont les centres des groupes dans chaque représentation $c_{k,r}$ et les degrés d'appartenance $u_{i,k,r}$ de l'individu i au groupe k dans la représentation r. β est un paramètre de flou et η est un paramètre qui module l'importance du désaccord.

Afin d'avoir des inerties comparables dans chaque représentation, il est nécessaire de réaliser une normalisation qui consiste -pour chaque représentation- à réduire chaque variable (variance à 1) et à les pondérer par $N_r^{-1/2}$.

Δ est un terme de désaccord tel que si toutes les représentations produisent la même organisation, ce terme doit être nul. Les centres $c_{k,r}$ n'étant pas comparables d'une représentation à l'autre, nous proposons :

$$\Delta = \frac{1}{|R| - 1} \sum_{r \neq r'} \sum_{x_i \in X} \sum_{k=1}^{K} (u_{i,k,r'}^{\beta} - u_{i,k,r}^{\beta}) ||x_{i,r} - c_{k,r}||^2$$

Dans la mesure où $u_{i,k,r}$ est d'autant plus grand que $||x_{i,r} - c_{k,r}||^2$ est petit, l'expression obtenue pour un couple (r, r') peut donc être vu comme une mesure de divergence entre les organisations obtenues dans r et r'.

Nous pouvons également montrer que l'expression de Q_{CoFKM} peut être réécrite sous une forme plus simple :

$$Q_{CoFKM} = \sum_{r \in R} \sum_{x_i \in X} \sum_{k=1}^{K} u_{i,k,r,\eta} ||x_{i,r} - c_{k,r}||^2 \quad (1)$$

avec $u_{i,k,r,\eta} = (1 - \eta) u_{i,k,r}^{\beta} + \frac{\eta}{|R|-1} (\sum_{\overline{r}} u_{i,k,\overline{r}}^{\beta})$. Le critère s'écrit donc comme une inertie pondérée, où dans chaque représentation, le poids $u_{i,k,r,\eta}$ est une moyenne pondérée entre les poids usuels ($u_{i,k,r'}^{\beta}$) de chaque représentation.

2.2. *Recherche d'une solution optimale*

Nous souhaitons obtenir une solution qui minimise le critère global Q_{CoFKM} sous la contrainte : $\forall r \in R, x_i \in X : \sum_{k} u_{i,k,r} = 1$. La résolution de ce problème d'optimisation sous contraintes par un lagrangien donne :

$$c_{k,r} = \frac{\sum_{x_i \in X} u_{i,k,r,\eta} x_{i,r}}{\sum_{x_i \in X} u_{i,k,r,\eta}} \text{ avec } u_{i,k,r,\eta} = (1 - \eta) u_{i,k,r}^{\beta} + \frac{\eta}{|R| - 1} \sum_{\overline{r}} u_{i,k,\overline{r}}^{\beta} \quad (2)$$

$$u_{i,k,r} = \frac{((1 - \eta) ||x_{i,r} - c_{k,r}||^2 + \frac{\eta}{|R|-1} \sum_{\overline{r}} ||x_{i,\overline{r}} - c_{k,\overline{r}}||^2)^{1/(1-\beta)}}{\Sigma_{k=1}^{K} ((1 - \eta) ||x_{i,r} - c_{k,r}||^2 + \frac{\eta}{|R|-1} \sum_{\overline{r}} ||x_{i,\overline{r}} - c_{k,\overline{r}}||^2)^{1/(1-\beta)}} \quad (3)$$

À chaque étape de l'algorithme, ces mises à jour assurent la minimisation du critère via la variable considérée et les autres variables fixées. Ceci garantit la convergence de l'algorithme.

Algorithm 1 CoFKM

Entrée : Ensemble d'individus X, Nombre de groupes K, Nombre de représentations $|R|$
Initialiser les K centres de groupe de manière commune pour chaque représentation r.
repeat
  for $r = 1$ **to** $|R|$ **do**
    for $k = 1$ **to** K **do**
      mettre à jour $c_{k,r}$ en utilisant l'équation (2)
      for $i = 1$ **to** $|X|$ **do**
        mettre à jour $u_{i,k,r}$ en utilisant l'équation (3)
      end for
    end for
  end for
until convergence
Affectation de l'individu x_i au groupe C_k si $k = \underset{h}{argmax} \sqrt[|R|]{\prod_{r \in R} u_{i,h,r}}$

Afin d'obtenir un unique regroupement, une règle d'affectation (tenant compte des différents degrés d'appartenance obtenus) est nécessaire. Nous proposons une règle simple qui consiste à calculer un degré global correspondant à la moyenne géométrique des degrés locaux $u_{i,k,r}$. Les individus sont alors affectés au groupe pour lequel leur degré d'appartenance global est le plus élevé.

Un récapitulatif des différentes mise à jour, initialisation et affectation est décrit dans l'algorithme 1

2.3. *Comparaisons*

Nous montrons dans un premier temps que notre modèle CoFKM généralise le modèle FKM appliqué à la concaténation des représentations (fusion *a priori*), puis dans un second temps, qu'il généralise également un modèle simple de fusion *a posteriori*, où le modèle FKM est appliqué indépendamment sur chaque représentation.

- Posons $\eta = \frac{|R|-1}{|R|}$. Dans ce cas, $u_{i,k,r,\eta} = \frac{1}{|R|} u_{i,k,r}^{\beta} + \frac{1}{|R|} \sum_{\overline{r}} u_{i,k,\overline{r}}^{\beta}$ correspond exactement à $u_{i,k}^{\beta}$ obtenu si les représentations avaient été concaténées. Par conséquent l'expression de $c_{k,r}$ correspond exactement à celle obtenue par FKM à partir de la concaténation des représentations. De la même manière, $||x_{i,r} - c_{k,r}||^2 + \sum_{\overline{r}} ||x_{i,\overline{r}} - c_{k,\overline{r}}||^2$ correspond à la distance entre x_i et c_k pour la concaténation des représentations. L'expression de $u_{i,k,r}$ correspond donc exactement à celle obtenue par FKM à partir de la concaténation des représentations.
- Posons maintenant $\eta = 0$, le terme de désaccord devient nul. Le critère obtenu correspond bien à optimiser les inerties locales à l'aide du modèle FKM.
  Nous avons observé empiriquement que la positivité du désaccord dépend de la valeur attribuée à η. Si $\eta > \frac{|R|-1}{|R|}$, alors le désaccord exprimé devient négatif. Nous suggérons donc, pour rester cohérent [1] , de choisir $0 \leq \eta \leq \frac{(|R|-1)}{|R|}$.

3. Expérimentations

Nous validons notre approche sur un jeu de données adapté : *multiple features* [2] . Nous observerons les apports empiriques de notre approche. Pour ces tests les paramètres sont : $\beta = 1.25$ et $\eta = \frac{|R-1|}{2|R|}$ (à mi-chemin entre fusions *a priori* et *a posteriori*).

1. un désaccord négatif ayant peu de sens.
2. disponible à l'adresse *http ://archive.ics.uci.edu/ml/*

	Précision (%)		Rappel (%)		FScore (%)	
CoFKM	91.95		92.07		**92.01**	
CoEM(v1)	68.44		73.96		71.07	
CoEM(v2)	27.04		64.71		37.87	
CoEM(v3)	78.80		82.83		80.73	
FKMconcat	90.20		91.75		90.93	
EMconcat(v1)	55.57		64.03		59.44	
EMconcat(v2)	20.11		70.21		31.03	
EMconcat(v3)	71.90		85.47		77.78	
	FScore (%)					
	fac	fou	kar	mor	pix	zer
FKM	66.88	33.65	23.03	55.69	71.52	42.36
EM(v1)	61.99	44,33	58.81	47.93	44.39	36.78
EM(v2)	21.18	18.13	19.25	38.42	21.75	18.55
EM(v3)	21.18	18.13	19.25	38.42	21.75	18.55

TAB. 1. *Comparaisons des modèles collaboratifs et de fusion a priori en utilisant les* 6 *représentations et résultats sur chaque vue indépendamment.*

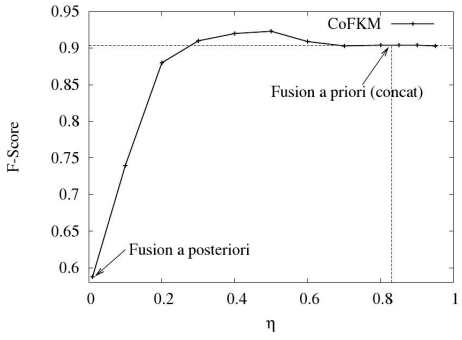


FIG. 1. *Influence du paramètre η sur le modèle CoFKM.*

Nous avons choisi ici d'utiliser l'information externe des étiquettes de classes pour calculer le F-score (ou F-mesure), critère d'évaluation externe classique, permettant de mesurer la pertinence d'un clustering par rapport à une classification de référence. Les résultats que nous obtenons correspondent à une moyenne de 100 exécutions pour lesquelles les méthodes sont comparées avec les mêmes initialisations. Un récapitulatif est présenté dans le tableau Tab. 1. On constate d'une part, que CoFKM surpasse les approches utilisant une unique représentation (quelque soit cette représentation) ; d'autre part, que CoFKM surpasse également l'approche Co-EM pour un mélange de gaussiennes dans 3 versions différentes (v1)-matrices de variances/covariances diagonales-, (v2)-matrices de la forme $\sigma_k.I$, différentes pour chaque composante du mélange-, (v3)-matrices de la forme $\sigma.I$, identiques pour chaque composante du mélange-. Nous avons également exploré les divers résultats obtenus selon différentes valeurs de η pour souligner l'intérêt que peut avoir le choix d'un "bon" η. La courbe (Fig.1) présente la performance du modèle selon différentes valeurs de η. On observe que l'on peut dépasser les résultats obtenus par la simple concaténation des représentations en choisissant convenablement η, et que l'on obtient toujours de meilleurs résultats que la simple fusion *a posteriori*. De manière empirique nous proposerons de choisir $\eta = \frac{|R|-1}{2.|R|}$ à égale distance d'un modèle de fusion *a posteriori* ($\eta = 0$) et d'un modèle de fusion *a priori* ($\eta = \frac{|R|-1}{|R|}$). Il s'agit du choix effectué pour les expérimentations précédentes.

4. Conclusion et perspectives

Le modèle collaboratif proposé généralise différentes solutions de fusion, permet de lui associer une solution algorithmique efficace (convergente) et se compose de peu de paramètres. Les premiers résultats expérimentaux viennent confirmer les observations théoriques précédentes, et les intuitions sur l'intérêt de cette approche au regard des solutions classiques de fusion.

Parmi les orientations futures de ce travail, nous envisageons de proposer d'autres formalisations du désaccord et de proposer une version "noyau" du modèle CoFKM (cf. [KUL 05]) pour permettre un traitement semi-supervisé de ces données. Enfin nous complèterons les expérimentations avec d'autres jeux de données adaptés.

5. Bibliographie

[BEZ 81] BEZDEK J., *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum, New York, 1981.

[BIC 05] BICKEL S., SCHEFFER T., Estimation of mixture models using Co-EM, *Proceedings of the ICML Workshop on Learning with Multiple Views*, 2005.

[KUL 05] KULIS B., BASU S., DHILLON I., MOONEY R., Semi-supervised graph clustering : a kernel approach, *ICML '05 : Proceedings of the 22nd international conference on Machine learning*, New York, NY, USA, 2005, ACM, p. 457–464.

[PED 02] PEDRYCZ W., Collaborative fuzzy clustering, *Pattern Recogn. Lett.*, vol. 23, n^o 14, 2002, p. 1675–1686, Elsevier Science Inc.

Classification floue de données intervallaires: Application au pronostic du cancer.

L. Hedjazi[1,2], T. Kempowsky-Hamon[1,2], M.-V. Le Lann[1,2], J. Aguilar-Martin[1,2].

[1]CNRS ; LAAS ;
7, avenue du Colonel Roche
F-31077 Toulouse, France
{ lhedjazi, tkempows, mvlelann, aguilar }@laas.fr
[2] Université de Toulouse ;UPS, INSA,INP, ISAE ; LAAS ;
F-31077 Toulouse, France

RÉSUMÉ. Afin de prendre en compte les différentes incertitudes/bruits ou réduire les grandes bases de données, la représentation par des données de type intervalle a connu une large utilisation ces dernières années. Cet article propose une approche de classification floue permettant de traiter simultanément ce type de données avec des données qualitatives. Pour cela, une mesure de similarité pour estimer le degré de similarité entre deux individus de type intervalle, représentés par des ensembles flous, est proposée. Dans cette méthode, les paramètres des classes sont représentés eux-aussi par des intervalles sans privilégier un point quelconque de l'intervalle (contrairement à ce qui est souvent fait en utilisant la moyenne par exemple). Dans le cas de données multi variables (soit de même type ou de types différents) et pour un traitement similaire de tous les types de données, une agrégation floue est effectuée. Celle-ci se base sur le concept des connectives mixtes, dont l'interpolation pour calculer l'appartenance d'un individu à chaque classe est réalisée par un index d'exigence. Cette méthode a été appliquée au cas du pronostic du cancer du sein.

MOTS-CLÉS : Analyse des données multi-variables, classification floue, similarité, pronostic du cancer.

1. Introduction

Dans la problématique du traitement du cancer, deux objectifs peut être distingués : le diagnostic du cancer et le pronostic de survie dans le cas d'un cancer préalablement diagnostiqué. Ces travaux s'attachent à développer un outil d'aide au pronostic permettant de mieux évaluer les possibilités de rechute ou de non rechute en fonction de données issues d'analyses ana-cyto-pathologiques de tissus. Dans le domaine du pronostic, les méthodes classiques de classification ont montré une réelle difficulté à obtenir des prédictions correctes en termes de rechute ou de non rechute [HOL 93]. Ceci est dû en tout premier lieu au fait que les bases de données sur le pronostic sont beaucoup moins importantes et que d'autre part, les données contenues dans ces bases sont de natures très différentes, elles peuvent être numériques, qualitatives ou même se présenter sous la forme d'intervalle. Au cours de ces travaux, une extension a été apportée à la méthode de classification floue LAMDA **(Learning Algorithm for Multivariable Data Analysis)** [AGU 82] permettant de traiter les données de type intervallaire en tant que telles et non plus en utilisant la moyenne des bornes de l'intervalle comme le font la plupart des autres méthodes.

2. Traitement des attributs de type intervallaire

Afin de prendre en compte les différentes incertitudes/bruits ou réduire les grandes bases de données, la représentation par des données de type intervalle a connu une large utilisation ces dernières années [BIL 08]. Dans cet article, une approche de classification floue des données intervallaires est proposée. Pour cela une mesure de « *similarité* », qui permet de traiter ce type d'attributs par la méthode LAMDA, a été proposée afin

d'obtenir les *DAMs (Degré d'Adéquation Marginal)* qui seront ensuite agrégés avec les autres attributs même de type différent.

2.1 Description de la similarité

Soient A et B deux intervalles avec $A=[a^{-}a^{+}]$ et $B=[b^{-}b^{+}]$ définis sur un univers de discours *X*, cet univers de discours est discrétisé et donné par la variable *x(n) notée* x_n . Pour calculer la similarité S entre les deux intervalles les données intervallaires sont représentées par des ensembles flous selon la relation :

$$S(A,B)=\frac{1}{2}\left[\left(\frac{\sum_{x_n}(\mu_{A\cap B}(x_n))}{\sum_{x_n}(\mu_{A\cup B}(x_n))}\right)+\left(1-\frac{\sum_{x_n}(\mu_{D}(x_n))}{\sum_{x_n}(\mu_{X}(x_n))}\right)\right] \quad (1)$$

Où D=|max(a^{-},b^{-})-min(a^{+},b^{+})| est la distance entre les deux intervalles A et B dans le cas de non intersection , + et – indiquent respectivement la borne supérieure et inférieure sur X. μ représente la fonction d'appartenance.

La démonstration que cette mesure satisfait les conditions de similarité est omise dans cet article mais il peut être remarqué facilement qu'il s'agit d'un couplage entre deux mesures de similarité : la première qui permet de mesurer la similarité en intersection [GOU 05] et l'autre qui permet de prendre en compte la mesure de similarité en fonction de la distance$[1-d(A,B)]$. Plus deux intervalles sont proches plus ils sont similaires.

2.2 Paramètres des classes

Pour utiliser la mesure définie dans la section précédente pour la classification, les paramètres des attributs de type intervalle pour chaque classe sont définis comme étant eux-aussi de type intervalle. Chaque individu est représenté par un vecteur X_I avec P descripteurs ou attributs $\{x_1,x_2,...,x_P\}$ dont chaque attribut intervallaire $x_i=[x_i^{-},x_i^{+}]$. Ainsi, un attribut *i* est représenté dans la classe *k* par $\rho_k^i=[\rho_{1k}^i,\rho_{2k}^i]$. Le calcul des paramètres de chaque attribut intervallaire se fait exactement de la même manière que pour les données numériques mais au lieu de considérer un attribut mono-dimensionnel, celui-ci est considéré comme bi-dimensionnel et la moyenne est calculée sur deux dimensions. Ceci équivaut à calculer la moyenne de chaque borne (min et max) de l'attribut intervallaire. Dans le cadre de l'apprentissage supervisé, en supposant que la classe *k* contient *m* individus, ses paramètres sont calculés par :

$$\rho_{1k}^i=moy(\sum_{j=1}^{m}x_{ij}^{-}) \quad et \quad \rho_{2k}^i=moy(\sum_{j=1}^{m}x_{ij}^{+}) \quad (2)$$

Où *moy* représente la fonction qui permet de calculer la moyenne, + et - représentent respectivement la borne inférieure et supérieure d'un attribut de type intervalle.

2.3 Choix des fonctions d'appartenance des attributs intervallaires

Dans cette étude, la sélection des fonctions d'appartenance μ_A et μ_B introduites dans la section (2.1) n'est pas abordée. Une fonction uniformément distribuée a été choisie pour représenter l'intervalle afin de ne pas privilégier un point quelconque (comme le centre) par rapport aux autres points de l'intervalle. Néanmoins, il est possible de choisir un autre type de fonction si la distribution des données intervallaires est connue a priori.

2.4 Choix des fonctions d'appartenance de X et D

Le même choix que pour les attributs a été effectué pour représenter μ_X et μ_D , c'est-à-dire une fonction uniformément distribuée. Cependant, lorsque l'univers de discours est assez large il est couteux, en termes de mémoire, de le représenter par une fonction d'appartenance. Pour cela une technique a été proposée afin d'éviter ce problème. Dans le cadre d'un apprentissage supervisé, cette dernière consiste en une simple normalisation des attributs intervallaires qui permet de choisir systématiquement l'univers de discours comme l'intervalle [0,1] et de limiter les attributs dans cet intervalle en effectuant la normalisation suivante (3) :

$$x_i^+ = \frac{\hat{x}_i^+ - x_{iMIN}^-}{x_{iMAX}^+ - x_{iMIN}^-} \quad et \quad x_i^- = \frac{\hat{x}_i^- - x_{iMIN}^-}{x_{iMAX}^+ - x_{iMIN}^-} \qquad (3)$$

Où, pour le descripteur *i* d'un élément, $\hat{x}_i = [\hat{x}_i^-, \hat{x}_i^+]$ est la valeur brute de l'attribut intervallaire et $x_i = [x_i^-, x_i^+]$ est la valeur normalisée. x_{iMAX}^+ et x_{iMIN}^- sont respectivement ses valeurs minimale et maximale.

2.5 Degré d'Adéquation Marginale (DAM) des attributs intervallaires

Le calcul des DAMs des attributs intervallaires se base sur la mesure de la similarité (1) entre deux intervalles introduite en 2.1 et donc l'appartenance de l'attribut x_i à la classe *k* est donnée par :

$$\text{DAM}(x_i, \rho_k^i) = S(x_i, \rho_k^i) \qquad (4)$$

Dans le cas d'attributs qualitatifs, les DAMs sont calculés à partir de la fréquence de chaque modalité [ISA 04]. Une fois que tous les DAMs sont calculés, le concept de connectifs mixtes flous est utilisé pour déterminer l'appartenance globale (DAG) de l'individu à la classe *k*. Ceci reste valable même si les attributs sont de types différents (intervallaire, qualitatif, quantitatif).

3. Base de données utilisée

La base de données utilisée est disponible dans la banque de données de l'Université d'Irvine (UCI) [MUR 95] et provient de l'Institut d'Oncologie du Centre Médical Universitaire de Ljubljana. L'étude concerne le **pronostic de rechute** des patientes atteintes du cancer du sein dont le but est de pronostiquer si les patientes risquent de rechuter au bout de cinq ans. Cette base contient un total de 286 patientes pour lesquelles 201 n'ont pas rechuté et 85 qui ont rechuté [CLA 87] avec des attributs de type mixte : **qualitatif** et sous forme **d'intervalles de valeurs**. Pour ces patientes, 9 attributs sont disponibles (six de nature symbolique, trois de type intervallaire) :

1. Age: 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90-99.
2. Ménopause: inférieur40, supérieur40, pré-ménopause.
3. Taille tumeur: 0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59.
4. Ganglions envahis : 0-2, 3-5, 6-8, 9-11, 12-14, 15-17, 18-20, 21-23, 24-26, 27-29, 30-32, 33-35, 36-39.
5. Ablation ganglions: oui, non.
6. Degré malignité: I, II, III
7. Sein: droit, gauche
8. Quadrant: sup.gauche, inf.gauche, sup.droit, inf. droit, central.
9. Irradiation: oui, non.

Une cross-validation (50% apprentissage, 50% test) a été appliquée afin d'obtenir des estimations exactes sur la précision de notre méthode de classification. Les résultats obtenus sont donnés dans la Table1. Les individus avec des données manquantes ont été exclus de cette analyse (9 individus). Afin de pouvoir comparer les résultats obtenus à ceux cités dans des travaux antérieurs [CLA 87], les 277 patientes ont été tout d'abord classées avec les 9 attributs tels qu'ils étaient donnés dans la base originelle : 6 attributs de type qualitatif dont le **Grade** (**degré de malignité** : descripteur n°6) avec 3 modalités (I, II ou III), et 3 attributs de type intervallaire. Une deuxième partie de l'étude a consisté à traiter ce grade comme une donnée intervallaire. (I: [3,5], II : [6,7] , III : [8,9]) ce qui permet de prendre en compte la progression du degré de malignité comme le font les cancérologues. Les résultats obtenus (91,33% en apprentissage, 90% en test) montrent l'efficacité de cette méthode. Une troisième partie de l'étude a consisté à ne prendre que les patientes n'ayant pas subi encore un traitement par irradiation (215) (traitement appliqué systématiquement dès la présence de ganglions atteints, ce qui sous-entend que les deux attributs : Irradiation et le nombre de ganglions atteints sont corrélés). L'objectif est donc ici la validation de la méthode justement pour aider les médecins sur la décision du traitement adéquat. Le résultat obtenu (3ème ligne de la **Table 1)** est **très satisfaisant**, **93%** de réussite pour l'apprentissage et **92.1%**

pour le test. En comparant avec d'autres techniques, principalement d'induction ou d'arbres de décision, (Assistant [CES 87], CN2 [CLA 87], C4.5 [QUI 93] et EXPLORE [KOR 97]), les résultats obtenus avec la méthode que nous proposons, apparaissent bien supérieurs : comparables aux meilleurs dans le cas de l'apprentissage (100% pour l'AQR [MIC 83], 92% pour Assistant) mais bien meilleurs dans le cas de la reconnaissance (maximum de 72% donné par l'AQR).

TABLE 1 *– Ljubljana 1988 : Résultats de pronostic selon les différents types d'attributs sélectionnés.*

Sélection d'attributs	Apprentissage (50%)	Test (50%)
3 intervallaires ,6 qualitatifs	91%	89.89%
8 attrib.+ Grade intervallaire	91.33%	90%
Patientes non irradiées (8 attributs)	**93%**	**92.1%**

4. Conclusion

Cette étude a démontré que la classification floue fournit des résultats très satisfaisants dans le domaine du pronostic du cancer du sein. Ces résultats ont été grandement améliorés par le développement d'une méthode spécifique au traitement de données de type intervallaire. Une comparaison de ces résultats avec ceux de la littérature montre qu'ils sont soit comparables soit très supérieurs (en particulier concernant le test). Ce type de données n'est pas exclusif au domaine du diagnostic du cancer, cette méthode peut donc être étendue à tout autre domaine en particulier pour le traitement de données entachées d'erreur ou d'incertitude.

5. Bibliographie

[AGU 82] AGUILAR J., LOPEZ DE MANTARAS R., The process of classification and learning the meaning of linguistic descriptors of concepts, *Approximate reasoning in decision analysis*, p. 165-175, 1982.

[BIL 08] BILLARD L., Some analyses of interval data, *Journal of Computing and Information Technology - CIT 16, 4,* p.225–233.,2008.

[CES 87] CESTNIK B., KONONENKO I., BRATKO I., Assistant-86: A knowledge elicitation tool for sophisticated users, *Progress in Machine Learning*. Sigma Press, p. 31-45, Wilmslow, 1987.

[CLA 87] CLARK P., NIBLETT T., Induction in Noisy Domains, *Progress in Machine Learning*, (Proceedings of the 2nd European Working Session on Learning), 1987, p. 11-30. Bled, Yugoslavia: Sigma Press.

[GOU 05] GOUSHUN H., YUNSHENG L., New Subsethood Measures and Similarity Measures of Fuzzy Sets, *IEEE International Conference on Communications, Circuits and Systems*, 2005.

[HOL 93] HOLTE R., Very simple classification rules perform well on most commonly used datasets, *Machine Learning 11 (1)*, 1993, p. 63-90

[ISA 04] ISAZA C., KEMPOWSKY T., AGUILAR-MARTIN J., GAUTHIER A., Qualitative data Classification Using LAMDA and other Soft-Computer Methods. Recent Advances in Artificial Intelligence Research and Development. IOS Press, 2004.

[KOR 97] KORS J.A., HOFFMANN A.L., Induction of decision rules that fulfil user-specified performance requirements. *Pattern Recognition Letters 18 (1997)*, 1187-1195.

[MIC 83] MICHALSKI R., LARSON J., Incremental generation of VL1 hypotheses: the underlying methodology and the description of program AQ11, Dept of Comp. Science report (ISG 83-5), Univ. of Illinois. Urbana. 1983.

[MUR 95] MURPHY P., AHA D., UCI repository of machine learning databases.. University of California. Irvine (http:// www.ics.uci.edu/~mlearn/MLRepository.html), 1995.

[QUI 93] QUINLAN J., C4.5: Programs for Machine learning. Morgan Kaufmann. San Mateo, CA,1993.

Modélisation des dépendances locales entre SNP à l'aide d'un réseau bayésien.

Raphaël Mourad[1], Christine Sinoquet[2], Philippe Leray[1]

[1]*LINA – Ecole Polytechnique de l'Université de Nantes, la Chantrerie, rue Christian Pauc, BP 50609, 44306 Nantes Cedex 3, France*
[2]*LINA – Université de Nantes, 2 rue de la Houssinière, BP 92208, 44322 Nantes Cedex 03, France*
{raphael.mourad,christine.sinoquet,philippe.leray}@univ-nantes.fr

RÉSUMÉ. Dans le cadre d'une étude sur une maladie génétique commune, la dystrophie valvulaire, nous cherchons à modéliser le déséquilibre de liaison existant entre les SNP proches sur l'ADN. Les modèles graphiques probabilistes apparaissent comme des outils intéressants pour une modélisation fine et biologiquement pertinente des dépendances existantes entre les SNP tant au niveau haplotypique que génotypique. La connaissance des haploblocs sur l'ADN nous permet de rendre compte en partie du déséquilibre de liaison présent localement. Cependant, nous montrons que des dépendances entre SNP appartenant à des haploblocs différents sont aussi présentes. A l'aide de réseaux bayésiens incorporant la connaissance des haploblocs sur l'ADN, nous proposons une méthode de modélisation de l'ensemble des dépendances locales du génome.

MOTS-CLÉS : réseaux bayésiens, déséquilibre de liaison, génétique d'association, bloc haplotypique.

1. Introduction

Dans le contexte des maladies génétiques communes, les GWAS (Genome Wide Association Studies), ou études d'association portant sur le génome entier, ont pour objectif de découvrir et de localiser les mutations causales dans le génome humain. A l'heure actuelle, ce type d'études échoue à trouver l'ensemble des facteurs génétiques impliqués dans ces maladies de nature multifactorielle.

Notre objectif in fine est de développer une nouvelle méthode de GWAS à l'aide de réseaux bayésiens. Ces derniers présentent de multiples avantages: l'apprentissage de modèles à partir de données, l'intégration de données hétérogènes, la possibilité de traiter un très grand nombre de variables et une grande souplesse de modélisation avec la prise en compte de connaissances expertes. Dans cette perspective, des travaux ont été initiés avec la construction de réseaux de Markov [VER 06] et de réseaux bayésiens hiérarchiques à variables latentes [NEF 06].

Une des voies les plus prometteuses pour l'amélioration des méthodes de génétique d'association est la prise en compte des dépendances existantes entre les différents SNP (Single Nucleotide Polymorphism). En génétique, ces dépendances sont appelées LD (Linkage Desequilibrium). Différentes approches ont été développées comme les méthodes à base d'haplotypes [SCH 04], de tag-SNP [STR 04], ou la régression logistique ridge [MAL 08].

Dans cet article, nous proposons une modélisation fine des dépendances locales existantes entre les SNP, qui intègre des connaissances biologiques, à l'aide d'un réseau bayésien.

2. Rappels

2.1 Les SNP.

Les SNP sont des marqueurs génétiques correspondant à des sites particuliers dans le génome. Chez l'Homme, il existerait une dizaine de millions de SNP répartis sur l'ensemble du génome [HAP 05]. Les SNP se présentent ainsi comme des marqueurs très intéressants pour mesurer les variations génétiques chez l'Homme.

2.2 Le déséquilibre de liaison.

Le déséquilibre de liaison est l'association non aléatoire entre les allèles de deux loci différents (ou plus) [HAR 97]. Il faut savoir que ces loci peuvent se trouver sur deux chromosomes différents et que le déséquilibre de liaison se mesure sur les haplotypes et non sur les génotypes. Dans cet article, nous nous focalisons sur les dépendances locales qui existent entre SNP proches sur le chromosome. Celles-ci sont structurées en blocs haplotypiques, appelés aussi haploblocs, sur le génome [HAP 05].

2.3 Réseaux bayésiens.

Les réseaux bayésiens sont des modèles graphiques probabilistes [NAÏ 07]. Concrètement, ils sont définis par un graphe orienté sans circuit représentant les relations de dépendance et d'indépendance dans le groupe de variables étudiées et par une distribution de probabilités conditionnelles associée à chaque variable. Différents algorithmes d'apprentissage de la structure du réseau bayésien existent, notamment les algorithmes à base de score. Ces algorithmes explorent l'espace des graphes en cherchant à maximiser un score, tel que le BIC (Bayesian Information Criterion). Par exemple, l'algorithme MWST (Maximum Weight Spanning Tree) se restreint à la recherche de l'arbre ou de la forêt optimale. La recherche gloutonne, quant à elle, utilise des opérateurs d'ajout, d'inversion et de suppression d'arc pour la recherche du graphe optimal.

3. Modélisation des dépendances locales.

3.1 Exploration des dépendances entre SNP sur des données réelles.

Nous aimerions mieux comprendre comment se structurent les dépendances locales entre SNP. Pour cela, nous avons récupéré et étudié, sur 90 individus, une séquence génotypique de 81kb du chromosome 1 extraite du projet HAPMAP. Après prétraitement par le logiciel Gevalt, cette séquence nous offre 24 SNP utilisables. Ce logiciel nous a permis d'inférer les haplotypes pour chaque SNP, de découvrir les haploblocs et de représenter les déséquilibres de liaison entre SNP [DAV 07]. Ce logiciel est basé sur une approche de résolution par maximun de vraisemblance et utilise, pour cela, l'algorithme EM (Espérance Maximisation). Les résultats sont présentés en figures 1a et 1b.

Sur la figure 1a, nous observons que la séquence se découpe en 5 haploblocs. En les comparant avec la matrice de déséquilibre de liaison présente sur la figure 1b, nous nous apercevons que les haploblocs parviennent à rendre compte d'une bonne partie des dépendances existantes entre les SNP. Cependant, nous observons que certaines dépendances (encerclées en pointillés verts) ne sont pas capturées par les haploblocs.

Nous avons ensuite modélisé les dépendances entre SNP avec un réseau bayésien. Afin que puissent apparaître nettement l'équivalent des haploblocs dans le graphe, l'apprentissage de structure du réseau bayésien a été restreint à la recherche d'une forêt optimale. Le logiciel GeNIe SMILE© a été utilisé pour cet apprentissage. Cette approche nous a effectivement permis de retrouver des séquences de dépendances entre SNP et de visualiser ainsi l'équivalent de certains haploblocs sur le graphe.

La recherche des dépendances entre SNP a été réalisée à la fois sur les données haplotypiques (figure 1c) et sur les données génotypiques (figure 1d). Globalement, les deux réseaux bayésiens sont similaires. Nous observons que certains haploblocs ou parties d'haploblocs (entourés en pointillés rouges) sont directement retrouvés dans le graphe de la figure 1c. Par exemple, les SNP de l'haplobloc 2 se retrouvent directement dépendants dans le graphe. La figure 1c montre aussi d'autres situations où les dépendances entre SNP d'un même haplobloc sont quasiment directes: les SNP de l'haplobloc 5 sont presque tous dépendants entre eux via le SNP 59 qui joue le rôle d'intermédiaire. Pour ce qui concerne le réseau bayésien des données génotypiques, nous retrouvons à peu près les mêmes dépendances. Ainsi, les connaissances que nous avons du déséquilibre de liaison au niveau

haplotypique seraient généralisables au niveau génotypique. Enfin, nous observons comme dans la figure 1c des dépendances directes ou indirectes entre SNP d'un même haplobloc.

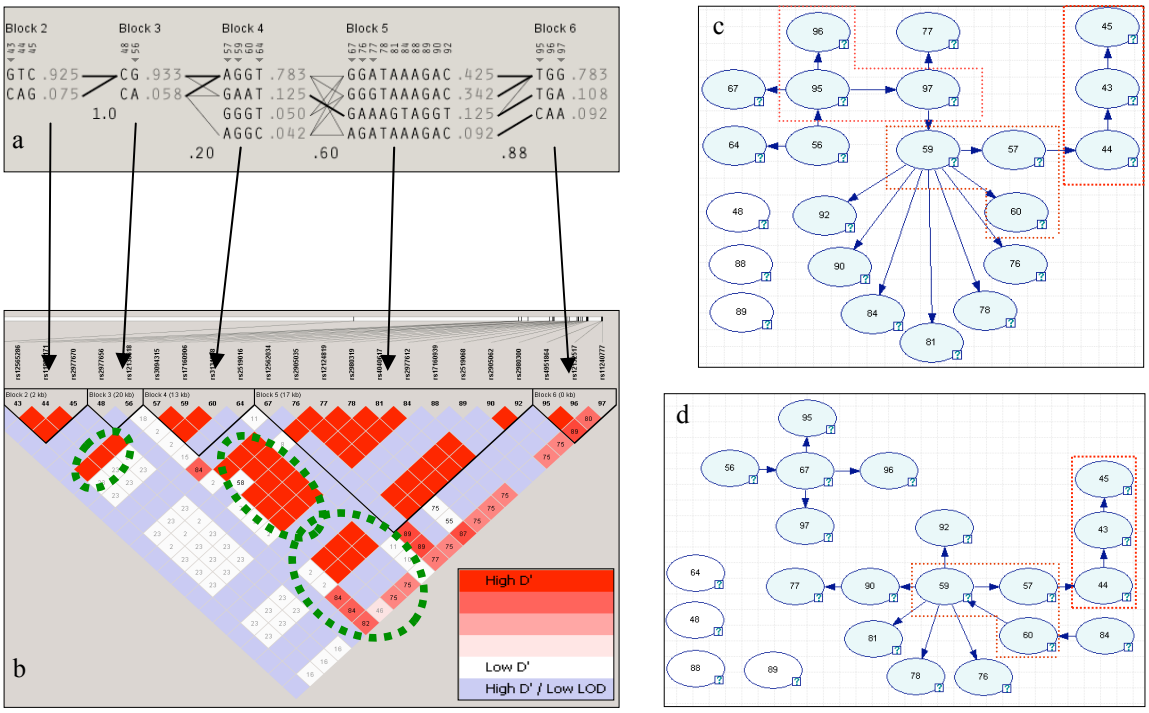


FIGURE 1 *– a) Les 5 haploblocs inférés à partir de la séquence de 81kb, sur 90 individus. b) La matrice des déséquilibres de liaison mesurés pour chaque paire de SNP. Plus la case est rouge, plus le LD est fort. La limite des haploblocs est surlignée en noir. c) Le réseau bayésien construit à partir des haplotypes. d) Le réseau bayésien construit à partir des génotypes.*

Nous avons vu que le découpage en haploblocs rend compte d'une partie du déséquilibre de liaison entre SNP proches. Ainsi, cette connaissance pourrait nous servir lors de l'apprentissage de la structure du réseau bayésien. En outre, nous avons pu constater dans la matrice des déséquilibres de liaison et sur les réseaux bayésiens, que certaines dépendances entre SNP de blocs différents ne sont pas appréhendées par cette approche.

3.2 Modélisation proposée.

Notre approche de modélisation doit pouvoir s'appliquer à tous les SNP génotypés lors d'une GWAS. Généralement, plusieurs centaines de milliers de SNP doivent être analysés dans ce type d'étude. Afin de résoudre ce problème de passage à l'échelle, Nefian a proposé de rechercher les dépendances à l'intérieur de fenêtres de quelques SNP consécutifs sur l'ADN [NEF 06].

Nous pensons que cette approche ne prend pas en compte la structure biologique complexe des dépendances : les blocs de LD sont de taille variable sur l'ADN. C'est pourquoi nous proposons d'utiliser la connaissance des haploblocs afin d'initialiser notre réseau bayésien. En outre, il existe des dépendances entre SNP appartenant à des haploblocs différents mais proches. Les réseaux bayésiens hiérarchiques à variables latentes utilisés par Nefian seraient un bon moyen de modéliser finement le LD. Nous avons repris ce modèle en y apportant quelques modifications afin de modéliser les dépendances intra-haploblocs et inter-haploblocs.

La méthode que nous proposons est développée pour les données génotypiques (figure 2), mais la transposition aux données haplotypiques est directe. Le réseau bayésien est d'abord initialisé en reliant tous les SNP d'un haplobloc à une variable latente appelée « bloc ». La structure ainsi formée est appelée « génobloc ». L'orientation de l'arc est contrainte de la manière suivante: bloc → SNP. Ainsi, le réseau bayésien vérifie que les SNP d'un même génobloc seront indépendants conditionnellement au bloc. Une fois tous les génoblocs construits, l'algorithme SEM (Structural Expectation Maximisation) pour la recherche gloutonne est utilisée

[NEF 06]. Il permet de coupler l'apprentissage de paramètres d'un réseau bayésien à variables latentes à l'apprentissage de sa structure. Nous proposons un algorithme de recherche gloutonne modifié: les opérations de suppression/ajout d'un arc ne sont possibles qu'entre un SNP et la variable bloc dans le sens bloc → SNP. Après que la structure de chaque génobloc ait été apprise, des dépendances entre génoblocs proches sont recherchées à l'aide d'un deuxième algorithme de recherche gloutonne modifiée: les opérations de suppression/ajout d'un arc ne sont possibles qu'entre les variables latentes blocs. L'utilisation d'une fenêtre glissante encadrant plusieurs génoblocs à la fois pour la recherche des dépendances entre les génoblocs semble une voie intéressante.

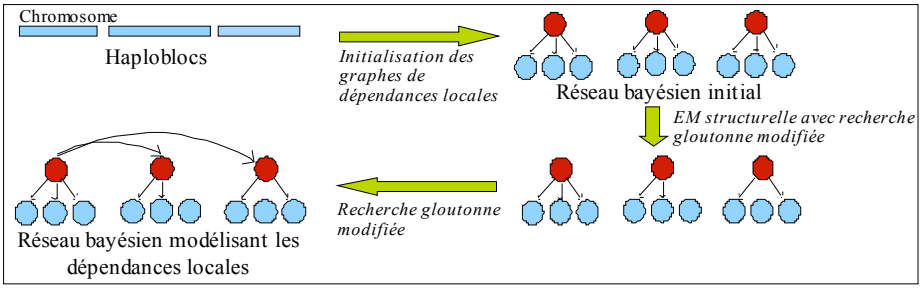


FIGURE 2 *– Les différentes étapes de la méthode proposée pour la modélisation des dépendances locales entre SNP. Les SNP sont représentés par des nœuds bleus clairs et les variables latentes sont représentées par des nœuds rouges foncés.*

4. Conclusion et perspectives.

Nous proposons une nouvelle méthode de modélisation fine des dépendances locales du génome basée sur la connaissance des haploblocs. L'intérêt de notre méthode en termes de qualité d'apprentissage et de rapidité sera à étudier ultérieurement. En outre, une validation expérimentale à grande échelle est à envisager afin de conforter nos résultats préliminaires sur la structuration des dépendances locales entre SNP.

La prochaine étape de notre travail intégrera, en plus, les dépendances globales entre SNP distants ou présents sur des chromosomes différents. Une fois la modélisation complète du déséquilibre de liaison réalisée, notre travail consistera à rechercher les associations entre les SNP et la maladie, à l'aide du réseau bayésien créé.

5. Bibliographie

[DAV 07] DAVIDOVICH O., KIMMEL G., SHAMIR R., GEVALT: An integrated software tool for genotype analysis, *BMC Bioinformatics*, vol. 8, 2007, p. 1-8.

[HAP 05] THE INTERNATIONAL HAPMAP CONSORTIUM, A haplotype map of the human genome, *Nature*, vol. 437, 2005, p. 1299-1320.

[HAR 97] HARTL D. L., CLARK A. G., *Principles of population genetics*, Sinauer associates, Inc, 1997.

[MAL 08] MALO N., LIBIGER O., SCHORK, N. J., Accommodating linkage disequilibrium in genetic-association analyses via ridge regression, *American Journal of Human Genetics*, vol. 82, 2008, p. 375-385.

[NAÏ 07] NAÏM P., WUILLEMIN P. H., LERAY P., POURRET O., BECKER A., *Réseaux bayésiens*, Eyrolles, Paris, 3ème édition, 2007.

[NEF 06] NEFIAN A. V., Learning SNP using embedded Bayesian networks, *IEEE Computational Systems*, Bioinformatics Conference 2006.

[SCH 04] SCHAID D. J., Evaluating association of haplotypes with traits, *Genetic Epidemiology*, vol. 27, 2004, p. 348-364.

[STR 04] STRAM D. O., Tag SNP selection for association studies, *Genetic Epidemiology*, vol. 27, 2004, p. 365-374.

[VER 06] VERZILLI C. J., STALLARD N., WHITTAKER J. C., STONE C., Bayesian graphical models for genomewide associations studies, *The American Journal of Human Genetics*, vol. 79, 2006, p. 100-112.

Classification sous contraintes géographiques

Ayale Daher, Thierry Dhorne, Valérie

Université Européenne de Bretagne
Lab-STICC UMR 3192

RÉSUMÉ. *Dans ce papier, nous posons un problème original d'extraction de connaissances dans des données électorales. L'enjeu est de proposer des outils performants pour restituer, sous forme de cartes facilement interprétables, l'information d'une base de données volumineuses et comportant une dimension géographique. Ceci passe par une étape de classification non supervisée qui permet de synthétiser l'information en un nombre fini de classes homogènes et géographiquement cohérentes. Après un bref état de l'art, nous proposons un modèle original reposant sur une caractérisation parcimonieuse de la covariance des variables observées.*

MOTS-CLÉS : *classification automatique, méthodes de clustering, covariance spatiale.*

1. Introduction

Classer des données dans des groupes est depuis longtemps une préoccupation en analyse de données, cette approche a été utilisée dans des domaines variés tels que la vision par ordinateur, le traitement du langage, l'exploration du Web, le marketing, le diagnostic médical, ou la bio-informatique.Chaque domaine d'application possède ses propres contraintes, objectifs, et développements spécifiques, comme par exemple la gestion de gros volumes d'information pour la fouille de données. La classification en général relève de disciplines comme les mathématiques et plus particulièrement les statistiques, la théorie des graphes, et bien sûr l'informatique.

Pour visualiser, analyser ou expliquer des données spatiales de grande dimension, il peut être utile de regrouper l'information en un nombre raisonnable de régions de l'espace thématiquement homogènes, par exemple pour réaliser des cartes.

Par exemple lors d'une élection présidentielle,pour diffuser des résultats de votes, les médias ont l'habitude de produire des cartes à 2 couleurs faisant ressortir les tendances gauche-droite (ou le score des 2 candidats présents au second tour). On se pose alors la question de savoir s'il serait possible de cartographier une information plus complète représentant de façon synthétique les scores de plusieurs candidats par exemple. En effet, de telles cartes permettraient de séparer la composante géographique des votes (opposition Nord/Sud ou ville/campagne) d'une composante thématique non liée à la géographie et ainsi d'aider à la compréhension du comportement des électeurs. Une classification qui ne prendrait pas du tout en compte la localisation géographique pourrait entraîner une cartographie en forme de mosaïque souvent confuse à interpréter.

De façon plus générale, notre objectif final est d'apporter des éléments de réponse à cette question pratique *Comment prendre en compte de façon pertinente la dimension géographique dans des méthodes de classification non supervisée ?* Cependant, dans ce papier, nous présentons uniquement un bref état de l'art et des perspectives du modèle proposé.

2. Classification sous contraintes géographiques

Dans cette partie, nous proposons un rapide état de l'art des méthodes permettant de partitionner des entités géographiques, en groupes thématiquement homogènes avec une certaine cohérence spatiale. Si des algorithmes classiques de classification non supervisée sont utilisés pour partitionner des données spatiales, la classification obtenue sera en général très morcelée. Pour éviter ce morcellement, il faut considérer l'information géographique contenue dans les données. Plusieurs méthodes ont été proposées dans la littérature pour pendre en compte l'information géographique. On peut les regrouper en trois catégories :

1. les méthodes dans lesquelles un algorithme est adapté afin d'imposer une contrainte de contiguïté aux éléments d'une même classe,

2. les méthodes qui transforment le jeux de données ou une matrice de distance entre individus pour intégrer l'information spatiale,

3. les méthodes où la classification est basée sur une loi de probabilité de mélange qui prend en compte la dimension spatiale. Cette dernière d'approche est très utilisé en segmentation d'image.

2.1. *Approche probabiliste avec contraintes spatiales*

Une alternative aux méthodes 1. et 2., consiste à proposer un modèle probabiliste pour décrire la structure spatio-thématique des données. On modélise alors le fait d'appartenir à une classe pour un individu par une variable aléatoire discrète Z définie par $\{1, ..., K\}$ avec K le nombre de classes. Cette variable n'est pas observée. Puis on décrit la distribution des populations de chacune des classes par un modèle $f_k(x) = P(X = x|Z = k)$. On peut alors introduire la structure spatiale à deux niveaux :

- soit au niveau de la variable latente Z ; les variables observées $X(s)$ avec s la position géographique sont alors généralement indépendantes conditionnellement à la variable latente[AMB 96].

- soit au niveau des modèles conditionnels f_k et dans ce cas les variables latentes $Z(s)$ sont généralement indépendantes.

La première solution est utilisée quand la structure spatiale est relativement simple, comme par exemple en traitement d'image. Dans les techniques de segmentation où l'information n'est plus géographique mais uniquement spatiale (imagerie satellite ou médicale), la modélisation par les champs de Markov permet de prendre en compte les dépendances spatiales entre les pixels d'une image[CHA 00]. L'idée, est que la répartition spatiale de la variable latente est un champ aléatoire markovien : la valeur de la classe en chaque point, tout en étant conditionnée par la distribution de probabilité des valeurs observées, dépend des valeurs de ses voisins et le graphe de voisinage est relativement simple puisque les données sont indexées sur une grille régulière. [ZAN 07] proposent aussi d'utiliser ce type de modèle imposant une structure Markovienne à la variable latente pour faire de la classification sur des graphes. On peut aussi citer les travaux de [FRA 06] qui introduisent un modèle hiérarchique incluant un champ markovien caché pour classer des génotypes.

La seconde approche consiste à modéliser la structure spatiale ou géographique en supposant que les observations sont des réalisations de champs aléatoires. On décrit alors directement la structure de dépendance entre les variables $X(s)$ aux différents sites s. Si la dimension de X est faible, il est assez naturel de modéliser la structure de dépendance par une covariance spatiale. On suppose alors implicitement que, dans la classe k, X est un champ gaussien de moyenne m_k et de covariance Σ_k. Si la dimension de X est faible (de l'ordre de quelques unités), on peut par exemple utiliser des modèles classiques de la géostatistique pour Σ_k[CRE 93]. Mais si la dimension de X est élevée, on doit envisager d'autres types de modèle dans lesquels on puisse caractériser à la fois la dépendance entre les différentes variables observées et la dépendance spatiale de façon parcimonieuse.

3. Introduction du Modèle

Nous proposons ici un modèle original, basé sur un petit nombre de paramètre et permettant de caractériser d'une part la covariance entre les variables X observés, même quand la dimension de X est élevée et d'autre part la dépendance spatiale entre les différents individus. Nous considérons que les observations sont des réalisations d'un processus gaussien dont la moyenne dépend de la classe de l'individu de même que la covariance. Nous supposons de plus que deux individus appartenant à deux classes différentes sont indépendant l'un de l'autre. La covariance totale Σ de chaque classe s'écrit sous la forme d'un modèle linéaire comme la somme de trois composantes.

$$\Sigma_T = \sigma^2 \ I_p + \alpha \ A + \beta \ C$$

avec :

- σ^2 : modélise une dispersion (dépendance) moyenne entre les variables.
- α et β sont des réels.
- I_p est la matrice unité
- A est la matrice de contiguïté qui est une matrice binaire, symétrique, formée d'éléments diagonaux nuls. La contiguïté entre deux individus se définit par le fait qu'elles ont une frontière commune et chaque terme de cette matrice est égale à 1 si les individus sont contiguës à l'ordre 1 et 0 sinon (par convention, un individu n'est pas contiguë avec lui-même : $As_is_j = 0, \forall s_is_j$). Cette notion de contiguïté peut être généralisée : deux individus s_is_j sont contiguës à l'ordre k si k est le nombre minimal de frontières à traverser pour aller de i à j.

$$A(s_i, s_j) = \begin{cases} 1 & \text{si } s_i \text{ est voisin d'ordre 1 de } s_j \\ 0 & \text{sinon} \end{cases}$$

- C est la matrice de classes qu'on souhaite déterminer

$$C(s_i, s_j) = \begin{cases} 1 & \text{si } s_i \text{ et } s_j \text{ sont dans la même classe} \\ 0 & \text{sinon} \end{cases}$$

Pour capter l'interdépendance entre individus, il faut considérer leurs positions relatives. Pour cela, on doit spécifier la topologie du système spatial en construisant une matrice de distance. Cette matrice de distance est une matrice carrée, ayant autant de lignes et de colonnes qu'il y a de zones géographiques.

Dans la suite notre choix s'est porté sur la distance de Mahalanobis, car elle tient en compte de la corrélation de données et que la matrice de covariance est la matrice identité, cette distance est alors la même que la distance euclidienne et se définit ainsi :

$$d(x_i, x_j) = \sqrt{(x_i - x_j)^t S^{-1} (x_i - x_j)}$$

où S désigne la matrice de variance / covariance. Notons que $S = I$ (matrice identité), on se ramène alors à la distance euclidienne.

Ces matrices de distances sont souvent utilisées en raison de leur simplicité mais apparaissent restrictives pour ce qui est de leur définition de la connexion spatiale entre individus. Une autre possibilité consiste à utiliser des matrices de similarité.

$$sim = \exp(-\gamma(D))$$

Dans le cas multivarié spatial, c'est-à-dire qu'il y a une dépendance entre les variables et entre les individus, on fait l'hypothèse simplifiée que $D = X^t M X$ avec $M = \lambda I_p$ est la distance thématique et M matrice de covariance.

4. Inférence

Dans cette section, nous exploitons le cadre d'un algorithme itératif en 2 étapes pour estimer les paramètres.

Étape Prédiction : on affecte chaque individu à une classe (notion de distance).

Étape Minimisation : on estime les paramètres du modèle : $\theta = (\sigma^2, \alpha, \beta, \lambda)$ et sont obtenus en minimisant ce critère :

$$\widehat{(\sigma^2, \alpha, \beta}, \lambda) = argmin \ \ ||exp - \gamma(D) - (\sigma^2 I_p + \alpha A + \beta C||_2$$

On itère ces 2 étapes suivantes jusqu'à la convergence.

La formulation de la dépendance est extrêmement simple mais a l'avantage d'introduire très peu de paramètres.

5. Application

Nous avons choisi de nous intéresser aux résultats des votes du premier tour à des élections au suffrage universel. Une bonne analyse de tels résultats va permettre d'estimer les reports de voix sur les candidats présents au second tour, mais aussi de mieux connaître la population des électeurs.

Ces données sont par ailleurs intéressantes : *par leur nature* : les variables aléatoires sont discrètes ; *par leur dimension* : 12 candidats au premier tour dans l'élection présidentielle de 2007 par exemple ; *par leur volume* : le nombre d'individus est important puisque l'individu est le bureau de vote ; *par leur structure* : la dimension géographique est déterminante et nécessite de définir un graphe de voisinage souvent complexe comportant une information multi-échelle (commune, canton, département, région,...) ; on remarque en effet que les individus peuvent être à différentes échelles allant du bureau de vote au département par exemple.

6. Conclusion

Après un état de l'art rapide, nous avons proposé un modèle original reposant sur une caractérisation parcimonieuse de la covariance des variables observées permettant de prendre en compte la dépendance entre les variables et une structure de dépendance spatiale réduite aux plus proches voisins. La notion de voisinage peut reposer par exemple sur un graphe de contiguïté. La somme de 2 derniers termes modélise ainsi une covariance spatiale décroissante par une fonction constante par morceaux. On remarque que ce modèle découple implicitement l'information spatiale de l'information thématique.

7. Bibliographie

[AMB 96] AMBROISE C., Approche probabiliste en classification automatique et contraintes de voisinage, PhD thesis, Compiègne :UTC, Laboratoire Heudiasyc(UMR CNRS 5699), 1996.

[CHA 00] CHALMOND B., *Éléments de modélisation pour l'analyse d'images*, vol. 33, Berlin : Springer-Verlag, mathématiques et applications édition, 2000.

[CRE 93] CRESSIE N., *Statistics for spatial data*, vol. 900p, John Wiley, New York, 1993.

[FRA 06] FRANCOIS O., ANCELET S., GUILLOT G., Bayesian Clustering using Hidden Markov Random Fields in Spatial Population Genetics, *Genetics*, , n^o 174, 2006, p. 805-817.

[ZAN 07] ZANGHI H., AMBROISE C., MIELE V., Fast Online Graph Clustering via Erdos Renyi Mixture, *rapport technique*, , 2007.

Moran and Geary indices for Multivariate Time Series exploratory analysis

Cedric Frambourg, Ahlame Douzal-Chouakria, Jacques Demongeot

TIMB TIMC-IMAG (CNRS-UMR 5525), Université Joseph Fourier Grenoble 1, France
University of Joseph Fourier Grenoble 1, Laboratory TIMC -IMAG- CNRS UMR 5525, IN3S INstitut de l'INgénierie de l'INformation de Santé, Faculté de Médecine, 38706 LA TRONCHE Cedex, Tel : (33) (0)4 56 52 00 68, Fax : (33) (0)4 56 52 00 22 Ahlame.Douzal@imag.fr

RÉSUMÉ. *This paper focuses on the exploratory analysis of a set of multivariate time series. Two principal component analyses for contiguous data based on Moran and Geary spatial autocorrelation are studied. Here are discussed and illustrated the specification of the Moran and Geary analyses to explore multivariate time series through a synthetic dataset.*

MOTS-CLÉS : *Multivariate Time series, Principal Component Analysis*

1. Introduction

It is quite often that we are faced with datasets where an a priori relationship structure is defined on the statistical units. A pioneering work to include an a priori structure was proposed by Lebart called the local analysis [LEB 69]. Many other studies follow. Wartenberg and Banet et al. propose a factorial analysis based mainly on decomposing the total variance through the neighboring and non neighboring units [WAR 85, BAN 84]. Le Foll generalizes the local analysis to weighting neighborhoods [LEF 82, THI 95]. Mom proposes an new operator generalizing the discriminant factorial analysis to account for a priori neighboring structure [MOM 88]. A smooth analysis and a local differences analysis were proposed by Benali et al.[BEN 89]. We start the paper with the definition of some basic statistics for a spatial autocorrelation measure in Section 2. In Section 3, we introduce two principal component analyses of contiguous data, based on the spatial autocorrelation. Finally, in Section 4, we compare these exploratory approaches through a synthetic dataset and show their main specifications to explore multivariate time series.

2. The spatial autocorrelation measure

Let us consider, for instance, the observation of n values $\mathbf{x}= (x_1, ..., x_n)$ through n geographical sites. One be interested in the influence of the spatial structure on the observed values. May all pairs (i, j) of neighbor values are independent, then a non-spatial autocorrelation is revealed ; otherwise, some dependency exists and has to be estimated. In the case of a general contiguity structure (spatial, temporal, etc.), below we present briefly two main autocorrelation statistics : Moran and Geary indices.

Moran index

Moran proposes an index *I* to estimate how much the observed data are affected by the contiguity structure :

$$I = \frac{1}{\sum_{i,j\in[1,n]} w_{ij}} * \frac{\sum_{i=1}^{n}\sum_{j\in V_i} w_{ij} * z_i * z_j}{\frac{1}{n}\sum_{i=1}^{n} z_i^2} \quad (1)$$

where $z_i = (x_i - \bar{\mathbf{x}})$ are the centered values ; $W = [w_{ij}]$ describes the $(n \times n)$ neighborhood matrix, w_{ij} is the neighborhood weight between the sites i and j, with $\sum_{i,j\in\{1,n\}} w_{ij} = 1$. We note V_i the set of i's neighbors. Under a normalisation constraint and similarly to the classical correlation, I varies in [-1,1]. This $I = 1$ in the case of neighborhoods composed of similar values (i.e., high positive dependency between neighbor values), $I = -1$ in the case of reverse neighbor values (i.e., high negative dependency between neighbor values), and $I = -\frac{1}{(n-1)}$ if there is non dependency between neighbor values.

Geary index

In a different manner, Geary estimates the autocorrelation based on the differences between neighbor values :

$$c = \frac{n-1}{4 * \sum_{i,j\in[1,n]} w_{ij}} \frac{\sum_{i=1}^{n} \sum_{j\in V_i} w_{ij} * (x_i - x_j)^2}{\sum_{i=1}^{n} z_i^2} \tag{2}$$

This $c = 0$ in the case of neighborhoods composed of similar values (i.e., high positive dependency between neighbor values), and it is close to 1 when there is non dependency between neighbor values.

3. Exploratory analysis of contiguous data

We focus in this section on the factorial analysis of contiguous data. We present two main principal components analyses (PCA) each of which includes neighboring information differently.

Let $X = [x_{ij}]$ be an $(n \times p)$ matrix describing n individuals through p variables $X_1, ..., X_p$. Let $W = [w_{ij}]$ be a symmetric $(n \times n)$ matrix describing the neighboring relationship between the n individuals with $\sum_{i,i'} w_{ii'} = 1$. Let $Y = [y_{ij}] = (I - 1_n 1_n^t D)X$ be the D-centered data with $D = [1/n, ..., 1/n]$ the individual diagonal weighting matrix. Finally, we note $V_T = [v_{jk}^T] = Y^t DY$ the classical variance/covariance matrix and $Z = [z_{ij}] = YD_{1/s}$ the standardized data matrix with $D_{1/s} = diag[1/s_1, ..., 1/s_p]$ and $s_j = \sqrt{v_{jj}^T}$.

3.1. *Moran index based PCA : Moran-PCA*

The first approach extending the principal components analysis to contiguous data was proposed by Wartenberg (1985). It aims to extract factorial axes maximizing Moran index. Let $V_I = [v_{jk}^I]$ and $R_I = [r_{jk}^I]$ be the considered variance/covariance and correlation matrices respectively :

$$V_I = Y^t WY \text{ with } v_{jk}^I = Y_j^t WY_k = \sum_{i,i'} w_{ii'} y_{ij} y_{i'k}, \tag{3}$$

$$R_I = Z^t WZ \text{ with } r_{jk}^I = \frac{Y_j^t WY_k}{\sqrt{Y_j^t DY_j}\sqrt{Y_k^t DY_k}} = \frac{\sum_{i,i'} w_{ii'} y_{ij} y_{i'k}}{\sqrt{\frac{1}{n}\sum_i y_{ij}^2}\sqrt{\frac{1}{n}\sum_i y_{ik}^2}}. \tag{4}$$

The diagonal terms of R_I are the Moran indices of the p variables. Unlike a classical correlation matrix, R_I is not positive-definite. Negative eigenvalues are equally important as positive eigenvalues. On the factorial plan, the latter helps to work out heterogeneous regions (i.e. neighborhoods with reverse values), the former works out homogeneous regions (i.e., neighborhoods with similar values).

3.2. *Geary index based PCA : Geary-PCA*

The second approach was proposed by Banet et al. (1984). Its aim is to look for factorial axes maximizing the Geary index. Let $V_c = [v^l_{jk}]$ and $R_c = [r^l_{jk}]$ be the considered variance/covariance and correlation matrices respectively :

$$V_c = Y^t(N-W)Y \text{ with } v^L_{jk} = Y^t_j(N-W)Y_k = \frac{1}{2}\sum_{i,i'} w_{ii'}(y_{ij}-y_{i'k})^2 \quad (5)$$

$$R_c = Z^t(N-W)Z \quad (6)$$

$$\text{with } r^L_{jk} = \frac{Y^t_j(N-W)Y_k}{\sqrt{Y^t_j DY_j}\sqrt{Y^t_k DY_k}} = \frac{1}{2}\frac{\sum_{i,i'} w_{ii'}(y_{ij}-y_{i'j})(y_{i'j}-y_{i'k})}{\sqrt{\frac{1}{n}\sum_i y^2_{ij}}\sqrt{\frac{1}{n}\sum_i y^2_{ik}}} \quad (7)$$

where $N = [w_1, ..., w_n]$ is a diagonal neighborhood weighting matrix with $w_i = \frac{1}{2}\sum_j (w_{ij} + w_{ji})$. The first axis helps to work out highly heterogeneous regions. On contrast to the Moran based approach, on the factorial plan, neighbor units of a same heterogeneous regions are expected to be projected far from each other.

4. Application

4.1. *Data description*

To study the specifications of the above exploratory and discriminant factorial analyses, we consider a synthetic dataset which consists of 3 classes of multivariate time series (7 per class) described by 4 variables X_I, X_C, X_R, and X_M. The first variable describes a global behavior (i.e., maximizes the Moran index), the second one alternates on neighbor values (i.e., maximizes the Geary index), X_R has a random behavior, and X_M is a mix of a global and an alternate behaviors as described in Figure 1 where we display the plot of two series in each classes. Note that variables X_I, X_C, and X_M describe similar behaviors within classes and quite different behaviors across the 3 classes.

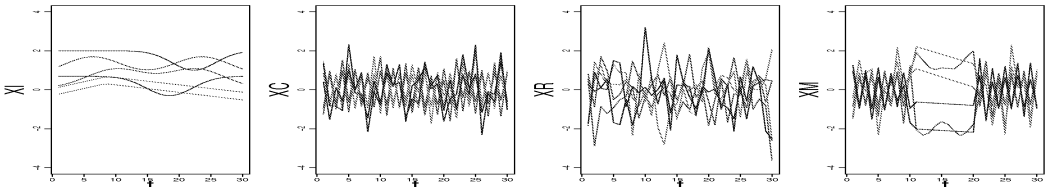


FIGURE 1. *Synthetic multivariate time series*

4.2. *Principal components analysis*

We have performed the two principal components analyses detailed in Section 3. Let us discuss the results obtained in Figure 2 showing the time series projection on the first two factorial axes.
On the one hand, we can see that the PC1 behavior of the classical PCA mainly expresses a combination of the behaviors of X_M and X_C. As the classical PCA looks for the axes maximizing the variance criteria, only the variables with a high variability have been involved in building the first axes. Note that the classical PCA ignores the temporal information which may lead to building axes based on random and non informative variables (e.g., X_R mainly build the axis 2).

One the other hand, the Moran-PCA approach, which looks for the projections maximizing the Moran statistic succeeds to identify X_I as the main variable in building the first factorial axis. The second axis is built by X_M as it is partially composed of a global behavior. On the contrary of the classical PCA, Moran-PCA succeeds to detect and isolate the random variable X_R. The Geary-PCA approach looks for axes maximizing the Geary statistic to select those behaviors alternating on neighbor values. This approach succeeds to detect X_C as maximizing the criteria (i.e., mainly building axis 1) and X_M to a lesser extent as it is partially composed of an alternate behavior and X_R as mainly building axis 2 . Geary-PCA fails to isolate the random variable from the other informative features.

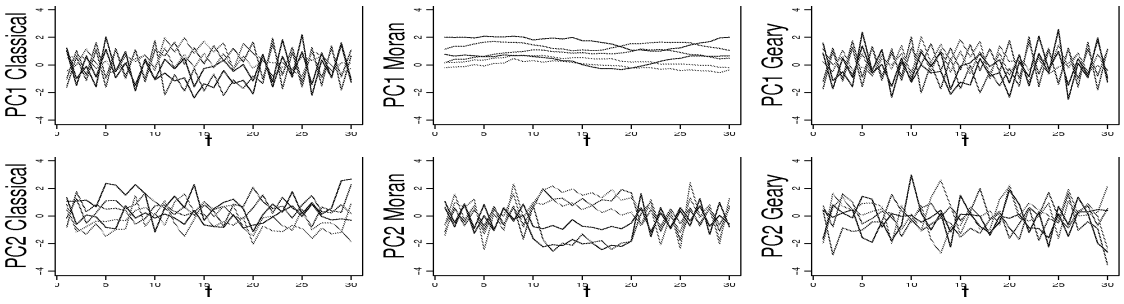


FIGURE 2. *PCA : First components behavior through the PCA approaches*

5. Conclusion

This paper concern is to study the specifications of two main exploratory analyses of contiguous data. Both approaches focus on the within time series progression regardless to the between time series distribution. Future work will address the extension of Moran and Geary analyses to the exploration of the between structure, then to discriminate time series classes.

6. Bibliographie

[BAN 84] BANET T. A., LEBART L., Local and partial principal components analysis and correspondence analysis., *COMPSTAT, Proceedings in Computational Statistics, Physica Verlag, Vienna,.*, , 1984, p. 113-118.

[BEN 89] BENALI H., ESCOFIER B. ., Smooth factorial analysis and factorial analysis of local differences, *Multiway data analysis, North-Holland, Amsterdam.*, , 1989, p. 327-339.

[LEB 69] LEBART L., L'analyse statistique de la contiguïté., *Publication de l'I.S.U.P*, vol. XVIII, 1969, p. 81-112.

[LEF 82] LE FOLL Y., Pondération des distances en analyse factorielle., *Statistiques et Analyse des données*, vol. 7, 1982, p. 13-31.

[MOM 88] MOM A., Méthodologie statistique de la classification de réseaux de transport, PhD thesis, U.S.T.L., Montpellier., 1988.

[THI 95] THIOULOUSE J., CHESSEL D., CHAMPELY S., Multivariate analysis of spatial patterns : a unified approach to local and global structures., *Environmental and Ecological Statistics.*, vol. 2, 1995, p. 1–14.

[WAR 85] WARTENBERG D., Multivariate spatial correlation : a method for exploratory geographical analysis., *Geographical Analysis.*, vol. 17(4), 1985, p. 263–283.

*N*ew LISA indices for spatio-temporal data mining

Catherine d'Aubigny, Gérard d'Aubigny

LJK (UMR 5224, CNRS),
Département de Statistique, Équipe MS^3,
Université Pierre Mendès France, Grenoble 2, France.
{*catherine.daubigny,Gerard.d-Aubigny*}*@upmf-grenoble.fr*

RÉSUMÉ. *One of the first steps of any spatial data mining process consists in "letting data speak for themselves" in a way to discover interesting patterns, and to suggest potential relationships and hypotheses, which take into account the spatial aspects of the data. Typically, the aim is to learn more about each individual datum by relating it to the values observed at neighbouring locations, in a way to discover either local departures from spatial stationarity assumptions, outliers, or boundaries between regions (changes of spatial regimes). The most often used indices for this aim where popularized by Getis and Ord (1992) and Anselin (1995). They are called LISA for Local Indices of Spatial Autocorrelation. Thus, the EDA (Exploratory Data Analysis) step consists in simultaneaously calculating LISA, and mapping either data or LISA in an exploratory framework. Our contribution yields a new class of LISA, which might reveal themselves useful because of their intuitive way to blend of a-spatial and spatial components in the case of spatial data and because of their simple generalization to the case of spatio-temporal data, based on standard methodological principles.*

MOTS-CLÉS : *Spatial autocorrélation, Getis, Geary, Moran, Index numbers, LISA, spatio-temporal data.*

1. Introduction

Starting with the work of Anselin (1995), the definition of Local Indices of Spatial Autocorrelation (or LISA) has been classically constrained to respect a decomposition property. As a matter of fact, Anselin and followers considered it natural to impose an additivity property to local association indices : their sum should be proportional to the global index from which they come. This principle makes it simple to built local indices, but it suffers some weaknesses. In particular, interpretability qualities are not guaranteed. Moreover, the known statistical properties of LISA result directly from those of the generative association index submitted to an addidive decomposition. We did contest the uncritical use of the Moran spatial regression index in several papers, see e.g. d'Aubigny (2005), and as a consequence we proposed modification of it. But we had also to propose new LISA taking into account the noticed weaknesses of the classical Moran index, as programmed in most Spatial Analysis packages, like CRIMESTAT (2009), GEODA (2005), or various R pakages.

While the original proposal by Moran (1950) seems hard to contest, it was limited to the analysis of binary response variables, observed over some spatial domain. What is now called the Moran index is in fact a generalisation of Moran's ideas to the case of numerical variables, due to Cliff and Ord (1973, 1980). It is our feeling that, while interesting for statistical inference purposes - since then interpretation is not important - this generalization suffers several inconsistencies, when scrutinized from the EDA (Exploratory Data analysis) view point. One aspect of the problem is due to the fact that Cliff and Ord (1973) tried to generalize to the spatial setting the arguments used by Durbin and Watson (1950) when the adapted the the Neyman ratio to time series analysis. Cliff and Ord seem to have neglected complexities generated by space. In particular, they did not take pay attention to the induced unbalance of the design of observation that departs the spatial case from the simpler time series setting.

The data mining point of viewtries to articulate statistical calculus and graphics. Thus, geometrical consistency becomes important in a way to get simple mapping tools taylored for visual detection of interesting patterns in the data. So, we established some corrections to the initial Moran index in a way to allow geometry to come into play. While section 1 is devoted to mathematical reminders, the corrected indices are presented in section 2. Section 3 is devoted to the definition of new LISA which decompose the corresponding corrected Global index. One useful property of such proposals lies in their very simple adaptation to the analysis of spatio-temporal data. The problems uncountered in such a situation are complicated by the fact that an analyst must tackle both the variability over time and the variability over space. It is usual practice in descriptive statistics to built mixte indices in a way to try to understand the sources of variation over each of these variability sources when the other is fixed.

2. Spatial information and Euclidean embedding of a graph

Let $G = (V, E)$ denote a simple graph with vertex set $V = \{1, 2, \cdots, n\}$ and edge set E. At least two obvious ways exist for specifying G by means of a matrix : the adjacency one $\mathbf{A}$ - it is $n \times n$ with general term a_{ij} - and the incidence one ∇ - it is $|E| \times |V|$ and its definition needs a fixed orientation of each edge $e = (s, t)$ - :

$$a_{ij} = \begin{cases} 1 & \text{if } (i,j) \in E, \\ 0 & \text{otherwise.} \end{cases} \qquad \nabla_{ex} = \begin{cases} -1 & \text{if } x = s, \\ +1 & \text{if } x = t, \\ 0 & \text{otherwise.} \end{cases}$$

The set of functionals $f : V \mapsto \mathbb{R}$ may be interpreted as a vector space F isomorphic to $\mathbb{R}^n$, and the map $f \mapsto \nabla f$ is known as the co-boundary mapping of the gaph G. Its value $(\nabla f)(e)$ is the difference of the values of f at the two end-points of the edge e (considering orientation). So one can interpret ∇ as a kind of difference (or "discrete differential") operator induced by G. The definition of second order "derivatives" is possible by extending this process to cycles C of Eulerian graphs G. One needs fix an orientation of G such that C is properly oriented so that every edge points "forward" along C.This is always possible and then, the cycle C may pass through each vertex x multiple times. For the i-th pass, the incoming edge is $e_1^{(i)} = (s_1^{(i)}, x)$, and the outcoming one is $e_2^{(i)} = (x, s_2^{(i)})$. So, one can define the "2nd derivatives" along C as

$$(\partial^2_{C,i} f)(x) \triangleq (\nabla f)(e_2^{(i)}) - (\nabla f)(e_1^{(i)}) = [f(s_2^{(i)}) - f(x)] - [f(x) - f(s_1^{(i)})] = f(s_1^{(i)}) + f(s_2^{(i)}) - 2f(x) \quad (1)$$

$(\partial^2_{C,i} f)(x)$ is independent of the orientation on G, and if each pass of C through x is interpreted as a different dimension, a natural definition of a "Laplace -Beltramy operator" results in summing "2nd derivatives" over passes :

$$(\mathbf{\Delta} f)(x) \triangleq \sum_{\text{passes i of } C \text{ through } x} (\partial^2_{C,i} f)(x) \sum_{i=1}^{d(x)/2} [f(s_1^{(i)}) + f(s_2^{(i)}) - 2f(x)] = \sum_{s \in \mathcal{N}(x)} [f(s) - f(x)]$$

where $\mathcal{N}(x)$ denotes the set of neighbors of x in G ($a_{sx} = 1$), and $d(x) = |\{e \in E |\ x \in e\}|$ is the outer degree of the vertex x. This definition of $\mathbf{\Delta}$ can be extended to arbitrary graphs. Notice that in the graph theory literature, one defines the (combinatorial) *Laplacian operator* $\mathcal{L} : F \mapsto F^*$ with the opposite sign :

$$\mathcal{L} f(x) = (-\mathbf{\Delta} f)(x) = \sum_{s \in \mathcal{N}(x)} [f(x) - f(s)]$$

where F^* is the algebraic dual vector space of F. In matrix form, one gets the so-called (combinatorial) *Laplacian matrix* of G :

$$\mathbf{L} = {}^t\nabla\nabla \quad \Leftrightarrow L_{uv} = \sum_{e \in E} \nabla_{eu} \nabla_{ev} = \begin{cases} -1 & \text{if } (u,v) \in E, \\ d(u) & \text{if } u = v, \\ 0 & \text{otherwise.} \end{cases}$$

Let us denote by $\mathbf{D}_{a+}$ the diagonal matrix of order n with general term the outer degree of the node s_i : $d(s_i) = \sum_{j=1}^n a_{ij} \triangleq a_{i+}$. Then, one can easily check that matrices $\mathbf{L}$ and $\mathbf{A}$ are tightly related : $\mathbf{L} = \mathbf{D}_{a+} - \mathbf{A}$, and $\mathbf{L}$ is

semi-definite positive by construction. Taking into account spatial relationships through contiguity graphs yields rather elementary mathematics, mainly because A is then implicitly defined as a symmetric $\{0,1\}$ matrix. But, as usual in data mining, any variable X observed at each node $s \in V$ may be interpreted as an element of an n dimensional vector space, denoted by F above, and the incidence matrix ∇ maps its representation vector $\mathbf{x}$ as the simple contrasts vector $\nabla\,\mathbf{x}$, where $\nabla\,\mathbf{x}$ lives in $\mathbb{F} \triangleq F \wedge F = \wedge^2 F$, $dim(F) = n$ and $n^* \triangleq dim(\mathbb{F}) = \frac{n(n-1)}{2}$:

$$\mathbf{x} = \big(x_i\big)_i \hookrightarrow \nabla\,\mathbf{x} = \big(x_i - x_j\big)_{i,j} \triangleq d\mathbf{x} \in \mathbb{F} \tag{2}$$

In full generality, the dimensionality of $\mathbb{F}$ should be n^2 for a non symmetric incidence matrix A, and reduction from $n(n+1)/2$ to $n(n-1)/2$ results from the assumption that the graph G admits no loop. This condition is not mandatory. It is usually assumed in a way to secure the identifiability of models' parameters. Fixing an Euclidean structure on $\mathbb{F}$ is mandatory for data visualisation purposes. The easiest way to do it consists in retaining the identity metric on $\mathbb{F}$, that is $D_Q = I_{n^*}$. In this case, the analysis will take into account only topological relationships described by the graph G. In a way to integrate metric relations among measurement locations, one can generalize the above approach to the case of a weighted graphs (G, q), where $q(e)$ denotes the weight attributed to the edge e. Then, $\mathbb{F}$ is equiped with a diagonal metric D_Q associated with the symmetric matrix $\mathbf{Q}$ with generic term $q_{ij} = q(e) \in \mathbb{R}^+$ if $e = (i,j) \in E$ and $q_{ij} = 0$ if $(i,j) \notin E$. Without loss of generality, we assume here that the weights are normalized : $q_{ii} = 0\ \forall i,\ \sum_i \sum_j q_{ij} = 1$, and we write $q_{i+} = \sum_j q_{ij}$. By classical arguments, F and $\mathbb{F}$ are two isometric Euclidean spaces, when F is equiped with the semi-metric

$$\mathbf{L}_Q \triangleq^t \nabla \mathbf{D}_Q\, \nabla = \mathbf{D}_{q_+} - \mathbf{Q} = \mathbf{D}_{q_+}\,(\mathbf{I} - \mathbf{D}_{q_+}^{-1}\,\mathbf{Q}) \tag{3}$$

Moreover, for any $j \in \{1, 2, \cdots, n\}$, $x_j = X(s_j)$ gives the value of the variable X observed at location s_j, and since $\sum_j \frac{q_{ij}}{q_{i+}} = 1$, the general term of $\widetilde{\mathbf{x}} \triangleq \mathbf{D}_{q_+}^{-1}\mathbf{Q}\mathbf{x}$ is the mean of X over the neighbourhood of s_i :

$$\widetilde{X}(s_i) \triangleq \widetilde{x}_i = \sum_j \frac{q_{ij}}{q_{i+}} x_j = \sum_j \frac{q_{ij}}{q_{i+}} X(s_j)$$

3. Definition of LISA

Let us denote by X the observed variable, $\bar{x} = \frac{1}{n}\sum_s x_s$ its a-spatial sample average and $z_s = \frac{x_s - \bar{x}}{s(X)}$ its corresponding standardized transform. The original definition of a global spatial regression index proposed by Moran (1950) writes :

$$I_M \triangleq \frac{\sum_{st} q_{st}(x_s - \bar{x})(x_t - \bar{x})}{\frac{1}{n}\sum_s (x_s - \bar{x})^2} = \sum_{st} q_{st} z_s z_t \tag{4}$$

Its interpretation is not easy, because of two sources of inconsistency. First, the numerator does depend on a centring with help of the ordinary average of observations $\bar{x}$, failing to integrate the Euclidean geometry induced by the weighted graph (G, q). Second, the denominator uses an a-spatal index of dispersion. As a result the ratio depends in an unpredictable way of the spatial information $\mathbf{Q}$. Nevertheless, Anselin (1995) shows that it may be interpreted as the estimated slope (in an ordinary Least squares sense) of the calibration model $Y = \alpha_0 + \alpha_1 x + \xi$ where $\mathbf{Y} = \mathbf{Q}\mathbf{x}$ is interpretable as the mean of spatial neighbours of each location point, only if $\mathbf{Q}$ is chosen stochastic, that is to say such that each line sums to 1. One generally uses the notation $\mathbf{Q} = \mathbf{W}$ in that case. The consistency problems are less pregnant when one uses Geary's contiguity ratio c, defined as, see Geary(1952) :

$$c \triangleq \frac{(n-1)}{2n} \frac{\sum_{st} q_{st}(x_s - x_t)^2}{\frac{1}{n}\sum_s (x_s - \bar{x})^2} = \frac{(n-1)}{2n} \sum_{st} q_{st}(z_s - z_t)^2 \tag{5}$$

because c does not depend on the choice of a mean. In both cases, Anselin (1995) defined local indices of spatial correlation, in a very simple additive way :

$$I_M = \sum_s I_s, \qquad I_s = z_s \sum_t q_{st} z_t, \qquad c = \sum_s c_s, \qquad c_s = \frac{(n-1)}{2n} \sum_t q_{st}(z_s - z_t)^2 \tag{6}$$

But the interpretation of these local indices is not especially clear. However, d'Aubigny (2005) proposed to correct the two Global indices, in a way to preserve geometric interpretation. In case of no spatial correlation, each observation point remains weighted by the outer degree of the corresponding node of the graph G. So, The diagonal metric D_{q+} defines an Euclidean geometry on F. The corresponding average writes $\bar{x} = \sum_s q_{s+} x_s$, and we can define two corrected indices of spatial association, corrsponding respectively to I_M and c :

$$J_M \triangleq \frac{\sum_{st} q_{st}(x_s - \bar{x})(x_t - \bar{x})}{\sum_s q_{s+}(x_s - \bar{x})^2} \qquad J_C \triangleq \frac{\frac{1}{2}\sum_{st} q_{st}(x_s - x_t)^2}{\sum_s q_{s+}(x_s - \bar{x})^2}$$

Then, our proposal of new LISA results from the following rewriting (illustrated here on J_M) :

$$J_M \triangleq \frac{\sum_s q_{s+}(x_s - \bar{x})^2 \sum_t \frac{q_{st}}{q_{s+}} \frac{(x_t - \bar{x})}{(x_s - \bar{x})}}{\sum_s q_{s+}(x_s - \bar{x})^2} = \sum_s cv_s j_s = \sum_s J_s, \quad Js \triangleq cv_s j_s \tag{7}$$

This equation shows that the corrected Moran's index combines two sources of variation :

$$j_s = \frac{(\widetilde{x}_s - \bar{\widetilde{x}})}{(x_s - \bar{x})}, \text{ and } cv_s = \frac{q_{s+}(x_s - \bar{x})^2}{\sum_s q_{s+}(x_s - \bar{x})^2} \geq 0, \text{ so that } \sum_s cv_s = 1.$$

The a-spatial information cv_s gives the contribution of observation s to the variance of X, while the spatial component of the index j_s compares the value observed in location s to the average of its values at neighbouring locations. Thus, the Global index is a sum of n local ones, which appear as products of a spatial and an a-spatial component. J_M appears as a combination of easily interpretable components, and its structure respects the definition of so called *index numbers*. So, knowledge about its behavior will benefit from their documented usage by most national statistics institutes, for the monitoring of economic politics, based on the follow-up of number indices over time. One factor of complexity of the analysis of the time series $\{J_M(t) |\ t = 1, 2 \cdots, t_{max}\}$, results from the mixture of two sources of variability. One is purely a-spatial, and described by the multidimensional time series $\{cv_s(t) |\ t = 1, 2 \cdots, t_{max}\}$, and the second one is clearly influenced by spatial features, and described by the multidimensional time series $\{j_s(t) |\ t = 1, 2 \cdots, t_{max}\}$. Statistical analysis classically compares one of the two components of the product to a reference period - say $t = 0$ - with help of *relative* indices, namely the *index of relative variability localized at site* s : $VR_s^0(t) = \frac{cv_s(t)}{cv_s(0)}$, or the *index of relative predictability localized at site* s : $JR_s^0(t) = \frac{J_s(t)}{J_s(0)}$. Thus, it becomes possible to study the evolution of these relative indices over time, by calculation e.g. the well known *Laspeyres index* of spatial prédictability $LJR^0(t)$, or its alternative, the Laspeyres index of localized variability $LVR^0(t)$, defined as :

$$LJR^0(t) = \sum_s \frac{cv_s(0)J_s(0)}{J_M(0)} \times JR_s^0(t), \qquad LVR^0(t) = \sum_s \frac{cv_s(0)J_s(0)}{J_M(0)} \times VR_s^0(t)$$

The index $LJR^0(t)$ is a weighted average of relative indices of spatial predictability, calculated at time t, where the the weights are the contributions of each location to the (corrected) Moran index, evaluated at time 0. In the case of $LVR^0(t)$, one calculates a weighted average of relative indices of localized variability measured at time t, with weights equal to the contribution of each location to the (corrected) Moran index calculated at time 0. This strategy is not unique and the theory of index numbers classically considers e.g. the corresponding Paasche indices. Variations along these strategies may reveal themselves very useful to follow-up the evolution of local spatial dependencies as well as local spatial heterogeneity over the period of time of interest.

4. Bibliographie

Anselin, L. (1995). Local indicators of spatial association - LISA. *Geographical Systems*, 3, 1-13..

Aubigny (d'), G. (2005). Dépendance spatiale et autocorrélation. Chapitre 2 in *Analyse statistique des données spatiales*, Droesbeke, J.-J., M. Lejeune and G. Saporta (Eds.). Editions Technip, Paris.

Cliff, A. D. and J. K. Ord (1973). *Spatial autocorrelation*, London, Pion.

Getis, A. and J. K. Ord (1992). The analysis of spatial association by use of distance statistics *Geographical Analysis*. 24, 189-206.

K-mean clustering of misaligned functional data

Laura M. Sangalli, Piercesare Secchi, Simone Vantini, Valeria Vitelli

MOX - Dipartimento di Matematica
Politecnico di Milano
Piazza Leonardo da Vinci 32
20133 Milano, Italy
laura.sangalli@polimi.it, piercesare.secchi@polimi.it, simone.vantini@polimi.it, valeria.vitelli@polimi.it

RÉSUMÉ. *We deal with the problem of classification of functional data, in the context of possible misalignment of the data. We describe a k-mean alignment algorithm which jointly clusters and aligns a set of functional data. We illustrate the procedure via application to a real dataset concerning three-dimensional cerebral vascular geometries.*

MOTS-CLÉS : *Classification of functional data, alignment of functional data, k-mean algorithm*

1. Introduction

Recent years have seen an explosive growth in the recording of data having a functional nature. This is essentially due to the development of many devices which are able to provide two and three-dimensional images and other measures of quantities of interest, captured in time and space (e.g., diagnostic medical scanners, satellite and computer vision devices). The analysis of these data poses new challenging problems and requires the development of novel statistical techniques.

Here we shall deal with the issue of efficient classification of functional data. Complications arise in this context, due to the possible misalignment of the data. Misalignment, that is peculiar to this kind of data, acts in fact as a confounding factor.

Figure 1, left, gives an illustrative example of misalignment of functional data. It displays the first derivatives x', y', z' of the three spatial coordinates of the centerlines of the Internal Carotid Artery (ICA) of 65 patients : the phase variability, which results in the evident misalignment of the data, is here due to the different dimensions and proportions of patients' skulls. These data, obtained from reconstructions of three-dimensional angiographic images, have been collected and analyzed within the AneuRisk Project [1], a joint research program that aims at evaluating the role of vascular geometry and hemodynamics in the pathogenesis of cerebral aneurysms. One of the task of the Project consisted in the classification of ICA's with different morphological shapes : such a classification could indeed be helpful in the determination of the risk level of a given patient, since the shape of the ICA influences the pathogenesis of cerebral aneurysms through its effects on the hemodynamics. When classifying these data, to obtain meaningful results, the phase variability must be suitably elicited, by appropriately aligning the data.

We describe a k-mean alignment algorithm that jointly clusters and aligns a set of functional data. When applied to the analysis of AneuRisk dataset, the procedure is able to identify two prototype shapes of ICA's that are described in the medical literature. Sangalli et al. [SAN 08] illustrate the efficiency of the algorithm via simulations studies and applications to other real datasets.

1. The project involves MOX Laboratory for Modeling and Scientific Computing and LABS Laboratory of Biological Structure Mechanics (Politecnico di Milano), Istituto Mario Negri (Ranica), Ospedale Niguarda Ca' Granda and Ospedale Maggiore Policlinico (Milano), and is supported by Fondazione Politecnico di Milano and Siemens Medical Solutions Italia.

2. Defining phase and amplitude variabilities

The variability among two or more curves can be thought of as having two components : *phase variability* and *amplitude variability*. Heuristically, phase variability is the one that can be eliminated by suitably aligning the curves, and amplitude variability is the remaining variability among the curves once they have been aligned. Consider a set $\mathcal{C}$ of (possibly multidimensional) curves $\mathbf{c}(s) : \mathbb{R} \to \mathbb{R}^d$. Aligning $\mathbf{c}_1 \in \mathcal{C}$ to $\mathbf{c}_2 \in \mathcal{C}$ means finding a warping function $h(s) : \mathbb{R} \to \mathbb{R}$, of the abscissa parameter s, such that the two curves $\mathbf{c}_1 \circ h$ and $\mathbf{c}_2$ are the most similar (with $(\mathbf{c} \circ h)(s) := \mathbf{c}(h(s))$). It is thus necessary to specify a similarity index $\rho(\cdot, \cdot) : \mathcal{C} \times \mathcal{C} \to \mathbb{R}$ that measures the similarity between two curves, and a class W of warping functions h (such that $\mathbf{c} \circ h \in \mathcal{C}$, for all $\mathbf{c} \in \mathcal{C}$ and $h \in W$) that indicates the allowed transformations for the abscissa. Aligning $\mathbf{c}_1$ to $\mathbf{c}_2$, according to (ρ, W), means finding $h^* \in W$ that maximizes $\rho(\mathbf{c}_1 \circ h, \mathbf{c}_2)$. This procedure decouples phase and amplitude variability without loss of information : phase variability is captured by the optimal warping function h^*, whilst amplitude variability is the remaining variability between $\mathbf{c}_1 \circ h^*$ and $\mathbf{c}_2$.

Note that the choice of the couple (ρ, W) is problem-specific, since it translates the similarity concept in the context under study, defining what is meant by phase variability and by amplitude variability. For instance, in the applied problem introduced in the previous section, two vessel centerlines can be considered similar if they are identical except for shifts and dilations along the three main axes. For this reason, when analyzing the AneuRisk dataset, Sangalli et al. [SAN 09] introduced the following bounded similarity index between two curves $\mathbf{c}_1 \in L^2(S_1 \subset \mathbb{R}; \mathbb{R}^d)$ and $\mathbf{c}_2 \in L^2(S_2 \subset \mathbb{R}; \mathbb{R}^d)$, where $\mathbf{c}_1' \in L^2(S_1 \subset \mathbb{R}; \mathbb{R}^d)$, $\mathbf{c}_2' \in L^2(S_2 \subset \mathbb{R}; \mathbb{R}^d)$ and $S_{12} = S_1 \cap S_2$ has positive Lebesgue measure :

$$\rho(\mathbf{c}_1, \mathbf{c}_2) = \frac{1}{d} \sum_{p=1}^{d} \frac{\int_{S_{12}} c_{1p}'(s) c_{2p}'(s) ds}{\sqrt{\int_{S_{12}} c_{1p}'(s)^2 ds} \sqrt{\int_{S_{12}} c_{2p}'(s)^2 ds}} \tag{1}$$

with c_{ip} indicating the pth component of $\mathbf{c}_i$, $\mathbf{c}_i = (c_{i1}, \ldots, c_{id})$. This index in fact assumes its maximal value 1 if the two curves are identical except for shifts and dilations of the components :

$$\rho(\mathbf{c}_1, \mathbf{c}_2) = 1 \quad \Leftrightarrow \quad \text{for } p = 1, \ldots, d, \ \exists A_p \in \mathbb{R}^+, B_p \in \mathbb{R} : \ c_{1p}(s) = A_p c_{2p}(s) + B_p \quad \forall s \in S_{12} \, .$$

The choice of this similarity index comes along with the following choice for the class W of warping functions of the abscissa :

$$W = \{h : h(s) = ms + q \ \text{ with } \ m \in \mathbb{R}^+, q \in \mathbb{R}\} \tag{2}$$

i.e., the group of strictly increasing affine transformations.

It is important to note that the couple formed by the similarity index ρ and the class of warping functions W must satisfy some minimal requirements, that ensure that the clustering and aligning problem is well-posed and the corresponding procedure is coherent. For instance, ρ and W must be consistent in the sense that, if two curves $\mathbf{c}_1$ and $\mathbf{c}_2$ are simultaneously warped along the same warping function $h \in W$, their similarity does not change :

$$\rho\left(\mathbf{c}_1, \mathbf{c}_2\right) = \rho\left(\mathbf{c}_1 \circ h, \mathbf{c}_2 \circ h\right), \quad \forall \, h \in W.$$

This guarantees that it is not possible to obtain a fictitious increment of the similarity between two curves $\mathbf{c}_1$ and $\mathbf{c}_2$ simply by warping them simultaneously to $\mathbf{c}_1 \circ h$ and $\mathbf{c}_2 \circ h$.

3. Curve clustering when curves are misaligned

Consider the problem of clustering and aligning a set of N curves $\{\mathbf{c}_1, \ldots, \mathbf{c}_N\}$ with respect to a set of k template curves $\underline{\boldsymbol{\varphi}} = \{\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_k\}$ (with $\{\mathbf{c}_1, \ldots, \mathbf{c}_N\} \subseteq \mathcal{C}$ and $\underline{\boldsymbol{\varphi}} \subseteq \mathcal{C}$). For each template curve $\boldsymbol{\varphi}_j$ in $\underline{\boldsymbol{\varphi}}$, define the domain of attraction

$$\Delta_j(\underline{\boldsymbol{\varphi}}) = \{\mathbf{c} \in \mathcal{C} : \sup_{h \in W} \rho(\boldsymbol{\varphi}_j, \mathbf{c} \circ h) \geq \sup_{h \in W} \rho(\boldsymbol{\varphi}_r, \mathbf{c} \circ h), \forall \, r \neq j\}, \quad j = 1, \ldots, k.$$

Moreover, define the labelling function

$$\lambda(\underline{\boldsymbol{\varphi}}, \mathbf{c}) = \min\{r : \mathbf{c} \in \Delta_r(\underline{\boldsymbol{\varphi}})\}\,.$$

Note that $\lambda(\underline{\boldsymbol{\varphi}}, \mathbf{c}) = j$ means that the similarity index obtained by aligning $\mathbf{c}$ to $\boldsymbol{\varphi}_j$ is at least as large as the similarity index obtained by aligning $\mathbf{c}$ to any other template $\boldsymbol{\varphi}_r$, with $r \neq j$. Thus $\boldsymbol{\varphi}_{\lambda(\underline{\boldsymbol{\varphi}},\mathbf{c})}$ indicates a template the curve $\mathbf{c}$ can be best aligned to and hence $\lambda(\underline{\boldsymbol{\varphi}}, \mathbf{c})$ a cluster the curve $\mathbf{c}$ should be assigned to.

Now, if the k templates $\underline{\boldsymbol{\varphi}} = \{\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_k\}$ were known, then clustering and aligning the set of N curves $\{\mathbf{c}_1, \ldots, \mathbf{c}_N\}$ with respect to $\underline{\boldsymbol{\varphi}}$ would simply mean to assign $\mathbf{c}_i$ to the cluster $\lambda(\underline{\boldsymbol{\varphi}}, \mathbf{c}_i)$ and align it to the corresponding template $\boldsymbol{\varphi}_{\lambda(\underline{\boldsymbol{\varphi}},\mathbf{c}_i)}$, for $i = 1, \ldots, N$. Unfortunately, the k template curves are not known and need to be themselves estimated from the data, leading to a complex optimization problem. We do not attempt a direct solution to this problem, but propose to deal with it by the following k-mean alignment algorithm.

4. k-mean alignment algorithm

The algorithm is iterative. Let $\underline{\boldsymbol{\varphi}}^{[q-1]} = \{\boldsymbol{\varphi}_1{}^{[q-1]}, \ldots, \boldsymbol{\varphi}_k{}^{[q-1]}\}$ be the set of templates after iteration $q-1$, and $\{\mathbf{c}_1{}^{[q-1]}, \ldots, \mathbf{c}_N{}^{[q-1]}\}$ be the N curves aligned and clustered to $\underline{\boldsymbol{\varphi}}^{[q-1]}$. At the qth iteration the algorithm performs the following three steps.

Template identification step. For $j = 1, \ldots, k$, the template of the jth cluster, $\boldsymbol{\varphi}_j{}^{[q]}$, is estimated using all curves assigned to cluster j at iteration $q-1$, i.e. all curves $\mathbf{c}_i{}^{[q-1]}$ such that $\lambda(\underline{\boldsymbol{\varphi}}^{[q-1]}, \mathbf{c}_i{}^{[q-1]}) = j$.

Assignment and alignment step. The set of curves $\{\mathbf{c}_1{}^{[q-1]}, \ldots, \mathbf{c}_N{}^{[q-1]}\}$ is clustered and aligned to the set of templates $\underline{\boldsymbol{\varphi}}^{[q]} = \{\boldsymbol{\varphi}_1{}^{[q]}, \ldots, \boldsymbol{\varphi}_k{}^{[q]}\}$: for $i = 1, \ldots, N$, the i-th curve $\mathbf{c}_i{}^{[q-1]}$ is aligned to $\boldsymbol{\varphi}_{\lambda(\underline{\boldsymbol{\varphi}}^{[q]},\mathbf{c}_i{}^{[q-1]})}$ and the aligned curve $\tilde{\mathbf{c}}_i{}^{[q]} = \mathbf{c}_i{}^{[q-1]} \circ h_i{}^{[q]}$ is assigned to cluster $\lambda(\underline{\boldsymbol{\varphi}}^{[q]}, \mathbf{c}_i{}^{[q-1]}) \equiv \lambda(\underline{\boldsymbol{\varphi}}^{[q]}, \tilde{\mathbf{c}}_i{}^{[q]})$.

Normalization step. For $j = 1, \ldots, k$, all the $N_j{}^{[q]}$ curves $\tilde{\mathbf{c}}_i{}^{[q]}$ assigned to cluster j are warped along the warping function $(\bar{h}_j{}^{[q]})^{-1}$, where $\bar{h}_j{}^{[q]} = \frac{1}{N_j{}^{[q]}} \sum_{i:\lambda(\underline{\boldsymbol{\varphi}}^{[q]},\tilde{\mathbf{c}}_i{}^{[q]})=j} h_i{}^{[q]}$, obtaining $\mathbf{c}_i{}^{[q]} = \tilde{\mathbf{c}}_i{}^{[q]} \circ (\bar{h}_j{}^{[q]})^{-1} = \mathbf{c}_i{}^{[q-1]} \circ h_i{}^{[q]} \circ (\bar{h}_j{}^{[q]})^{-1}$. This ensures that the average warping undergone by curves assigned to cluster j is the identity transformation, thus avoiding the drifting apart of clusters or the global drifting of the overall set of curves.

The algorithm is stopped when, in the assignment and alignment step, the increments of the similarity indexes are all lower than a fixed threshold. Technical details for the implementation of this k-mean alignment algorithm, with the couple (ρ, W) defined in (1) and (2), are given in [SAN 08].

5. An application to the analysis of three-dimensional cerebral vascular geometries

In this section we classify the AneuRisk data by means of the k-mean alignment algorithm. Figure 1 shows the boxplot of the similarity indexes between the original centerlines and the estimated template centerline, and the boxplots of the similarity indexes between the k-mean aligned centerlines and the associated estimated templates, for $k = 1, 2, 3$. Note that 1-mean alignment leads to a large increase in the similarities, with respect to the similarities of the original curves, but a further reasonable gain in the similarities can be obtained by setting $k=2$. Thus, the procedure suggests the presence of 2 amplitude clusters. Figure 1, center, shows the first derivatives of the three spatial coordinates of the 2-mean aligned centerlines. Figure 2 gives a three-dimensional visualization of the estimated template centerlines of the two clusters. It is very interesting to note that the algorithm identifies two prototype shapes of ICA's that are described in the medical literature (see e.g., [KRA 82]). The cluster displayed in orange can in fact be interpreted as the cluster of Ω-shaped ICA's (i.e., ICA's having one siphon in their distal part) and the cluster displayed in green as the one of S-shaped ICA's (i.e., ICA's having two siphons in their distal part).

An analysis of these data by simple k-mean clustering without alignment does not instead give interesting insights. Figure 3 displays the two templates of the clusters obtained by 2-mean clustering without alignment.

Notice that the two templates appear to have almost the same morphological shape, and seem to differ mainly in their phase. For these data, the simple k-mean clustering without alignment is driven by phase variability and fails to identify different morphological shapes.

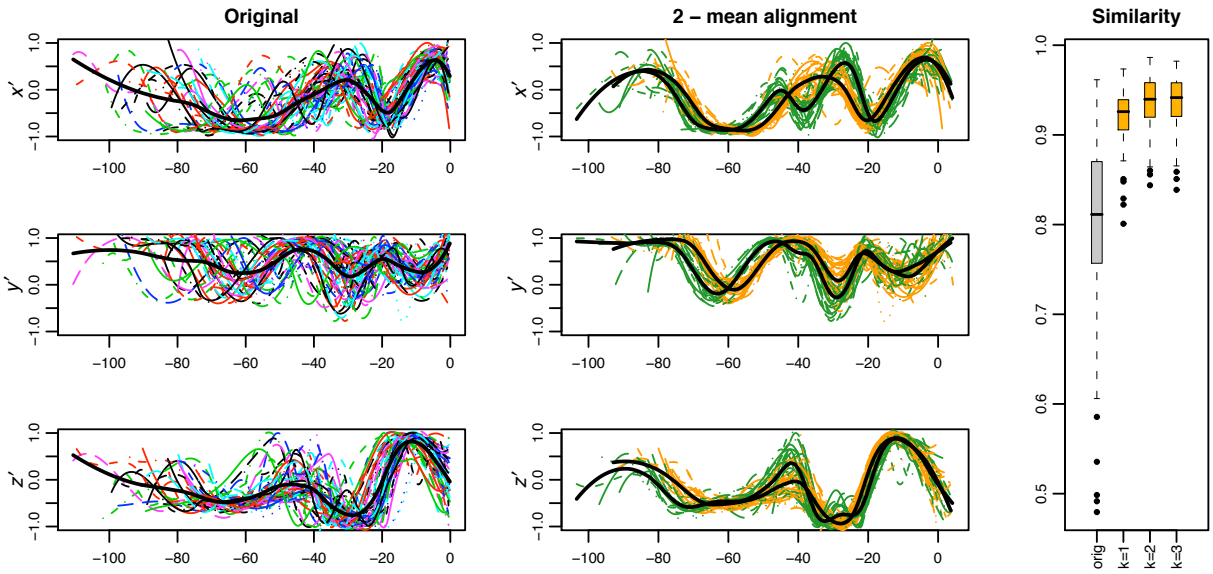


FIGURE 1. *Left : first derivatives x', y', z' of the three spatial coordinates of ICA centerlines. Center : first derivatives of 2-mean aligned ICA centerlines. Right : boxplots of similarity indexes for original curves and k-mean aligned curves, $k = 1, 2, 3$.*

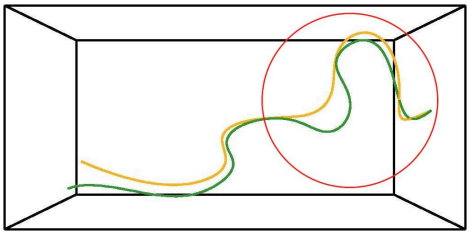


FIGURE 2. *Three-dimensional image of the estimated templates of the 2 clusters found by 2-mean alignment of ICA centerlines.*

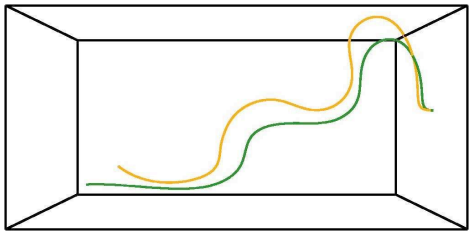


FIGURE 3. *Three-dimensional image of the estimated templates of the 2 clusters found by simple 2-mean clustering without alignment of ICA centerlines.*

6. Bibliographie

[KRA 82] KRAYENBUEHL H., HUBER P., YASARGIL M. G., *Krayenbuhl/Yasargil Cerebral Angiography*, Thieme Medical Publishers, 2 édition, 1982.

[SAN 08] SANGALLI L. M., SECCHI P., VANTINI S., VITELLI V., K-means alignment for curve clustering, rapport n^{o}13/2008, 2008, MOX, Dipartimento di Matematica, Politecnico di Milano, http ://mox.polimi.it/it/progetti/pubblicazioni.

[SAN 09] SANGALLI L. M., SECCHI P., VANTINI S., VENEZIANI A., A Case Study in Exploratory Functional Data Analysis : Geometrical Features of the Internal Carotid Artery, *J. Amer. Statist. Assoc.*, vol. 104, n^o 485, 2009, p. 37–48.

Multiple Comparison Procedures for Treatment Classification within ANOVA layout

Livio Corain, Luigi Salmaso, Francesca Solmi

Department of Management and Engineering - University of Padova
Str. S. Nicola 3, 36100 Vicenza, Italy. Email: livio.corain@unipd.it, salmaso@gest.unipd.it

Department of Statistics - University of Padova
Via Cesare Battisti 241, 35121 Padova, Italy. Email: solmi@stat.unipd.it

ABSTRACT. *The topic of Multiple Comparison Procedures (MCPs) is receiving nowadays a growing theoretical and applicative attention in both industrial and scientific problems. This issue arises within ANOVA layout when we consider a response variable of interest and we wish to compare more than two treatments in order to find out possible significant differences among them. Accordingly, the hypothesis testing problem is connected with a multiple set of individual statistical tests giving rise to the so called multiplicity issue, that is the problem to control the inferential errors of the whole set of null hypotheses of interest. This paper aims at proposing the use of the closed testing approach instead of standard MCPs, showing the advantages of using the closed testing approach in order to improve the capability of right treatment classification. A simulation study has been developed where a number of different data scenarios has been considered, and the closed testing approach has been compared with the standard Bonferroni and powerful Shaffer procedures. The advantage of using the closed testing approach has been confirmed.*

KEYWORDS: *Closed testing, Familywise Error, Multiple testing.*

1. Multiple comparison procedures

The topic of Multiple Comparison Procedures (MCPs) is receiving nowadays a growing theoretical and applicative attention in both industrial and scientific problems. This problem is not only of theoretical interest but also it has a recognized practical relevance. In fact, especially for industrial research, a ranking in terms of performance of all investigated products/prototypes is a very natural goal. For example, a chemical company operating in the field of detergents deals with the problem of detecting the "best" detergent from a set of proposed prototypes. In this context the experimental performance is the reflectance, i.e. the percentages of removed soil from a piece of fabric which, before being put in the washing machine, is soiled with certain types of soil which are considered to be representative of everyday domestic laundry.

The MCPs are commonly used to control the so called Familywise Error Rate (FWER). The strong form of FWER is the probability of rejecting al least one true null hypothesis H_{0i} contained in a subset of true null hypothesis S ; that is formally :

$$\text{FWE}(S) = \Pr(\text{Reject al least one } H_{0i},\ i \in S | H_{0i} \text{ is true for all } i \in S).$$

In this paper we motivate the use of the closed testing approach ([MAR 76]) in comparison to standard controlling FWER MCPs, showing the advantages of the use of this method in order to increase the power of the multiple comparison procedure. In fact the closed testing method allows us to take advantage from the logical dependences among the minimal hypotheses of interest and so to adjust the *p*-values often in a less conservative way than the other standard procedures.

Moreover, an analysis of recent literature has underlined the lack of exhaustive comparative studies among standard parametric methods, such as classical Bonferroni-Holm ([HOL 79]) or Shaffer ([SHA 86]) methods, and the procedures based on the closed testing approach. At first we will present the multiplicity issue then we will present in details the closed testing approach, discussing its principal features, and finally an exhaustive simulation study will be developed in order to compare the closed testing approach with some standard MCPs.

In what follows we will refer to the one-way ANOVA problem in which n units are randomly assigned to C treatments or groups and a numeric response variable X is observed. Groups are supposed to have the same size (balanced design). Units belonging to the jth group are presumed to receive a treatment at the jth level. We will refer to the so called fixed effect model, i.e.

$$Y_{ij} = \mu_j + \varepsilon_{ij} = \mu + \tau_j + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim IID(0, \sigma^2),\ i = 1, ..., n,\ j = 1, ..., C, \qquad [1]$$

where τ_j and Y_{ij}, are respectively the effect of the j-th treatment and the response variable for the ith replicate and the jth treatment. The random term ε_{ij} represents the experimental error with zero mean, variance σ^2 and unknown continuous distribution P. The usual side-conditions for effects are given by the constrains $\sum_j \tau_j = 0$.

The resulting inferential problem of interest is concerned with the following hypotheses : $H_0 : \{\tau_j = 0 \,\forall\, j\}$, against $H_1 : \{\exists j : \tau_j \neq 0\}$. Note that this hypothesis is referred to a global test ; if H_0 is rejected, it is of interest to perform inference on each pair-wise comparison between couples of treatments, i.e. $H_{0(jh)} : \tau_j = \tau_h,\ j, h = 1, ..., C, j \neq h$, against $H_{1(jh)} : \tau_j \neq \tau_h$; with reference to model 1, an equivalent representation of $H_{0(jh)}$ is the following : $H_{0(jh)} : \mu_j - \mu_h = 0, j, h = 1, ..., C, j \neq h$, against $H_{1(jh)} : \mu_j - \mu_h \neq 0$. Accordingly, the hypothesis testing problem is connected with a multiple set of $C(C-1)/2$ individual statistical tests giving rise to the so called multiplicity issue, that is the problem to control the inferential errors of the whole set of null hypotheses of interest.

Note that we are considering the usual two-sided alternatives. In case, by the purpose of the investigation or prior reasons, it would be of interest to refer to one-sided tests, we remark that the analysis should take into account for the so called Type III error ([SHA 02]), involving the incorrectly inference on the direction of the effect.

2. The closed testing approach

The closed testing approach represents an interesting solution for multiple comparisons problems also in a treatment classification context. The strength of the method is its quite general definition together with its capability of taking coherent decisions : this approach is able to maintain some logical consistence in the taken decisions on the hypotheses of interest, being it defined such as the coherence property holds. The property of coherence is said to be satisfied when, in a hierarchical family of hypotheses, the acceptance of a given null hypothesis implies the acceptance of each null hypothesis which is included in that accepted hypothesis ([MAR 76]). The closed testing approach has been first proposed in [MAR 76] ; let us consider the whole of the minimal hypotheses $\{H_i, i = 1, ..., k\}$, and let us define as *closure* the whole $\left\{H_p = \bigcap_{i \in p} H_i, p \in \{1, ..., k\}\right\}$ of all the not empty intersections of the H_i. Thus the closed testing procedure operates according to the following steps :

i. test each minimal hypothesis H_i at an α significance level ;

ii. test each intersection hypothesis H_p of the closure whole with an appropriate significance level. Obviously these composite tests have to be made with appropriate type of test statistics, so for a test that includes two or more hypothesis an F test statistic or a simile one might be used ;

iii. assure the control of FWER rejecting each initial null hypothesis if the test related to the single hypothesis is significant and also all the hypotheses that include it are significant too.

Note that the above given definition is quite general. A theorem in [MAR 76] demonstrates that the procedure controls the FWER in the strong sense.

3. Simulation study and results

A simulation study has been developed in order to compare the closed testing method with other standard parametric procedures, in the different data scenarios ; we have compared the procedures in terms of right treatment classification capability. The methods of Bonferroni (hereafter B, in order to give a baseline for the right classification percentages), Shaffer (hereafter S2, in order to have a powerful competitor) and closed testing (CT) have been considered in the study. As regards the testing of the single partial null hypotheses we have worked with standard t tests, in order to treat the problem from a parametric point of view, and two ad hoc procedures have been written for the implementation of the closed testing approach. It is important to underline that the testing of all pair-wise comparisons implies that we have to do with a strong dependence structure of the minimal hypotheses.

We have performed the procedures in different settings as regards the number of treatments to compare ($C = 4$ and $C = 5$), the sample size ($n = 4$ and $n = 8$) and the distribution of the random error (standard normal, exponential and Student's t with 2 degrees of freedom). Moreover we have simulated some appropriate experimental settings in order to consider different situations also in terms of the number of true null hypotheses in the family. Interesting results have been obtained in the study ; Tables 1 and 2 summarize the behaviour of the procedures in different data scenarios, in which we report the percentage of right classification (All-RCP) of the whole model, and the actual value for the FWER (under H_0).

Table 1. *Results of the simulation study in the case of $\alpha = 0.05$, $C = 4$ treatments to compare, sample size $n = 8$, different settings and distributions of the data : all-treatments right classification percentage (All-RCP) and actual FWER.*

		Normal errors			Exponential errors			Students t errors		
		B	S2	CT	B	S2	CT	B	S2	CT
Setting 1 (None true H_0)	All-RCP	0.5668	0.6807	0.6935	0.5907	0.7097	0.7202	0.2018	0.2537	0.2903
Setting 2 (1 true H_0 on 6)	All-RCP	0.4790	0.5726	0.6080	0.5102	0.6108	0.6430	0.1336	0.1654	0.2080
	FWER	0.0050	0.0280	0.0330	0.0090	0.0420	0.0440	0.0050	0.0140	0.0180
Setting 3 (3 true H_0 on 6)	All-RCP	0.2213	0.2700	0.3450	0.2783	0.3430	0.4143	0.0493	0.0600	0.0870
	FWER	0.0200	0.0370	0.0580	0.0210	0.0400	0.0600	0.0100	0.0130	0.0340

Table 2. *Results of the simulation study in the case of $\alpha = 0.05$, $C = 5$ treatments to compare, sample size $n = 8$, different settings and distributions of the data : all-treatments right classification percentage (All-RCP) and actual FWER.*

		Normal errors			Exponential errors			Students t errors		
		B	S2	CT	B	S2	CT	B	S2	CT
Setting 1 (None true H_0)	All-RCP	0.6238	0.7302	0.6918	0.6279	0.7333	0.6992	0.2421	0.3006	0.2723
Setting 2 (1 true H_0 on 10)	All-RCP	0.5803	0.6718	0.7009	0.5906	0.6823	0.7088	0.1907	0.2368	0.2574
	FWER	0.0040	0.0260	0.0360	0.0050	0.0270	0.0350	0.0050	0.0100	0.0110
Setting 3 (3 true H_0 on 10)	All-RCP	0.4650	0.5434	0.5966	0.4676	0.5473	0.5976	0.1076	0.1360	0.1479
	FWER	0.0130	0.0350	0.0220	0.0130	0.0370	0.0340	0.0120	0.0200	0.0140
Setting 4 (6 true H_0 on 10)	All-RCP	0.1792	0.2300	0.2505	0.2140	0.2563	0.2700	0.0352	0.0443	0.0527
	FWER	0.0290	0.0410	0.0410	0.0120	0.0180	0.0410	0.0190	0.0210	0.0160

In general the obtained results give a very strong indication for the advantage of using the closed testing procedure in the considered data scenarios both for the case of 4 and 5 treatments to be compared. The closed testing method reports higher percentages of right classification of the model for all the considered settings for the data. Moreover the percentage of right treatments classification of all the considered procedures is shown to

decrease in the case of nonnormal distribution of the data, and also as the simulated sample size decreases ; this behaviour is surely due to the use of parametric tests.

In the light of these results, the application of the closed testing approach in order to cope with classification problems is recommended. Moreover, as regards possible future works, the general definition of the closed testing method allows the application of nonparametric methods (for example permutation tests and nonparametric combinations of dependent permutation tests, [PES 01]) to test the minimal and composite hypotheses of the family ; it could be of interest to compare parametric and nonparametric solutions in those cases in which data can not attend the strong assumptions for the application of standard parametric instruments, or the dependence structure of the minimal hypotheses is complicated and difficult to cope with.

Acknowledgements : This research has been supported by University of Padova grant CPDA088513.

4. References

[BEN 95] BENJAMINI Y., HOCHBERG Y., "Controlling the false discovery rate : A practical and powerfull approach to multiple testing", *Journal of the Royal Statistical Society (Ser B.)*, vol. 57, 1995, p. 289-300.

[DON 04] DONOGUE R. J., "Implementing Shaffer's multiple comparison procedure for a large number of groups", *Recent Developments in Multiple comparison Procedures, Institute of Mathematical Statistics*, vol. 47, 2004, p. 1-23.

[HOC 87] HOCHBERG Y., TAMHANE A. C., *Multiple comparison procedures*, Wiley, 1987.

[HOL 79] HOLM S., "A simple sequentially rejective multiple test procedure", *Scandinavian Journal of Statistics*, vol. 6, 1979, p. 65-70.

[HOR 04] HORN M., DUNNET C. W., "Power and sample size comparisons of stepwise FWE and FDR controlling test procedures in the normal many-one case", *Recent Developments in Multiple comparison Procedures, Institute of Mathematical Statistics*, vol. 47, 2004, p. 48-64.

[KES 02] KESELMAN H. J., A. C. R., WILCOX R. R., "Pairwise multiple comparison tests when data are nonnormal", *Educational and Psychological Measurement*, vol. 62, 2002, p. 420-434.

[MAR 76] MARCUS R., PERITZ E., GABRIEL K. R., "On closed testing procedures with special reference to ordered analysis of variance", *Biometrika*, vol. 63, 1976, p. 655-660.

[PES 01] PESARIN F., *Multivariate permutation tests with applications in biostatistics*, Wiley, 2001.

[RAF 02] RAFTER J. A., ABELL M. L., BRASELTON J. P., "Multiple comparison methods for means", *Society for Industrial and Applied Mathematics*, vol. 44, 2002, p. 259-278.

[RAM 78] RAMSEY P. H., "Power differences between pairwise multiple comparisons. Comments", *Journal of the American Statistical Association*, vol. 73, 1978, p. 479-487.

[SEA 97] SEAMAN M. A., "Tables for pairwise multiple comparisons using Shaffer's modified sequentially-rejective procedure", *Communications in Statistics - Simulation and Computation*, vol. 26, 1997, p. 687-705.

[SHA 86] SHAFFER J. P., "Modified sequentially rejective multiple test procedure", *Journal of the American Statistical Association*, vol. 81, 1986, p. 826-831.

[SHA 95] SHAFFER J. P., "Multiple hypothesis testing", *Annual Review of Psychology*, vol. 46, 1995, p. 561-584.

[SHA 02] SHAFFER J. P., "Multiplicity, directional (Type III) errors, and the null hypothesis", *Psychological methods*, vol. 7, 2002, p. 356-369.

[VAL 01] VALLEJO SECO G., MENÈNDEZ DE LA FUENTE I. A., ESCUDERO J. R., "Pairwise multiple comparisons under violation of the independence assumption", *Quality & Quantity*, vol. 35, 2001, p. 71-76.

[WES 93] WESTFALL P. H., S. Y. S., *Resampling-based multiple testing : Examples and methods for p-values adjustment*, Wiley, 1993.

[WES 97] WESTFALL P. H., "Multiple testing of general contrasts using logical constraints and correlations", *Journal of the American Statistical Association*, vol. 92, 1997, p. 299-306.

[WES 99] WESTFALL P. H., TOBIAS R. D., ROM D., WOLFINGER R. D., Y. H., *Multiple comparisons and multiple tests using SAS*, SAS Institute Inc., 1999.

Dynamic clustering of data described by multivariate distributions using the Jensen-Shannon dissimilarity

Francesca Condino[1,2], Antonio Irpino[3], Rosanna Verde[3], Filippo Domma[4]

[1]*University of Naples Federico II, Italy*
[2]*Institute of Neurological Sciences, National Research Council, Italy*
[3]*Second University of Naples, Italy*
[4]*University of Calabria, Italy*

ABSTRACT. In multi-valued analysis context, data can be described by intervals or distributions. In order to compare data described by distributions, we investigate the Jensen-Shannon divergence measure and its properties. Because it is based on joint distribution functions, this measure allows to consider the dependence between marginal distributions inside the observations. We propose to use copula functions as a tool for investigating the dependence structure between marginals. Copula enables to model the marginal distributions and the dependence structure separately and this feature allows to have a more flexible procedure to modeling joint distributions. Moreover, through copula, we can consider useful information, often available in symbolic data-set, about dependence among variables. Then, we propose to use Jensen-Shannon dissimilarity in dynamic clustering algorithm to find the best partition according to a criterion function, in order to guarantee cohesion inside groups. Indeed, in this context, we show that it is possible simultaneously minimize the Jensen-Shannon divergence within clusters and maximize the Jensen-Shannon divergence between clusters. We prove that it is possible to obtain the barycentre of each cluster as a mixture of densities or mass functions belonging to the cluster itself.

KEYWORDS : Shannon-Jensen Divergence, Multivalued data, Dynamic clustering, Copula function.

1. Introduction

The concept of clustering has been extended now-a-days to the patterns described by unconventional data called symbolic data [BOC 00]. In this context, data can be described by intervals or distributions, so we have a set E of objects described by p multi-valued variables. A common task in this data analysis framework is the detection of homogeneous groups of objects in the set E such that objects belonging to the same group show high similarity degree, in contrast to objects from different groups. In this context, it is necessary to find a dissimilarity measure to evaluate the degree of proximities between two objects [VER 08]. Often, in doing this, the multi-valued variables describing each observation are considered independent, but it is not always true. In the present paper, we proposed to compare data described by multivariate distributions using a suitable dissimilarity measure, which allow us to take in account the dependence between variables. So, copula arise as natural tool to model dependence. The concept of copula initially introduced by Sklar [SKL 59], is now extensively applied in many fields. The main idea behind this concept is in the modelling of the dependence structure and the univariate margins separately, to obtain the multivariate distribution of two or more random variables. After presenting distribution data in section 2, we introduce the Jensen-Shannon divergence as a dissimilarity measure among objects, and we explore the properties and we show some preliminary results. In section 4 we use the Jensen-Shannon divergence in dynamic clustering context. Section 5 is dedicated to the computation procedure.

2. Distribution data

Let *T* a table of data with *n* lines and *p* columns. Suppose that the i^{th} row $\{i=1,...,n\}$ corresponds to an object or a concept denoted by ω_i. Let suppose that ω_i is described by *p* univariate distributions $F_i^{(1)}(\cdot),...,F_i^{(p)}(\cdot)$, known or estimated through classical methods, and let be $f_i^{(1)}(\cdot),...,f_i^{(p)}(\cdot)$ the corresponding density functions. Furthermore, we assume that variables are dependent and we use copula function to obtain multivariate distributions for each object.

3. The Jensen-Shannon dissimilarity

In order to find a dissimilarity measure to evaluate the degree of proximities between ω_i and $\omega_{i'}$ objects, we propose to use the Jensen-Shannon (JS) divergence. This measure is based on the so-called Kullback-Leibler (KL) divergence, defined for continuous variables as:

$$d_{KL}(f_i \mid f_{i'}) = \int_{\mathbf{X}} f_i(\mathbf{x}) \log \frac{f_i(\mathbf{x})}{f_{i'}(\mathbf{x})} d\mathbf{x} \tag{1}$$

where: $f_i(\cdot)$ and $f_{i'}(\cdot)$ are the joint density functions corresponding respectively to the rows *i* and *i'*.

The JS divergence between ω_i and $\omega_{i'}$ can be written as:

$$d_{JS}(f_i, f_{i'}) = \pi d_{KL}(f_i \mid m) + (1-\pi) d_{KL}(f_{i'} \mid m) \tag{2}$$

where: $\pi \in [0,1]$ is a mixing parameter and $m = \pi \cdot f_i + (1-\pi) \cdot f_{i'}$ is a mixture of densities $f_i(\cdot)$ and $f_{i'}(\cdot)$. It is easy to prove that $d_{JS}(f_i, f_{i'}) \geq 0$ and equality holds when $f_i = f_{i'}$. In addition it is symmetric, i.e. $d_{JS}(f_i, f_{i'}) = d_{JS}(f_{i'}, f_i)$, and then it is a bonafide measure of dissimilarity between $f_i(\cdot)$ and $f_{i'}(\cdot)$.

JS divergence is related to the entropy concept. Indeed, expression (2) can be rewritten as

$$\begin{aligned} d_{JS}(f_i, f_{i'}) &= H(\pi f_i + (1-\pi) f_{i'}) - \pi H(f_i) - (1-\pi) H(f_{i'}) = \\ &= H(m) - \pi H(f_i) - (1-\pi) H(f_{i'}) \end{aligned} \tag{3}$$

where $H(f) = -\int_{\mathbf{X}} f(\mathbf{x}) \log f(\mathbf{x})\, d\mathbf{x}$ denotes the Shannon entropy [SHA 71] for the generic density *f*.

Extending the JS divergence to the case of *n* density functions, we get:

$$d_{JS}(f_i,...,f_n) = H\left(\sum_{i=1}^{n} \pi_i f_i\right) - \sum_{i=1}^{n} \pi_i H(f_i) = H(m) - \sum_{i=1}^{n} \pi_i H(f_i) \tag{4}$$

with $\pi_i \in [0,1] \quad \forall i$, and $\sum_{i=1}^{n} \pi_i = 1$.

Because joint entropy $H(f_i)$ can be decomposed in a first part, due to marginal entropies $H\left(f_i^{(j)}\right)$, and in a second one, due to mutual information [PAP 91], and because mutual information is equal to negative copula entropy, considering copula approach, we can write entropy $H(f_i)$ as:

$$H(f_i) = \sum_{j=1}^{p} H\left(f_i^{(j)}\right) + H(c_i) \tag{5}$$

where c_i is the copula density function corresponding to the row *i* and $H(c_i)$ is the copula entropy.

Moreover, we have $H(m) = H\left(\sum_{i=1}^{n} \pi_i c_i \prod_{j=1}^{p} f_i^{(j)}\right)$ and we can rewrite the JS dissimilarity in terms of copula functions and marginal distributions.

4. Dynamic clustering based on Jensen-Shannon dissimilarity

The dynamic clustering algorithm (DCA) is a non-hierarchical iterative algorithm characterized by two steps: the construction of clusters (allocation step) and the identification of a representative object, or prototype, (representation step) for each cluster. These two steps aim to partition objects by optimizing an adequacy criterion that measures the fitting between the clusters and their corresponding representatives (prototypes). The aim of DCA is then to find a partition $P^* = (C_1,...,C_k)$ of *E* in *k* cluster and a vector $L^* = (G_1,...,G_k)$ of *k* prototypes so that a criterion function Δ is optimized:

$$\Delta(P^*,L^*) = Min\{\Delta(P,L) \mid P \in P_k, L \in L_k\} \tag{6}$$

with P_k the set of all possible partitions of *k* dimension and L_k the set of all possible prototypes vectors. The criterion Δ is defined as:

$$\Delta(P,L) = \sum_{h=1}^{k} \sum_{\omega_i \in C_h} D(\omega_i, G_h) \tag{7}$$

where $D(\omega_i, G_h)$ is the dissimilarity measure between the generic object ω_i and the prototype G_h. In our context, $D(\omega_i, G_h)$ will be a weighted sum of KL distances between each joint density function f_i describing the object $\omega_i \in C_h$ and the mixture of densities describing all objects belonging to C_h:

$$\Delta(P,L) = \sum_{h=1}^{k} \sum_{\omega_i \in C_h} \pi_i d_{KL}(f_i \mid m_h) \tag{8}$$

with $m_h = \sum_{\omega_i \in C_h} \frac{\pi_i}{\pi^{(h)}} f_i$ and $\pi^{(h)} = \sum_{\omega_i \in C_h} \pi_i$. If there are no reasons to chose different weights for the objects, π_i can be supposed constant and equal to $1/n$.

It can be shown that optimizing criterion Δ is equivalent to minimizing JS divergence within cluster and contextually maximizing JS divergence between clusters. Indeed, one of the JS divergence's properties is that the total divergence, i.e. the divergence among all considered objects, can be decomposed in two quantities, one relates to the dissimilarities in each cluster and the other reflecting the dissimilarities among cluster:

$$d_{JS}(f_1,...,f_n) = \sum_{h=1}^{k} \sum_{\omega_i \in C_h} \pi_i d_{KL}(f_i|m_h) + \sum_{h=1}^{k} \pi^{(h)} d_{KL}(m_h|m) = d_{JS}^W + d_{JS}^B \tag{9}$$

Moreover, the mixture m_h is such that minimizes JS divergence within h^{th} cluster.

The final partition's quality can be evaluated using the index obtained by ratio between d_{JS}^{B} and total JS divergence, analogously to the way proposed by Chavent et al. [CHA 03]

5. Computation procedure

According to (4), the JS divergence among objects belonging to h^{th} cluster has the following expression:

$$d_{JS}^{(h)} = H(m_h) - \sum_{\omega_i \in C_h} \frac{\pi_i}{\pi^{(h)}} H(f_i) \tag{10}$$

and after simple passages, we obtained the following expression for $d_{JS}^{(h)}$:

$$d_{JS}^{(h)} = H(m_h) - \sum_{\omega_i \in C_h} \frac{\pi_i}{\pi^{(h)}} \left[\sum_{j=1}^{p} H\left(f_i^{(j)}\right) + H(c_i) \right] \tag{11}$$

Then, we can easily compute JS dissimilarities among objects in a cluster, computing copula entropy, marginal entropies and mixture entropy separately. To obtain these quantities, numerical integration procedure, based on adaptive methods, can be used. Subsequently, the d_{JS}^{W} quantity can be computed.

The proposed clustering algorithm allows us to find simultaneously the best partition of symbolic objects, according to the chosen criterion, and a suitable model to describing dependence inside observations.

6. Bibliography

[BOC 00] BOCK H.H., DIDAY E., Analysis of Symbolic Data, Expanatory methods for extracting statistical informations from Complex data, *Studies in Classification, Data Analysis and Knowledge Organization*, Springer Verlag, 2000.

[CHA 03] CHAVENT M., DE CARVALHO F.A.T., LECHEVALLIER Y., VERDE R., Trois nouvelles méthodes de classification automatique des données symbolique de type intervalle, *Revue de Statistique Appliquée*, vol. 4, 2003, p. 5-29.

[PAP 91] PAPOULIS A., *Probability, Random Variables and Stochastic Process,* McGraw-Hill,1991.

[SHA 71] SHANNON C.E., WEAVER W., *La teoria matematica delle comunicazioni,* Etas Kompass, 1971.

[SKL 59] SKLAR A., Fonctions de répartition à n dimension et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris*, vol. 8, 1959, p. 229-231

[VER 08] VERDE E., IRPINO A., Comparing Histogram Data Using Mahalanobis-Wasserstein Distance, *Proceeding in Compstat 2008: Proceedings in Computational Statistics,* Heidelberg, Physica-Verlag Springer, 2008.

Correspondence analysis with linear constraints of cross-classification tables using orthogonal polynomials

Pietro Amenta

Department of Analysis of Economic and Social Systems, University of Sannio
Via delle Puglie, 82
82100, Benevento, Italy
amenta@unisannio.it

ABSTRACT. *Within the context of the non-iterative procedures for performing correspondence analysis with linear constraints, a strategy is proposed to impose linear constraints in analyzing a contingency tables with one or two ordered sets of categories. At the heart of the approach is the partition of the Pearson chi-squared statistic which involves terms that summarize the association between the nominal/ordinal variables using bivariate moments based on orthogonal polynomials. Linear constraints are then included directly on suitable matrices which reflect the most important components overcoming the problem to impose linear constraints based on subjective decisions. A possible use of this constrained two-way approach for sliced three ordered sets of categories is also suggested.*

KEYWORDS : *Ordered Correspondence Analysis, Emerson's orthogonal polynomials, linear constraints.*

1. Introduction

Correspondence analysis is a widely used tool for obtaining a graphical representation of the dependence between the rows and columns of a contingency table, and it is usually performed by applying a singular value decomposition to the standardised residuals of a two-way contingency table. This decomposition ensures that the maximum information regarding the association between the two categorical variables are accounted for in a factorial plane of a correspondence plot. However, such a plot can identify those categories that are similar but does not clarify how some categories are different. In addition, the interpretation of the multidimensional representation of the row and column categories may be greatly simplified if additional information (as linear constraints) about the row and column structure of the table is available. In the classical analysis, Böckenholt and Böckenholt (hereafter B&B) [BOC 90] considered this problem (see also [BEH 09] [TAK 09] [AME 08a]). This additional information is usually imposed by making use of orthogonal polynomials which are suitable for subdividing total variation of the scores into linear, quadratic, cubic, etc., components. For instance, to obtain a linear order for the standard scores, B&B eliminates the effects of the quadratic and cubic trend by means suitable constraint matrices. Unfortunally, these constraints are commonly selected on the basis of subjective decisions without taking into account if the effects of the linear, quadratic and cubic trend are or not statistically significant, respectively. Aim of this paper is to consider a suitable extension of the B&B's approach to contingency tables with more than one ordered sets of categories. This is achieved by using the additional information about the structure and statistically significant associations of the data given by the correspondence analysis proposed by Beh [BEH 97].

2. Correspondence analysis of ordinal cross-classifications based on the moment decomposition

Consider a two-way contingency table $\mathbf{N}$ describing the joint distribution of two categorical variables where the $(i,\, j)$th cell entry is given by n_{ij} for $i = 1, ..., I$ and $j = 1, ..., J$ with $n = \sum_{i,j} n_{ij}$. The $(i,\, j)$th element

of the probability matrix $\mathbf{P}$ is defined as $p_{ij} = n_{ij}/n$ so that $\sum_{i,j} p_{ij} = 1$. Suppose that $\mathbf{N}$ has ordered row and column categories with row and column marginal probabilities given by $p_{i.} = \sum_j p_{ij}$ and $p_{.j} = \sum_i p_{ij}$, respectively, with $\mathbf{D}_I = \text{diag}(p_{i.})$ and $\mathbf{D}_J = \text{diag}(p_{.j})$. Usually, Correspondence Analysis of cross-classifications regarding the association between two categorical variables can be usually performed by applying Singular Value Decomposition (SVD) on the Pearson ratio's of the table [GOO 96]. That is, for the $I \times J$ correspondence matrix $\mathbf{P}$ then its Pearson ratio α_{ij} is decomposed so that $\alpha_{ij} = p_{ij}/(p_{i.}p_{.j}) = 1 + \sum_{m=1}^{\min(I,J)-1} \lambda_m a_{im} b_{jm}$ with a_{im} and b_{jm} $\{m = 1, \ldots, K = \min(I, J) - 1\}$ singular vectors associated with the i'th row and j'th column category, respectively. Let λ_m be the m'th singular value of the ratio. Moreover, we have that $\sum_j p_{.j} b_{jm} = \sum_i p_{i.} a_{im} = 0$ and $\sum_j p_{.j} b_{jm} b_{jm'} = \sum_i p_{i.} a_{im} a_{im'} = 1$ for $m = m'$, 0 otherwise. Using the matrix notation, the above least squares estimates are obtained by a SVD of the matrix $\mathbf{\Pi} = \mathbf{D}_I^{-1/2}(\mathbf{P} - \mathbf{D}_I \mathbf{1}\mathbf{1}^T \mathbf{D}_J)\mathbf{D}_J^{-1/2} = \mathbf{A}\mathbf{\Lambda}\mathbf{B}^T$ with $\mathbf{A}^T\mathbf{A} = \mathbf{I}$, $\mathbf{B}^T\mathbf{B} = \mathbf{I}$ and $\mathbf{\Lambda}$ diagonal matrix where singular values λ_m are in descending order.

Using a different approach, *Double Ordered Correspondence Analysis* [BEH 97] decomposes the (i, j)th Pearson ratio α_{ij} so that $\alpha_{ij} = 1 + \sum_{u=1}^{I-1} \sum_{v=1}^{J-1} a_u(i) b_v(j) z_{uv}$. For this method of decomposition, $z_{uv} = \sqrt{n} \sum_{i,j} a_u(i)\, b_v(j)\, p_{ij}$ is the (u, v)th generalised correlation [DAV 03] where $\{a_u(i) : u = 1, \ldots, I-1\}$ and $\{b_v(j) : v = 1, \ldots, J-1\}$ are the orthogonal polynomials [EME 68] for the i-th row and j-th column respectively. The bivariate association z_{uv} are collected in $\mathbf{Z} = \sqrt{n}\mathbf{A}_*^T \mathbf{P} \mathbf{B}_*$ where $\mathbf{A}_*$ contains the $I - 1$ non-trivial row orthogonal polynomials and $\mathbf{B}_*$ is the $J \times (J-1)$ matrix of the $J - 1$ non-trivial column orthogonal polynomials. The matrix $\mathbf{\Pi}$ can be then rewritten as $\mathbf{\Pi}_D = \mathbf{A}_*(\mathbf{Z}/\sqrt{n})\mathbf{B}_*^T$ with $\mathbf{A}_*^T \mathbf{D}_I \mathbf{A}_* = \mathbf{I}$ and $\mathbf{B}_*^T \mathbf{D}_J \mathbf{B}_* = \mathbf{I}$. It is possible to show [BEH 97] that the elements of $\mathbf{Z}$ (that is, the bivariate associations z_{uv}) are asymptotically standard normal and indipendent. In addition, [RAY 96] showed that the Pearson chi-squared statistic can be decomposed into the sum of squares of the generalized correlations so that $\chi^2/n = \sum_{u=1}^{I-1} \sum_{v=1}^{J-1} z_{uv}^2$ and sources of variation for the row and column profiles can be obtained. Observe that the above chi-squared index is partitioned into $(I-1)(J-1)$ terms, where the significance of each term can be compared with the χ^2 with one degree of freedom (dof). Sources of variation for the row and column profiles can be easily obtained. For instance, any difference in the row profiles in terms of their location is computed by $\sum_{v=1}^{J-1} z_{1v}^2$ while the row dispersion component is given by $\sum_{v=1}^{J-1} z_{2v}^2$ .The significance of each component can be compared with the χ^2 with $(J-1)$ dof. Similarly, location and dispersion column components can be computed. This approach to correspondence analysis uses then the bivariate moment decomposition to identify linear (location), quadratic (dispersion) and higher order moments. Note that this feature is not readily available by using classical SVD. An alternative approach to partitioning the Pearson chi-squared statistic for a two-way contingency table with one ordered set of categories (*Singly Ordered Correspondence Analysis*) [BEH 97], combines the approach of orthogonal polynomials for the ordered columns and singular vectors for the unordered rows, such that $\chi^2 = \sum_{u=1}^{M^*} \sum_{v=1}^{J-1} \mathbf{Z}_{(u)v}^2$ with $M^* \leq I - 1$ and where $z_{(u)v} = \sqrt{n} \sum_{i,j} p_{ij} a_{iu} b_v(j)$ are asymptotically standard normally distributed random variables. The parentheses around u indicates that the above formulas are concerned with a non-ordered set of row categories. Quantities $z_{(u)v}$ can be written in matrix notation as $\mathbf{Z} = \sqrt{n}\mathbf{A}^T \mathbf{P} \mathbf{B}_*$ where $\mathbf{A}$ is the $I \times (I-1)$ matrix of left singular vectors. The value of $z_{(u)v}$ means that each principal axis from a simple correspondence analysis can be partitioned into column component values. In this way, the researcher can determine the dominant source of variation of the ordered columns along a particular axis using the simple correspondence analysis. The Pearson ratio is given by $\alpha_{ij} = \sum_{u=0}^{M^*} \sum_{v=0}^{J-1} a_{iu}(z_{(u)v}/\sqrt{n}) b_v(j)$. Eliminating the trivial solution, the matrix $\mathbf{\Pi}$ can be also rewritten as $\mathbf{\Pi}_S = \mathbf{A}(\mathbf{Z}/\sqrt{n})\mathbf{B}_*^T$ with $\mathbf{A}^T\mathbf{A} = \mathbf{I}$ and $\mathbf{B}_*^T \mathbf{D}_J \mathbf{B}_* = \mathbf{I}$. For deeper information refer to [BEH 97] [BEH 08].

3. Correspondence analysis with linear constraints of ordinal cross-classifications

B&B [BOC 90] proposed a canonical analysis of contingency tables (CCALC) which takes into account additional information (as linear constraints) about the row and column categories of the table. Let $\mathbf{H}$ and $\mathbf{G}$ be the matrices of linear constraints of order $I \times K$ and $J \times L$ of ranks K and L, respectively, such that $\mathbf{H}^T\mathbf{X} = \mathbf{0}$ and $\mathbf{G}^T\mathbf{Y} = \mathbf{0}$ where $\mathbf{X}$ and $\mathbf{Y}$ are the standardized row and column scores. According to the principle of Restricted Eigenvalue Problem [RAO 73], constrained CCALC scores are obtained by a SVD of the matrix

$$\{\mathbf{I} - \mathbf{D}_I^{-1/2}\mathbf{H}(\mathbf{H}^T\mathbf{D}_I^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{D}_I^{-1/2}\}\mathbf{\Pi}\{\mathbf{I} - \mathbf{D}_J^{-1/2}\mathbf{G}(\mathbf{G}^T\mathbf{D}_J^{-1}\mathbf{G})^{-1}\mathbf{G}^T\mathbf{D}_J^{-1/2}\} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T$$

where $\mathbf{U}^T\mathbf{U} = \mathbf{I}$, $\mathbf{V}^T\mathbf{V} = \mathbf{I}$ and $\mathbf{\Lambda}$ is a diagonal matrix with eigenvalues λ in descending order. Standardized row and column scores are given by $\mathbf{X} = \mathbf{D}_I^{-1/2}\mathbf{U}$ and $\mathbf{Y} = \mathbf{D}_J^{-1/2}\mathbf{V}$, respectively, such that $\mathbf{X}^T\mathbf{D}_I\mathbf{X} = \mathbf{I}$, $\mathbf{Y}^T\mathbf{D}_J\mathbf{Y} = \mathbf{I}$, $\mathbf{1}^T\mathbf{D}_I\mathbf{X} = \mathbf{0}$ and $\mathbf{1}^T\mathbf{D}_J\mathbf{Y} = \mathbf{0}$. Classical approach to correspondence analysis is obtained when $\mathbf{H} = [\mathbf{D}_I|\mathbf{1}]$ (that is, $\mathbf{H}$ is obtained by column-linking $\mathbf{D}_I$ and $\mathbf{1}$) and $\mathbf{G} = [\mathbf{D}_J|\mathbf{1}]$ (absence of linear constraints). Constraints are usually imposed by making use of orthogonal polynomials which are suitable for subdividing total variation of the scores into linear, quadratic, cubic, etc., components. For instance, B&B in CCALC eliminate the effects of the quadratic and cubic trend in order to obtain a linear order for the standard scores by including suitable constraint matrices. We remark that these constraints are commonly selected on the basis of subjective decisions without taking into account if the effects of the linear, quadratic and cubic trend are or not statistically significant. Because the B&B and Beh's approaches decompose the same $(i,\ j)$th Pearson ratio α_{ij}, it is therefore possible to perform a suitable B&B approach working only on the statistically significant trends highlighted by a preliminary ordered correspondence analysis, in order to consider constraints for a two-way cross-classification with nominal and/or ordinal categorical variables extending in this way the B&B's approach performance. Let $\mathbf{\Pi}_D^*$ ($\mathbf{\Pi}_S^*$) be the matrix identifying the most important linear (location), quadratic (dispersion) or higher order moments obtained by the Double (Singly) Correspondence Analysis. For example : $\mathbf{\Pi}^*_{D[1:2,1:2]} = \mathbf{A}_{*[1:2]}\,(\mathbf{Z}/\sqrt{n})\,\mathbf{B}^T_{*[1:2]}$ is the matrix built by taking into account the linear and quadratic components for the row and the column categories, respectively ; $\mathbf{\Pi}^*_{\mathbf{S}[1:2]} = \mathbf{A}\,(\mathbf{Z}/\sqrt{n})\,\mathbf{B}^T_{*[1:2]}$ is the matrix built by taking into account the singular vectors of simple correspondence analysis using a singular value decomposition for the nominal row categories and the linear and quadratic components for the column categories.

Extensions of the B&B approach for singly (hereafter B&B.SO) [AME 08b] and double ordinal cross-classifications (hereafter B&B.DO) [D'A 09] are then obtained by performing the following SVD

$$\{\mathbf{I} - \mathbf{D}_I^{-1/2}\mathbf{H}(\mathbf{H}^T\mathbf{D}_I^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{D}_I^{-1/2}\}\mathbf{\Pi}^*\{\mathbf{I} - \mathbf{D}_J^{-1/2}\mathbf{G}(\mathbf{G}^T\mathbf{D}_J^{-1}\mathbf{G})^{-1}\mathbf{G}^T\mathbf{D}_J^{-1/2}\} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T$$

with $\mathbf{\Pi}^* = \mathbf{\Pi}^*_{\mathbf{S}}$ (B&B.SO) and $\mathbf{\Pi}^* = \mathbf{\Pi}^*_D$ (B&B.DO), respectively, where $\mathbf{U}^T\mathbf{U} = \mathbf{I}$, $\mathbf{V}^T\mathbf{V} = \mathbf{I}$ and $\mathbf{\Lambda}$ is a diagonal matrix with eigenvalues λ that are arranged in descending order. Standardized row and column scores are given by $\mathbf{X} = \mathbf{D}_I^{-1/2}\mathbf{U}$ and $\mathbf{Y} = \mathbf{D}_J^{-1/2}\mathbf{V}$, respectively, such that $\mathbf{X}^T\mathbf{D}_I\mathbf{X} = \mathbf{I}$, $\mathbf{Y}^T\mathbf{D}_J\mathbf{Y} = \mathbf{I}$, $\mathbf{1}^T\mathbf{D}_I\mathbf{X} = \mathbf{0}$ and $\mathbf{1}^T\mathbf{D}_J\mathbf{Y} = \mathbf{0}$. We highlight that the original B&B approach is obtained when we consider all the components of the Beh's chi-squared partition into bivariate moments for the row and column categories. Same result is obtained by considering singular vectors and column orthogonal polynomials. Moreover, classical correspondence analysis of $\mathbf{\Pi}_D^*$ (or $\mathbf{\Pi}_S^*$) is also performed when $\mathbf{H} = [\mathbf{D}_I|\mathbf{1}]$ and $\mathbf{G} = [\mathbf{D}_J|\mathbf{1}]$ (absence of linear constraints).

With this combined approach it is possible to consider linear constraints directly on suitable matrices that reflect the most important components. For instance, given a row/column significant location component, researchers can impose external constraints in order to restrict the spacing to be the same for some modalities in addition to those for ensuring linear order working directly on the obtained row/column significant location component. In addition, the power to identify hidden but statistically significant sources of variation, helps the researcher to improve the interpretation of the data matrix using suitable constraints for these significant components otherwise ignored. Finally, when [RAY 96] and [BEH 97] considered the partition of the chi-squared into bivariate moments, they restricted the analysis to the integer valued scores with the assumption that the ordered categories are equally spaced. In general this may not be the case. By this combined approach, suitable linear constraints can be then introduced in the Beh's analysis in order to take into account not-equally spaced categories without changing the scoring scheme.

This approach can be also useful when we have three ordered sets of categorical variables collected in a three-way contingency table $\mathbf{N}$ where the first two variables are defined according to section 2 and the third one, consisting of K tubes, is denoted with a subscript k. The (i, j, k)th element of the probability matrix $\mathbf{P}$ is then defined as $p_{ijk} = n_{ijk}/n$ so that $\sum_{i,j,k} p_{ijk} = 1$ with $n = \sum_{i,j,k} n_{ijk}$. The Pearson's mean-square contingency coefficient $\Phi^2(I, J, K) = \chi^2/n = \sum_{i,j,k}(p_{ijk} - p_{i..}p_{.j.}p_{..k})^2/(p_{i..}p_{.j.}p_{..k})$ can be partitioned [LAN 51] as $\Phi^2(I, J, K) = \Phi^2(I, J) + \Phi^2(J, K) + \Phi^2(I, K) + \text{int}(I, J, K)$ where $\text{int}(I, J, K)$ is the three way interaction. In terms of squared norm of arrays, this partition can be written [CAR 96] as $||\mathbf{\Pi}||^2 = ||\mathbf{\Pi}_{IJ.}||^2 + ||\mathbf{\Pi}_{I.K}||^2 + ||\mathbf{\Pi}_{..K}||^2 + ||\mathbf{\Pi}_{IJK}||^2$ where $\Phi^2(I, J, K) = ||\mathbf{\Pi}||^2$ with $\mathbf{\Pi} = (\Pi_{i,j,k})$ and $\Pi_{i,j,k} = (p_{ijk}/p_{i..}p_{.j.}p_{..k}) - 1$. The Pearson chi-squared statistic χ^2 can be also partitioned [BEH 98] with respect to the orthogonal polynomials $\{a_u(i)\}$,

$\{b_v(j)\}$ and $\{c_w(k)\}$ associated with the rows, columns and tubes, respectively, as $\chi^2 = \sum_{u=1}^{I-1}\sum_{v=1}^{J-1} z_{uv.}^2 + \sum_{u=1}^{I-1}\sum_{w=1}^{K-1} z_{u.w}^2 + \sum_{v=1}^{J-1}\sum_{w=1}^{K-1} z_{.vw}^2 + \sum_{u=1}^{I-1}\sum_{v=1}^{J-1}\sum_{w=1}^{K-1} z_{uvw}^2$ (or $\chi^2 = \chi_{IJ.}^2 + \chi_{I.K}^2 + \chi_{.JK}^2 + \chi_{IJK}^2$) with $z_{uvw} = \sqrt{n}\sum_{i,j,k} a_u(i)b_v(j)c_w(k)p_{ijk}$, and where each of the $\mathbf{Z}$ terms is asymptotically standard normal and independent. We can then follow the same strategy here proposed to analyze the three-way contingency table $\mathbf{N}$ by imposing linear constraints on each way. Obviously, our approach is only applicable extending two way correspondence analysis to three way case. This can be performed reducing such table to a two way table (I, JK) by the interactive coding, and checking if the condition (that is a negligible value of $\Phi^2(J, K)$) for the following lemma [CHO 88] occurs : $\Phi^2(I, JK) \approx \Phi^2(I, J) + \Phi^2(I, K) + \text{int}(I, J, K)$. In this case we have $\Phi^2(I, J, K) \approx \Phi^2(J, K) + \Phi^2(I, JK)$. We can then perform two constrained correspondence analyses : the former is the B&B.DO of the table (J, K) and the latter is the B&B.SO of table (I, JK). Following [CHO 88] we can also perform a B&B.SO of the table (I, JK) adding the constrained tables (I, J) and (I, K) as supplementary points. Finally, our approach can be used if other negligible terms of the $\Phi^2(I, J, K)$ partition occur.

4. Bibliographie

[AME 08a] AMENTA P., Generalized constrained co-inertia analysis, *Advances in Data Analysis and Classification*, vol. 2, 2008, p. 81–105.

[AME 08b] AMENTA P., SIMONETTI B., BEH E. J., Single Ordinal Correspondence Analysis with External Information, *Asian Journal of Mathematics and Statistics*, vol. 1,1, 2008, p. 34–42.

[BEH 97] BEH E. J., Simple correspondence analysis of ordinal cross-classifications using orthogonal polynomials, *Biometrical Journal*, vol. 39, 1997, p. 589-613.

[BEH 98] BEH E. J., DAVY P. J., Partitioning Pearson's chi-squared statistic for a completely ordered three-way contingency table, *Australian & New Zealand Journal of Statistics*, vol. 40(4), 1998, p. 465–477.

[BEH 08] BEH E. J., Simple Correspondence Analysis of Nominal-Ordinal Contingency Tables, *Journal of Applied Mathematics and Decision Sciences*, vol. 2008 - Article ID 218140, 2008, page doi :10.1155/2008/218140.

[BEH 09] BEH E. J., D'AMBRA L., Some interpretative tools for non-symmetric correspondence analysis, *Journal of Classification*, vol. 26, 2009, p. DOI : 10.1007/s00357-009- 9025-0.

[BOC 90] BOCKENHOLT U., BOCKENHOLT I., Canonical analysis of contingency tables with linear constraints, *Psychometrika*, vol. 55, 1990, p. 633–639.

[CAR 96] CARLIER A., KROONENBERG P., Decompositions and biplots in three-way correspondence analysis, *Psychometrika*, vol. 61, 1996, p. 355–373.

[CHO 88] CHOULAKIAN V., Analyse factorielle des correspondances de tableaux multiples, *Revue de statistique appliquee*, vol. 34, 4, 1988, p. 33–41.

[D'A 09] D'AMBRA A., AMENTA P., Correspondence Analysis with linear constraints of ordinal cross-classifications, *in review*, , 2009.

[DAV 03] DAVY P. J., RAYNER J. C. W., BEH E. J., Generalised correlations and Simpson's Paradox, *Current Research in Modelling, Data Mining and Quantitative Techniques*, V. Pemajayantha, R. W. Mellor, S. Peiris, and J. R. Rajasekera, Eds., University of Western Sydney, Sydney, Australia, 2003, p. 63–73.

[EME 68] EMERSON P. L., Numerical construction of orthogonal polynomials from a general recurrence formula, *Biometrics*, vol. 24, 1968, p. 696–701.

[GOO 96] GOODMAN L. A., A single general method for the analysis of cross-classified data : reconciliation and synthesis of some methods of Pearson, Yule, and Fisher, and also some methods of correspondence analysis and association analysis, *Journal of the American Statistical Association*, vol. 91, 1996, p. 408-428.

[LAN 51] LANCASTER H. O., Complex contingency tables treated by the partition of chi- square, *Journal of Royal Statistical Society B*, vol. 13, 1951, p. 242–249.

[RAO 73] RAO C. R., *Linear statistical inference and its applications*, Wiley, New York, 1973.

[RAY 96] RAYNER J. C. W., BEST D. J., Smooth extensions of Pearson's product moment correlation and Spearman's rho, *Statistics and Probability Letters*, vol. 30, 1996, p. 171–177.

[TAK 09] TAKANE Y., JUNG S., Regularized nonsymmetric correspondence analysis, *Computational Statistics and Data Analysis*, vol. 53, 2009, p. 3159–3170.

Applying Differential Geometric LARS Algorithm to Ultra-high Dimensional Feature Space

Luigi Augugliaro, Angelo M. Mineo

University of Palermo
Dipartimento di Scienze Statistiche e Matematiche,
90128 Palermo, Viale delle Scienze ed. 13
{*augugliaro,elio.mineo*}*@dssm.unipa.it*

RÉSUMÉ. *Variable selection is fundamental in high-dimensional statistical modeling. Many techniques to select relevant variables in generalized linear models are based on a penalized likelihood approach. In a recent paper, Fan and Lv (2008) proposed a sure independent screening (SIS) method to select relevant variables in a linear regression model defined on a ultrahigh dimensional feature space. Aim of this paper is to define a generalization of the SIS method for generalized linear models based on a differential geometric approach.*

MOTS-CLÉS : *LARS, Dimensionality reduction, Variable selection, Differential geometry.*

1. Introduction

Ultrahigh dimensional variable selection plays an important role in regression models applied in many areas of modern scientific research such as microarray analysis, genomics or proteomics. For this kind of problems the number of variables, say p, can be much larger than the sample size n. In this case, it is often assumed that only a small number of variables contributes to the response, which leads to assume the sparsity of the model. For this reason, many variable selection techniques for high dimensional statistical models have been proposed in literature ; most of them are based on a penalized likelihood approach. LASSO estimator proposed by [TIB 96], SCAD method [FAN 01] or L_1-regularization path following algorithm for generalized linear models proposed by [PAR 07] are only some of the most popular methods used to select relevant variables in a generalized linear model [MCC 89]. In a recent paper, [FAN 08] proposed a sure independent screening (SIS) method to select relevant variables in a linear regression model defined on a ultrahigh dimensional feature space. SIS method is based on two steps : in the first step, the authors use the information on marginal correlation to reduce the dimensionality from p to a relative large scale d, lower than the sample size n. In this step, variables that have weak correlation with the response variable are filtered out. In the second step, the authors use a lower dimensional model selection method, such as adaptive LASSO or SCAD method, for the low dimensional submodel defined in the previous step. Since SIS method is based on the geometrical theory underlying the linear regression model, it is closely related with the least angle regression method proposed by [EFR 04]. This observation suggests that a genuine generalization of the SIS method for generalized linear models could be founded on an adequate generalization of the LARS algorithm, namely, on an adequate generalization of the Euclidean geometric interpretation of a linear regression model. Based on this idea, aim of this paper is to define a new statistical method to select relevant variables in a generalized linear model defined in ultrahigh dimensional feature space, which is based on the differential geometrical theory underlying the dgLARS algorithm proposed by [AUG 09]. The rest of the paper is organized as follows. In section 2 we explain the theory underlying the dgLARS method and define the proposed method used to reduce the dimensionality from p to a given value d. In this way, we obtain a unified method to

select the relevant variables in a generalized linear model defined on a ultrahigh dimensional feature space. In section 3 we evaluate the proposed method by a simulation study and draw some conclusions.

2. Differential Geometric LARS (dgLARS) algorithm

Let $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)^T$ be a random variable vector having a probability density function

$$p_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\theta}, \lambda) = a(\mathbf{y}; \lambda) \exp\{\lambda(\mathbf{y}^T \boldsymbol{\theta} - k(\boldsymbol{\theta}))\}, \quad \mathbf{y} \in \mathcal{Y} \subseteq \mathbb{R}^n, \tag{1}$$

with respect to a σ-finite measure ν on $\mathbb{R}^n$, where $a(\cdot)$ and $k(\cdot)$ are specific given functions, $\boldsymbol{\theta}$ varies in the subset $\Theta \subseteq \mathbb{R}^n$ and λ varies in a subset Λ of $\mathbb{R}^+$. The model (1) is called *exponential dispersion model* [JØR 87]. We denote the mean value of $\mathbf{Y}$ by $\boldsymbol{\mu} = \mu(\boldsymbol{\theta})$, where $\mu(\boldsymbol{\theta}) = \partial k(\boldsymbol{\theta})/\partial \boldsymbol{\theta}$ is called *mean value mapping*. Since $\mu(\cdot)$ is a one-to-one function from $int\,\Theta$ onto $\Omega = \mu(int\,\Theta)$, the exponential dispersion model may be parametrized by $(\boldsymbol{\mu}; \sigma^2)$, where $\sigma^2 = \lambda^{-1}$ is called *dispersion parameter*. Following [AMA 85], the parameter space can be treated as a n-dimensional Riemannian manifold where $(\boldsymbol{\mu}; \sigma^2)$ plays the role of coordinate system and the Fisher information matrix is a Riemannian metric (see [CAR 92] for further details). A generalized linear model is completely specified by the following assumptions : (a) $\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$ is a set of n independent observations taken from (1) ; (b) for each random variable Y_i we have a column of covariates $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T \in \mathcal{X} \subseteq \mathbb{R}^p$, with $p < n$. These covariates are related to the mean value of $\mathbf{Y}$ by a known function such that $\mu_i = f(\mathbf{x}_i^T \boldsymbol{\beta})$, where $\boldsymbol{\beta} \in \mathcal{B} \subseteq \mathbb{R}^p$; in order to simplify our notation, we denote $\boldsymbol{\mu}(\boldsymbol{\beta}) = (f(\mathbf{x}_1^T \boldsymbol{\beta}), f(\mathbf{x}_2^T \boldsymbol{\beta}), \ldots, f(\mathbf{x}_n^T \boldsymbol{\beta}))^T$. We assume that $\boldsymbol{\mu}(\cdot)$ is an embedding with domain $\mathcal{B}$; (c) the dispersion parameter σ^2 does not depend on the vector of covariates. Given the assumption (b), $\boldsymbol{\mu}(\mathcal{B}) = \Omega_{\mathcal{B}} \subset \Omega$ is a Riemannian submanifold of Ω, then we can generalize the notion of angle between two given vectors. Let $\boldsymbol{\beta}(\gamma)$ be a differentiable curve, $\ell(\boldsymbol{\beta}(\gamma))$ be the log-likelihood function and $\partial_{\beta_i} \ell(\boldsymbol{\beta}(\gamma))$ be the derivative of $\ell(\boldsymbol{\beta}(\gamma))$ with respect to β_i. Following [KAS 97], we have the identity

$$\partial_{\beta_i} \ell(\boldsymbol{\beta}(\gamma)) = \left\langle \partial_{\beta_j} \boldsymbol{\mu}(\boldsymbol{\beta}(\gamma)); \mathbf{r}(\boldsymbol{\beta}(\gamma)) \right\rangle_{\boldsymbol{\mu}(\beta(\gamma))} \tag{2}$$

where $\langle \cdot, \cdot \rangle_{\boldsymbol{\mu}(\boldsymbol{\beta}(\gamma))}$ is the inner product between the current residual vector $\mathbf{r}(\boldsymbol{\beta}(\gamma))$ and the i-th base of the tangent space of $\Omega_{\mathcal{B}}$ at $\boldsymbol{\mu}(\boldsymbol{\beta}(\gamma))$. Using expression (2) we have the following differential geometric identity

$$\begin{aligned} \partial_{\beta_i} \ell(\boldsymbol{\beta}(\gamma)) &= \cos\left(\rho_i(\boldsymbol{\beta}(\gamma))\right) \cdot \|\mathbf{r}\left(\boldsymbol{\beta}(\gamma)\right)\|_{\boldsymbol{\mu}(\beta(\gamma))} \cdot \|\partial_{\beta_i} \boldsymbol{\mu}(\beta(\gamma))\|_{\boldsymbol{\mu}(\beta(\gamma))} \\ &= \cos\left(\rho_i(\boldsymbol{\beta}(\gamma))\right) \cdot \|\mathbf{r}\left(\boldsymbol{\beta}(\gamma)\right)\|_{\boldsymbol{\mu}(\boldsymbol{\beta}(\gamma))} \cdot (i(\beta_i(\gamma)))^{1/2}, \end{aligned} \tag{3}$$

where $\rho_i(\boldsymbol{\beta}(\gamma))$ is the local angle between $\mathbf{r}(\boldsymbol{\beta}(\gamma))$ and $\partial_{\beta_i} \boldsymbol{\mu}(\boldsymbol{\beta}(\gamma))$, $i(\beta_i(\gamma))$ is the expected Fisher information for $\beta_i(\gamma)$ and $\|\cdot\|_{\boldsymbol{\mu}(\boldsymbol{\beta}(\gamma))}$ is the norm defined on the tangent space of $\Omega_{\mathcal{B}}$ at $\boldsymbol{\mu}(\boldsymbol{\beta}(\gamma))$. Condition (3) shows that the gradient of the log-likelihood function does not generalize the notion of equiangular condition, since we are not considering the variation related to $(i(\beta_i(\gamma)))^{1/2}$. To overcome this problem, [AUG 09] propose a generalization of LARS algorithm based on the following condition

$$|r_u\left(\beta_i(\gamma)\right)| = |i^{-1/2}(\beta_i(\gamma)) \cdot \partial_{\beta_i} \ell(\boldsymbol{\beta}(\gamma))| = \cos\left(\rho_i(\boldsymbol{\beta}(\gamma))\right) \cdot \|\mathbf{r}(\boldsymbol{\beta}(\gamma))\|_{\boldsymbol{\mu}(\boldsymbol{\beta}(\gamma))}, \, \forall i \in \mathcal{A}, \tag{4}$$

where $r_u\left(\beta_i(\gamma)\right)$ is the i-th Rao score statistic evaluated in $\boldsymbol{\beta}(\gamma)$ and $\mathcal{A}$ is the active set, namely, the set of indices of covariates that are included in the actual model. The proposed algorithm can be formalized as follows. Let $\boldsymbol{\beta}(\gamma_0) \in \mathcal{B}$ be the point in which a new covariate is included in the active set, then the following conditions are satisfied

1. $|r_u(\beta_i(\gamma_0))| = \gamma_0, \quad \forall\, i \in \mathcal{A}$,
2. $|r_u(\beta_h(\gamma_0))| < |r_u(\beta_i(\gamma_0))|, \quad \forall\, h \in \mathcal{A}^c$ and $\forall\, i \in \mathcal{A}$.

Let $\mathbf{r}_u\left(\boldsymbol{\beta}_{\mathcal{A}}(\gamma)\right) = \left(r_u(\beta_1(\gamma_0)), \ldots, r_u(\beta_{k_{\mathcal{A}}}(\gamma_0))\right)^T$ be the vector of Rao score statistics evaluated in $\boldsymbol{\beta}_{\mathcal{A}}(\gamma)$, where $k_{\mathcal{A}} = |\mathcal{A}|$. The curve $\boldsymbol{\beta}_{\mathcal{A}}(\gamma)$ that satisfies the equiangular condition (4) can be computed as solution of the following system of $k_{\mathcal{A}}$ non-linear equations

$$\mathbf{r}_u\left(\boldsymbol{\beta}_{\mathcal{A}}(\gamma)\right) = \tilde{\mathbf{v}}(\gamma_0 - \gamma), \quad \gamma \in \mathbb{R}^+,$$

where $|\tilde{v}_i| = 1$ for any $i \in \mathcal{A}$. We follow the path until a new covariate is included in the active set, formally we search $\gamma^* \in \mathbb{R}^+$ such that

$$\exists\, h \in \mathcal{A}^c \;:\; |r_u(\beta_h(\gamma^*))| = |r_u(\beta_i(\gamma^*))|, \quad \forall i \in \mathcal{A}, \tag{5}$$

then we set $\mathcal{A} = \mathcal{A} \cup \{h\}$. To compute the solution path $\boldsymbol{\beta}_{\mathcal{A}}(\gamma)$ we use the predictor-corrector algorithm [ALL 90]. The basic idea underlying the predictor-corrector algorithm is to trace a curve, implicitly defined by a system of non-linear equations, by generating a sequence of points satisfying a chosen tolerance criterion. A predictor-corrector step was also used in [PAR 07] to compute path of the coefficients of a generalized linear model with L_1 penalty function. When we work in a ultrahigh dimensional feature space, namely $p \gg n$, the following expression

$$\frac{r_u^2\left(\beta_i(\gamma)\right)}{\|\mathbf{r}\left(\boldsymbol{\beta}(\gamma)\right)\|^2_{\boldsymbol{\mu}(\boldsymbol{\beta}(\gamma))}} = \cos^2\left(\rho_i(\boldsymbol{\beta}(\gamma))\right) \tag{6}$$

can be used to define a genuine generalization of the SIS method for generalized linear models. Then, we propose the following method to identify the relevant variables in a generalized linear model defined in a ultrahigh dimensional feature space. Following [FAN 08], let $\boldsymbol{\omega} = \left(\cos^2(\rho_1(\hat{\beta}_0)), \cos^2(\rho_2(\hat{\beta}_0)), \ldots, \cos^2(\rho_p(\hat{\beta}_0))\right)^T$, where $\hat{\beta}_0$ is the maximum likelihood estimate of the parameter of a generalized linear model with only the intercept. For a given value $d < n$, we sort the p componentwise magnitudes of the vector $\boldsymbol{\omega}$ in decreasing order and define the submodel

$$\mathcal{M}_d = \{i \in \{1, 2, \ldots, p\} : \cos^2(\rho_i(\hat{\beta}_0)) \text{ is among the first } d \text{ largest of all}\}.$$

This is a straightforward way to reduce the dimensionality from p to d. Then, submodel $\mathcal{M}_d$ is studied by using the dgLARS algorithm. We have evaluated the goodness of the proposed method by means of a simulation study.

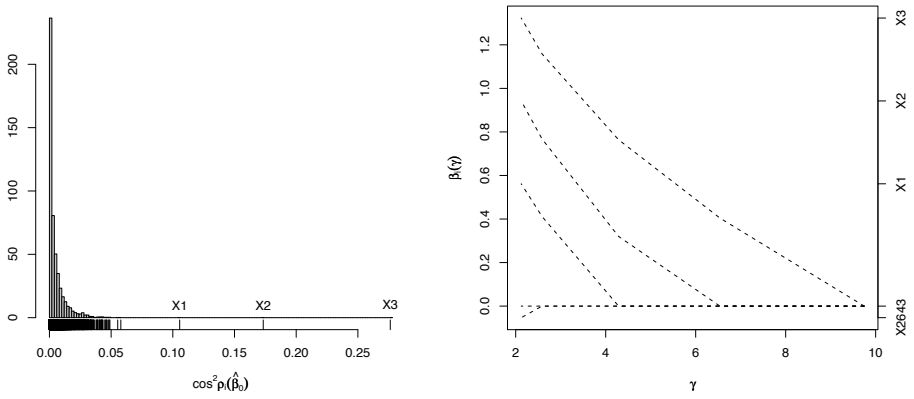
3. Simulation Study and Conclusions

We evaluated the behaviour of the proposed algorithm through a simulation study. We used a simulation setting similar to that one proposed by [FAN 08] to evaluate the SIS method. In particular, we simulated 1000 samples with 200 observations for 5000 predictors. These predictors are generated using the standard normal distribution. The hypothesized true model is the following logistic regression model

$$\text{logit}(E(Y_i)) = 1 + 2 \cdot x_{i1} + 3 \cdot x_{i2} + 4 \cdot x_{i3}.$$

We supposed that the response variable Y has a Bernoulli distribution and then we generated the Y_i values according to this setting. The dimension of the submodel $\mathcal{M}_d$ has been set to $d = 5$. Figure 2(a) shows the histogram of $\boldsymbol{\omega}$ for a sample. We can clearly see that the proposed method identifies the true variables used to simulate the logistic regression model. In figure 2(b) we have reported the path of the coefficients of the submodel $\mathcal{M}_d$ estimated with dgLARS algorithm. Similar results were obtained for the other samples, but results are not reported for the sake of brevity.

FIG. 1. *Figure (a) shows the histogram of* $\cos^2(\rho_i(\hat{\beta}_0))$ *obtained using a sample with* $n = 200$ *observations and* $p = 5000$ *variables. Figure (b) shows the path of the coefficients of the submodel* $\mathcal{M}_d$ *estimated by dgLARS algorithm.*



(a) (b)

In conclusion, in this paper we propose a new method to identify important variables for a generalized linear model defined on a ultrahigh dimensional feature space. The proposed method generalizes the approach proposed by [FAN 08] for linear regression models and it is based on the differential geometrical theory underlying the dgLARS algorithm proposed by [AUG 09]. Further investigation is necessary to develop a probabilistic theory for the choice of the optimal dimension of the submodel $\mathcal{M}_d$ evaluated by dgLARS algorithm.

4. Bibliographie

[ALL 90] ALLGOWER E., GEORG K., *Numerical Continuation Methods*, Springer, Berlin, 1990.

[AMA 85] AMARI S.-I., *Differential-Geometrical Methods in Statistics (Lecture Notes in Statistics, 28)*, Springer-Verlag, New York, 1985.

[AUG 09] AUGUGLIARO L., WIT E., Generalizing LARS algorithm using differential geometry, *to appear*, , 2009.

[CAR 92] DO CARMO M. P., *Riemannian Geometry*, Birkhäuser, Boston, 1992.

[EFR 04] EFRON B., HASTIE T., JOHNSTONE I., TIBSHIRANI R., Least angle regression, *The Annals of Statistics*, vol. 32, 2004, p. 407–451.

[FAN 01] FAN J., LI R., Variable selection via nonconcave penalized likelihood and its oracle propoerties, *Journal of the American Statistical Assotiation*, vol. 96, 2001, p. 1348–1359.

[FAN 08] FAN J., LV J., Sure independent screening for ultrahigh dimensional feature space, *Journal of the Royal Statitical Society, Series B.*, vol. 70, 2008, p. 849–911.

[JØR 87] JØRGENSEN B., Exponential dispersion models, *Journal of the Royal Statitical Society, Series B.*, vol. 49, 1987, p. 127–162.

[KAS 97] KASS R., VOS P., *Geometrical Foundation fo Asymptotic Inference*, John Wiley & Sons, Inc , New York, 1997.

[MCC 89] MCCULLAGH P., NELDER J., *Generalized Linear Models*, Chapman & Hall, London, 1989.

[PAR 07] PARK M. Y., HASTIE T., L_1-regularization path algorithm for generalized linear models, *Journal of the Royal Statitical Society, Series B.*, vol. 69, 2007, p. 659–677.

[TIB 96] TIBSHIRANI R., Regression shrinkage and selection via the lasso, *Journal of the Royal Statitical Society, Series B.*, vol. 58, 1996, p. 267–288.

Catégorisation de documents à l'aide de termes composés

J. Beney *, C.H.A. Koster **

**Université de Lyon, LCI, Département Informatique,*
INSA de Lyon F69621 Villeurbanne, jean.beney@insa-lyon.fr
***Radboud University, ICIS, Nijmegen, kees@cs.ru.nl*

RÉSUMÉ. *Nous présentons une expérience de classification de documents à l'aide de triplets construits par un analyseur syntaxique écrit en AGFL. Comparé à l'utilisation des mots simples, elle montre une nette amélioration de la qualité lorsque les triplets sont utilisés avec les mots, mais pas lorsqu'ils sont pris seuls.*

MOTS-CLÉS : *Classification supervisée, Linguistique, Analyse syntaxique, Termes complexes*

1. Introduction

Les expériences publiées d'utilisation de termes complexes pour la classification supervisée de documents ont rarement été positives, que ces termes soient extraits par analyse syntaxique profonde ou superficielle (expression régulières) ou soient simplement des paires de mots statistiquement cooccurents (pour une vue d'ensemble : [SPA 99]). Au mieux, l'utilisation de ces méthodes complexes donne la même qualité que la catégorisation par les formes de mots prises telles quelles, et parfois le résultat est moins bon.

Pourtant, les mots, mêmes techniques, étant souvent ambiguës, il semble naturel de penser qu'un terme composé est souvent moins ambigu et donc un meilleur représentant du propos d'un texte que chacun des mots qui le composent, surtout si ces deux composants sont reliés par une relation syntaxique.

Après une première expérience également décevante [KOS 02], nous avons repris les essais sur un plus large jeu de documents, avec une version améliorée de l'analyseur syntaxique de l'anglais et nous avons reconsidéré l'étalonnage des paramètres de la méthode d'apprentissage.

Nous présentons ci-dessous l'outil AGFL utilisé pour construire l'analyseur syntaxique, puis les résultats de l'expérience de catégorisation.

2. L'analyse syntaxique

2.1. *AGFL*

Issu de travaux sur l'analyse des langages pour la compilation, AGFL[1] (Affix Grammars on Finite Lattice [KOS 91]) est adapté au traitement des langues naturelles. Cet outil s'inscrit dans la lignée des travaux de Chomsky, en traitant des grammaires non contextuelles et en permettant de prendre en compte les *transformations* amenant à la grammaire de surface.

1. Voir : http ://www.agfl.cs.ru.nl/

La langue est décrite par des règles paramétrées[2] :

```
sujet negatif(NOMBRE, trois) :
        DetNeg(GENRE, NOMBRE),
        nom commun(GENRE, NOMBRE),
        adjectifs opt(GENRE, NOMBRE).
```

Les types de paramètres étant définis par des règles d'affixes :

```
GENRE : : mas | fem.
```

Le moteur accepte les ambiguïtés syntaxiques et peutdonc construire toutes les analyses possibles par la méthode descendante avec la variante du coin gauche, mais on peut se limiter à la *meilleure* analyse qui est définie à l'aide de *pénalités* appliquées à certaines règles, et de l'indication des fréquences des mots du lexique. Celui-ci est géré très efficacement (*try*) et contient les traits syntaxiques :

```
"pas de"        DetNeg(GENRE, sing|plur)
```

Il est possible de préciser le format de sortie que l'on désire obtenir à la place du résultat standard (arbre syntaxique). Par exemple, pour écrire le triplet reconnu par la la règle ci-dessus, on ajoutera à la fin de celle-ci :

```
... / "[", nom commun, ",ATTR, " adjectifs opt "]" .
```

2.2. *Les triplets syntaxiques*

L'analyse fournit un arbre de dépendances, représentant la structure de chaque phrase.

Pour l'exploiter, nous découpons cet arbre en triplets correspondant chacun à une relation entre deux mots. Les relations utilisées sont : sujet, objet, attribut, prédicat, modifieur, ainsi que toutes les prépositions. Par exemple, la phrase "Je prends la porte avec son chambranle" donnera les triplets :

[je, SUBJ, prends], [prends, OBJ, porte],[porte, AVEC, chambranle].

Comme nous ne cherchons pas à traiter le *sens* d'un texte mais uniquement à représenter son *propos*, certaines informations sont omises (quantification, temps, mode des verbes, ...). En effet, une phrase et sa négation parlent de la même chose. En conséquence, certains mots, inutiles pour notre objectif, ont été supprimés (déterminants, par exemple).

3. L'expérimentation

Le système utilisé est LCS avec la méthode Winnow symétrique [DAG 97, BEN 08]. Des expériences précédentes nous ont fait utiliser les paramètres suivants pour les mots : promotion 1,01, rétrogradation 0,99, 10 itérations sur les documents, seuil épais [0,5–2], force des termes LTC, sélection des termes par l'incertitude.

La force d'un terme dans un document est calculée par la formule LTC, qui est une variante de TF*IDF[3].

3.1. *Les données*

Les documents utilisés sont des brevets en anglais provenant de l'EPO (Office Européen des Brevets). 68 418 documents appartiennent à une ou plusieurs (1,43 en moyenne) des 44 classes (directoires de l'EPO). Il y a au minimum 2000 documents par classe. Ils ont en moyenne 6 000 mots et le vocabulaire comprend 557 790 mots différents.

L'analyse syntaxique fait apparaître 49 112 399 triplets différents, et un document en contient 9500 en moyenne.

2. Les exemples sont tirés d'une grammaire du français en cours d'écriture, tandis que l'expérimentation a utilisé l'EP4IR (English Phrases for Information Retrieval, voir le site d'AGFL).
3. term frequency, inverse document frequency [SAL 88].

3.2. *Les résultats*

Les documents ont été divisés en un ensemble d'apprentissage (80%) et un ensemble de test (20%). Nous donnons la valeur de F1 calculée sur ces deux ensembles ainsi que l'écart type calculé sur 10 divisions aléatoires (validation croisée).

Notons d'abord qu'un ré-étalonnage des paramètres de Winnow a apporté une amélioration de près de 1%. En particulier, le coefficient de promotion a été passé à 1,02 au lieu de 1,01.

F1	apprent.	gain/mots	test	gain/mots
mots	81,58(0,20)		65,86(0,23)	
triplets	91,85	10,27	65,60	-0,26
triplets réét.	96.59(0.03)	15.07(0.25)	66.15(0.28)	0.26(0.26)
triplets+mots	90,16	8,58	69,12	3,26
triplets+mots réét.	96,09(0,04)	14,51(0,19)	70,03(0,21)	4,17(0,16)

On remarquera surtout que la différence entre les triplets seuls et les mots n'est pas toujours positive, mais que l'ajout des triplets aux mots permet de gagner plus de 4%, ce qui est statistiquement significatif.

Cet exemple de termes (ordonnés par importance décroissante, pour la catégorie dir24), illustre la manière dont les triplets et les mots se partagent le rôle discriminant :

terme	poids
A :redundant	1900,6
A :V :testing	45,6
[N :instrument,ATTR, A :electronic]	32,9
N :panel	23,1
N :bits	16,5
[N :mode,ATTR N :test]	12,7
[N :system,ATTR N :memory]	12,6
V :accessed	10,9
[N :crystal,ATTR A :liquid]	10,2
[N :system,ATTR N :computer]	9,5
[N :converter,SUBJ V :comprising]	5,9

Il faut noter que les très grands poids des premiers termes sont compensés par leurs forces qui dépendent de l'inverse de leur fréquence.

4. Conclusion et travaux en cours

Les triplets, tels que nous les extrayons des documents, améliorent la classification si l'on utilise des paramètres adaptés : un coefficient de promotion plus grand donc un apprentissage plus rapide. Ceci suggère que cette représentation des document est moins sujette à un sur-apprentissage, ce qui est confirmé par le fait que l'optimum sur l'ensemble de test est obtenu avec un bien meilleur classifieur de l'ensemble d'apprentissage (96,09% au lieu de 81,58%). Les triplets sont donc des termes qui discriminent bien mieux les classes.

Mais ils sont encore insuffisants et leur alliance avec les mots simples donne encore un meilleur résultat.

De plus, ils sont très nombreux, ce qui allonge les temps d'apprentissage (40h) et de test, même si ceux-ci restent raisonnables pour l'exploitation (un apprentissage tous les mois, ou moins ; 0,38s pour classifier un document de 6 000 mots).

Remarquons que notre jeu d'essai est plus large que ceux utilisés précédemment. Ceci peut expliquer notre résultat positif : les triplets étant plus rares, il faut plus de documents pour repérer ceux qui sont utiles à la classification.

Nos efforts portent actuellement sur la réduction, avant l'apprentissage, du nombre de ces termes par diverses techniques de regroupement de ces termes : lemmatisation, suppression du type des mots ou des relations, ...

5. Bibliographie

[BEN 08] BENEY J., *Classification supervisée de documents*, Hermes-lavoisier, 2008.

[DAG 97] DAGAN I., KAROV Y., ROTH D., Mistake-Driven Learning in Text Categorization, *Proceedings of the Second Conference on Empirical Methods in NLP*, 1997, p. 55–63.

[KOS 91] KOSTER C. H. A., Affix grammars for natural languages, *In Attribute Grammars, Applications and Systems, International Summer School SAGA*, Springer-Verlag, 1991, p. 469–484.

[KOS 02] KOSTER C., SEUTTER M., Taming Wild Phrases, *Proceedings 25th European Conference on IR Research (ECIR'03)*, LNCS 2633, Springer, 2002, p. 161-176.

[SAL 88] SALTON G., BUCKLEY C., Term-weighting approaches in automatic text retrieval, *Information Processing and Management*, 1988, p. 513–523.

[SPA 99] SPARCK-JONES K., What is the role of NLP in text retrieval ?, T. S., Ed., *Natural Language Information Retrieval*, 1999, p. 1–25.

Essais de classification par l'intermédiarité

Marc Le Pouliquen [1], **Marc Csernel** [2] **et Stéphane Sire** [3]

1- Telecom Bretagne, Labsticc UMR 3192 , BP 832, 29285 Brest Cedex - France
marc.lepouliquen@telecom-bretagne.eu
2- Inria-Rocqencourt, BP-105- 78180 Le Chesnay - France
marc.csernel@inria.fr
3- Université de Bretagne Occidentale, LBMS EA 4325, 6 av. Le Gorgeu - CS93837 29238 Brest Cedex 3 - France
stephane.sire@univ-brest.fr

RÉSUMÉ. *Nous allons, dans cet article, regarder comment à partir d'informations sur l'intermédiarité ou la presque-intermédiarité de documents, nous pouvons obtenir des classifications ou des sériations. Dans un premier temps, nous modélisons une relation ternaire d'intermédiarité entre des documents ; pour cela, nous utilisons plusieurs caractérisations différentes, celle issue de la géométrie, celle liée à une structure définie par une relation binaire ou par un score, celle de la théorie des ensembles... A partir des relations obtenues, nous observons les structures classificatoires que l'on peut visualiser, ainsi que les différentes contraintes que l'on doit imposer aux relations pour satisfaire les structures choisies. Après cette partie plus théorique, nous utilisons notre modélisation pour trois applications concrètes. La première se situe dans le cadre de la classification automatique de courriels ; l'intermédiarité peut permettre de ranger chaque nouveau courriel dans une classification pré-établie. C'est dans l'édition critique que l'on retrouve la seconde application, l'intermédiarité permettant d'affiner la filiation des différentes versions d'un même manuscrit. Enfin, la troisième application permet, en archéologie, de réaliser la sériation de vestiges. Nous observons ainsi que les critères permettant d'obtenir les relations d'intermédiarité doivent souvent être relaxés pour permettre d'établir la classification désirée.*

MOTS-CLÉS : *Intermédiarité, classification de texte, sériation archéologique.*

1. Introduction

La notion d'intermédiarité est une notion assez naturelle dans de nombreux domaines :
– un lieu B est intermédiaire entre A et C s'il se trouve sur un chemin menant de A à C ;
– 1914 est intermédiaire entre 1870 et 1939 ;
– le vert est entre le jaune et le bleu ;
– Paul est entre Jean et Louis si Paul est le fils de Louis et le père de Jean ;
– Un ensemble B est entre les ensembles A et C si, par exemple, $A \subset B \subset C$ ou $C \subset B \subset A$.

Nous allons tenter, dans cet article, de modéliser cette relation par plusieurs caractérisations différentes. La première définition de l'intermédiarité, que nous utilisons, est une de celle issue de la géométrie qui est liée à une métrique. Elle est ici appliquée à la construction du stemma codicum en édition critique. La seconde caractérisation définit l'intermédiarité par un score liée à la structure des données. L'application est réalisée dans le cadre de la sériation d'une base de données de ponts métalliques. C'est de la théorie des ensembles que provient la dernière définition de l'intermédiarité, et c'est pour la classification automatique de courriels que nous l'utilisons.

2. Intermédiarité en géométrie, construction du *stemma codicum*

Pour la première modélisation de l'intermédiarité dans le cadre de la classification, nous nous sommes intéressés à la construction du *stemma codicum* [1] qui est une des phases de la réalisation d'une édition critique de manuscrits anciens. Les techniques employées sont dédiées à un corpus de manuscrits sanskrits (cf. [LEP 06]).

C'est à Don Quentin [QUE 26] que l'on doit l'idée d'utiliser la notion d'intermédiarité afin de dresser le stemma. En effet, il se propose de reconstituer des petites chaînes de trois manuscrits dont l'un est l'intermédiaire des deux autres puis, d'assembler ces petites chaînes afin d'inférer l'arbre complet. En s'inspirant de cette démarche, nous allons utiliser une des nombreuses caractérisations géométriques de l'intermédiarité, celle introduite par Menger [MEN 28] sous le nom de relation métrique d'intermédiarité de la façon suivante :

Définition 2.0.1 *Une relation ternaire B sur une ensemble E est dite relation métrique d'intermédiarité s'il existe une métrique d sur E telle que :* $(a, b, c) \in B \Leftrightarrow d(a, b) + d(b, c) = d(a, c)$

A partir de maintenant, pour exprimer le fait que b est intermédiaire entre a et c, on se contentera d'écrire (a, b, c) au lieu de $(a, b, c) \in B$.

Il faut désormais définir une métrique qui permet de comparer les manuscrits de telle sorte que l'intermédiarité au sens de la copie corresponde à celle de la métrique. Pour cela, prenons un exemple. Soient les 3 phrases suivantes correspondant aux trois mêmes phrases de différents manuscrits copiés les uns sur les autres.

A = `Voici une phrase courte inventée pour l'exemple`

B = `Voici une phrase inventée pour cet exemple`

C = `Voici une phrase créée pour cet exemple`

La logique veut que la phrase B soit intermédiaire au sens de la copie entre A et C. En effet, le copiste de B a omis `courte` et a modifié `l'` en `cet` et le copiste de C a remplacé `inventée` par `créée`. C'est en revanche peu probable que C soit l'intermédiaire, car le copiste de C a supprimé `inventée` qui est réintroduit par le copiste suivant. Soit la distance d définie par le nombre de mots insérés, supprimés ou substitués entre 2 textes alignés (une sorte de distance d'édition (Levenstein [LEV 66]) au niveau des mots). Dans notre cas, les calculs de d donnent :

A	Voici une phrase	courte	inventée	pour	l'	exemple
B	Voici une phrase		inventée	pour	cet	exemple
C	Voici une phrase		créée	pour	cet	exemple
$d(A, B) = 2$		$d(B, C) = 1$		$d(A, C) = 3$		

TABLE 1 – Alignement des trois phrases pour compter le nombre de variantes

On constate en effet que (A, B, C) puisque $d(A, C) = d(A, B) + d(B, C)$ et l'on n'a pas (A, C, B).

3. Intermédiarité définie par un score, application à une base de données de ponts métalliques

Nous cherchons à réaliser la sériation d'un ensemble de ponts métalliques dont une base de données a été réalisée par Stéphane Sire et Dominique Malfondet (cf. [SIR 07]). Pour cela, nous utilisons une méthodologie originale qui utilise l'intermédiarité.

1. Le **stemma codicum** consiste à trier les différentes versions d'un même texte afin d'établir un arbre (ou un graphe) de la filiation des manuscrits du corpus pour savoir lequel a été copié sur l'autre et détecter les chaînons manquants. Cette généalogie des manuscrits représente donc la filiation établie entre différentes versions d'un même texte.

Pour la modélisation de l'intermédiarité, nous définissons un score permettant d'établir, pour un pont choisi, s'il est intermédiaire entre deux autres ponts. Nous construisons ce score en favorisant les champs de la base de données caractéristiques d'une période temporelle. Nous pensons qu'à partir de l'ensemble des relations « temporelles » d'intermédiarité, nous allons retrouver une sériation elle-même « temporelle » des ponts métalliques. Il convient en premier lieu d'analyser la base de données pour repérer les champs caractéristiques d'une période de temps, on les appellera les champs temporels.

Trois hypothèses préfigurent alors la construction d'un score d'intermédiarité pour les champs temporels :

– Pour les champs de type numérique ou de type date

  (i) Si un pont p_2 est intermédiaire dans le temps entre les ponts p_1 et p_3 alors la valeur d'un champ temporel de p_2 se situe dans l'intervalle entre ceux de p_1 et p_3.

– Pour les champs de type texte

  (ii) Si un pont p_2 est intermédiaire dans le temps entre les ponts p_1 et p_3, alors si la valeur d'un champ temporel de p_1 est égale à la valeur du même champ temporel de p_3, cela implique que la valeur de celui de p_2 est aussi la même.

  (iii) Si un pont p_2 est intermédiaire proche dans le temps entre les ponts p_1 et p_3 alors la valeur d'un champ temporel de p_2 se retrouve soit dans p_1 soit dans p_3.

Autrement dit, pour que p_2 soit intermédiaire entre p_1 et p_3, il suffit que le champ de type texte de p_1 et p_3 s'accorde tour à tour avec celui de p_2 mais surtout que ces deux champs ne s'accordent jamais contre lui.

Prenons un exemple simple avec quatre ponts métalliques en arc.

Pour chaque pont, nous avons sélectionné quatre champs dans la base de données. Le premier champs `date fin travaux` est le champ de référence, il permet juste de vérifier les résultats. Les deux champs suivants, `portée de l'arc` et `matériau` peuvent être considérés comme temporels car la portée de l'arc et les matériaux évoluent avec les technologies donc avec le temps. Le dernier champ `nombre arches` dépendant essentiellement de la largeur du cours d'eau est peu représentatif de la chronologie.

	date fin travaux	portée de l'arc (en m.)	matériau	nbre arches
pont de Coalbrookdale (Royaume Uni)	1779	30,5	fonte	1
pont d'Austerlitz (Paris)	1806	32,36	fonte	5
pont du Carroussel (Paris)	1834	47,67	fonte	3
pont d'Arcole (Paris)	1856	80	fer	1

Le pont d'Austerlitz est intermédiaire proche entre le pont de Coalbrookdale et celui du Carroussel puisque, pour le champ `portée de l'arc`, on a bien $32,36 \in [30, 547, 67]$ et que pour l'autre champ temporel `matériau`, on respecte le (ii) et le (iii) car les trois ponts sont en fonte.

Le pont d'Arcole n'est pas intermédiaire entre le pont d'Austerlitz et celui du Carroussel. Le champ numérique `portée de l'arc` est tel que $80 \in [32, 3647, 67]$ et ne vérifie pas le (i). Le champ de type texte `matériau` ne vérifie pas le (ii) car les deux extrémités, le pont d'Austerlitz et celui du Carroussel sont tous les deux en fonte alors que celui d'Arcole est en fer. Aucun des deux champs temporels ne vérifie l'intermédiarité.

On peut maintenant définir pour chaque triplet p_1, p_2, p_3 un indice d'intermédiarité I_i constitué de deux sous indices $I_{i_{num}}$ et $I_{i_{text}}$.

– $I_{i_{num}}$ est le quotient du nombre de champs temporel de type numérique ou de type date qui ne vérifie pas le (i) par le nombre total de ces champs.
– $I_{i_{text}}$ est le quotient du nombre de champs temporel de type texte qui ne vérifie pas le (ii) par le nombre total de ces champs.

L'indice d'intermédiarité I_i est la somme pondérée par le pourcentage de champs temporels considérés par les deux sous indices $I_{i_{num}}$ et $I_{i_{text}}$. Si l'indice d'intermédiarité est nul alors p_2 est intermédiaire entre p_1 et p_3 et

s'il n'est pas nul, on détermine si le pont est « plus ou moins intermédiaire ». Ainsi, en reprenant l'exemple, on obtient

$I_{i_{Coalbrookdale,Austerlitz,Carroussel}} = 0,5 * \frac{0}{1} + 0,5 * \frac{0}{1} = 0$ et $I_{i_{Austerlitz,Arcole,Carroussel}} = 0,5 * \frac{1}{1} + 0,5 * \frac{1}{1} = 1$

Nous obtenons ainsi un ensemble de relations d'intermédiarité qui vont nous permettre de réaliser la sériation.

4. Intermédiarité en théorie des ensembles, mise en pratique pour la classification de courriels

La modélisation de l'intermédiarité, définie ici, a pour objectif de proposer une classification par apprentissage supervisé de courriels. C'est-à-dire que, pour apprendre comment classifier les courriels, nous disposons préalablement d'un premier ensemble de courriels dont nous connaissons déjà les classes réalisées par l'utilisateur. Ainsi, à chaque nouveau courriel reçu, le système doit être capable de proposer rapidement à l'utilisateur un répertoire qui correspond à la classe dans lequel il doit se situer.

Pour les premiers tests, et afin d'obtenir une méthode suffisamment rapide, nous avons commencé par travailler uniquement sur le titre du courriel. La méthode s'appuie sur une définition de l'intermédiarité introduite par Restle [RES 59]. Cette relation définit la notion d'intermédiarité au niveau des ensembles.

Définition 4.0.2 *Soient trois ensembles A,B et C. On considère que B est intermédiaire entre A et C ssi :*

(i) $A \cap \overline{B} \cap C = \emptyset$

(ii) $\overline{A} \cap B \cap \overline{C} = \emptyset$

Dans le cas de nos courriels, nous pouvons très bien les associer à un ensemble, l'ensemble constitué par les mots du titre. Nous pouvons alors obtenir un ensemble de relations d'intermédiarité entre le nouveau courriel et ceux déjà classés qui va nous permettre de le classifier.

5. Conclusion

Dans un deuxième temps, nous nous intéresserons aux possibilités que l'on a de construire une structure classificatoire (graphe, arbre, sériation,...) à partir d'un certain nombre de relations d'intermédiarité entre objets. On utilise alors différentes contraintes que l'on doit imposer aux relations pour satisfaire les structures choisies. Nous regarderons ensuite comment nous pouvons reconstruire la structure choisie à partir d'un ensemble de relations d'intermédiarité obtenues par la modélisation précédente.
La troisième partie est applicative. Dans le cadre des trois mises en pratique, nous regarderons ce que peut apporter l'utilisation d'un critère d'intermédiarité sur les techniques de classification.

6. Bibliographie

[LEP 06] LEPOULIQUEN M., BARTHÉLEMY J., BERTRAND P., Filiation de manuscrits sanskrits et arbres phylogénétiques, *RNTI-C-2 Classification : points de vue croisés, 2008, Actes de la XIIIe rencontre de la Société Francophone de Classification, http://lita.sciences.univ-metz.fr/~sfc06/*, 2006.

[LEV 66] LEVENSHTEIN V. I., Binary Codes capable of correcting deletions, insertions, and reversals, *Soviet Physics - Doklady*, vol. 10, n^o 8, 1966, p. 707-710.

[MEN 28] MENGER K., Untersuchungen über allgemeine Metrick, *Mathematische Annalen*, vol. 100, 1928, p. 75-163.

[QUE 26] QUENTIN H., *Essais de critique textuelle*, Picard, 1926.

[RES 59] RESTLE F., A metric and an ordering on sets, *Psychometrika*, vol. 24, 1959, p. 207-220.

[SIR 07] SIRE S., MALFONDET D., LIRON J.-M., Analyse Comparative de l'Émergence du Patrimoine Industriel (ACEPI) : aide informatique à la réflexion, la contribution et la communication, *5ème journée O.S.TIC, Institut de l'Homme et de la Technologie, École polytechnique de l'Université de Nantes*, 2007.

Index

La SFC organise, chaque année, les « Rencontres de la Société Francophone de Classification », qui ont pour objectifs de présenter des résultats récents, des applications originales en classification ou dans des domaines connexes, ainsi que de favoriser les échanges scientifiques à l'intérieur de la société et de faire connaître à divers partenaires extérieurs les travaux de ses membres. Durant ces rencontres qui rassemblent régulièrement une centaine de participants, est attribué le prix Simon Régnier, consacrant une contribution originale d'un jeune chercheur à la classification.

En 2009, les universités grenobloises, Joseph Fourier, Pierre Mendès France et Grenoble-INP se mobilisent pour ces seizièmes Rencontres de la Société Francophone de Classification qui se dérouleront à l'Université Joseph Fourier du 2 au 4 septembre 2009.