Comptes Rendus des 13ème Rencontres
de la Société Francophone de Classification

METZ, 6–8 septembre 2006

Université Paul Verlaine – Metz
Île du Saulcy
57045 METZ Cedex 1

**Actes publiés sous la direction de**
**Mohamed NADIF et François-Xavier JOLLOIS**

# Comité Scientifique

- **Présidente** : Le Thi Hoai An (Université de Metz - LITA)
- *Vice-président* : Mohamed Nadif (IUT de Metz - LITA)
- Gilles Bernot (Université d'Evry)
- Hans-Hermann Bock (Institut de la statistique, Aix la Chapelle, Allemagne)
- Marie-José Caraty (Université Paris 5)
- Guy Cucumel (Université du Quebec, Canada)
- Edwin Diday (Université Paris Dauphine)
- Yadolah Dodge (Université de Neuchâtel, Suisse)
- Bernard Fichet (Faculté de Médecine de Marseille)
- Gérard Govaert (Université de Technologie de Compiègne)
- André Hardy (Facultés universitaires Notre-Dame de la Paix, Namur, Belgique)
- Georges Hébrail (ENST de Paris)
- Pascale Kuntz (Université de Nantes)
- Yves Lechevallier (INRIA-Rocquencourt)
- Israël César Lerman (IRISA)
- Jean-Marie Monnez (Université de Nancy)
- Amedeo Napoli (LORIA Nancy)
- Fernando Nicolau (Université Nouvelle de Lisbonne, Portugal)
- Jean-Paul Rasson (Facultés universitaires Notre-Dame de la Paix, Namur, Belgique)
- Pham Dinh Tao (INSA-Rouen)
- Gilles Venturini (Université de Tours)
- Maurizio Vichi (Université "La Sapienza" de Rome, Italie)
- Djamel Zighed (Université de Lyon 2)

# Comité d'Organisation

- **Président** : Mohamed Nadif (IUT de Metz - LITA)
- *Vice-président* : François-Xavier Jollois (Université de Paris 5 - CRIP5)
- Damien Aignel (LITA)
- Brieuc Conan-Guez (IUT de Metz - LITA)
- Pierre Laroche (IUT de Metz - LITA)
- Yves Lemoine (IUT de Metz)
- Le Thi Hoai An (Université de Metz - LITA)
- Franck Marchetti (IUT de Metz - LITA)

# Préface

Après Montréal en 2005, la ville de Metz est ravie de vous accueillir pour les XIIIèmes Rencontres de la Société Francophone de Classification qui se tiennent sur le campus de l'Université Paul Verlaine, sur l'Ile du Saulcy du 6 au 8 septembre 2006.

Les actes de cette année témoignent encore une fois de la diversité et de l'intérêt de la classification. Plusieurs domaines d'application sont présents dans les divers papiers : bioinformatique, biopuces, données textuelles, image, etc.

Cette manifestation est organisée par le Laboratoire d'Informatique Théorique et Appliquée de Metz (LITA). Sans le soutien de notre université, de notre IUT et de nos partenaires publics et privés, cette manifestation n'aurait pas pu avoir lieu. Qu'ils soient ainsi remerciés ici, de même que les membres du comité d'organisation et du comité de programme qui nous ont aidés. Nous souhaitons porter une attention toute particulière à Damien Aignel, secrétaire du LITA, pour sa disponibilité et sa contribution pour l'organisation du colloque.

Enfin nous remercions tous les invités qui ont gentiment accepté d'intervenir et tous les auteurs des articles qui composent cet ouvrage qui, par leur travail, font prospérer la discipline.

Metz le 04 juillet 2006,

Les éditeurs

# Table des matières

**Conférences Plénières**

# Complex Statistical Models for Object Recognition and Mass Spectrometry

**Joachim M. Buhmann**

*Institute for Computational Science*
*ETH Zurich, Universitätstrasse 6*
*8092 Zurich*
*jbuhmann@inf.ethz.ch*

RÉSUMÉ. *Modern Artificial Intelligence uses graphical models with a complex dependency structure to generate data representations for applications like e.g. computer vision, robotics or computational biology. Model selection and parameter optimization in high-dimensional parameter spaces arise as core problems in the inference process where new solutions have to developed to reconcile the tradeoff between statistical precision and computational complexity. In this contribution, the challenges of these learning problems are demonstrated in the context of image categorization and of unsupervised learning in proteomics, i.e., de novo peptide sequencing from mass spectrometry data. Both application domains require highly structured statistical models and efficient algorithms to process gigabyte of data for structure detection as demonstrated in [OMM 06, FIS 05, FIS 06].*

MOTS-CLÉS : *Machine Learning, graphical models, HMM*

## 1. Introduction

Machine learning today faces challenging problems in data analysis areas where massive amounts of *unlabeled* data are collected in large scale screening experiments or by online monitoring. Computer vision as well as computational biology and bioinformatics share this common property as application domains that large amounts of unlabeled data are readily available and that intelligent processing of these data sets require sophisticated statistical models with an adapted model complexity. Issues like model and parameter selection dependent on the reliability of the data and the robustness of the models arise at a level of complexity which has rarely been encountered in the past. This contribution demonstrates how complex and heterogeneously structured statistical models can be learned from data with as little supervision as possible. The frist model is used to categorize image data and the second model extracts peptide sequence data from mass spectra.

## 2. Composition Systems for Computer Vision

The richness in appearence of objects in images requires a highly flexible data structure which uses parts and combination of parts to build internal models of image content. A conceptionally novel approach has been suggested and championed by S. Geman with his *composition systems* [GEM 98] where a complex scene is decomposed in objects, parts and parts of parts. This proposal for a data representation of images which has been developed for image parsing, e.g., for handwritten character recognition and licence plate reading.

The model can be best explained by considering recognition, see Figure 2(a). Given a novel image, salient image regions are detected in a first stage using a scale invariant Harris interest point detector. Each region is then described by localized histograms. In a next step a perceptual grouping of these local part descriptors is performed to obtain a set of possible candidate compositions. This grouping leads to a sparse image representation based on

]

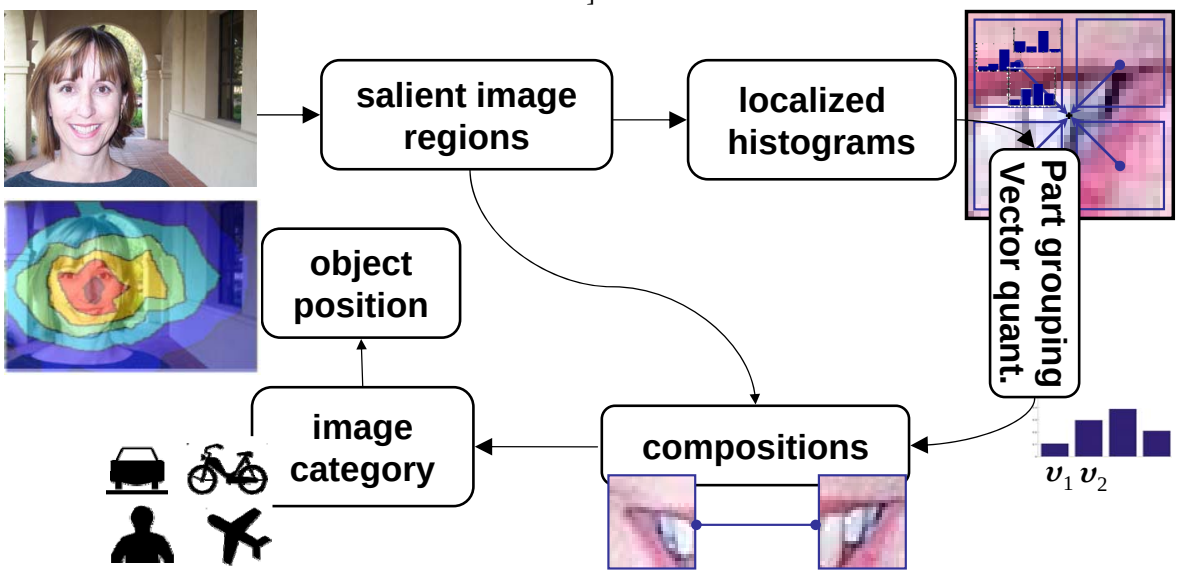



**FIG. 1.** *Processing pipeline for the image categorization system*

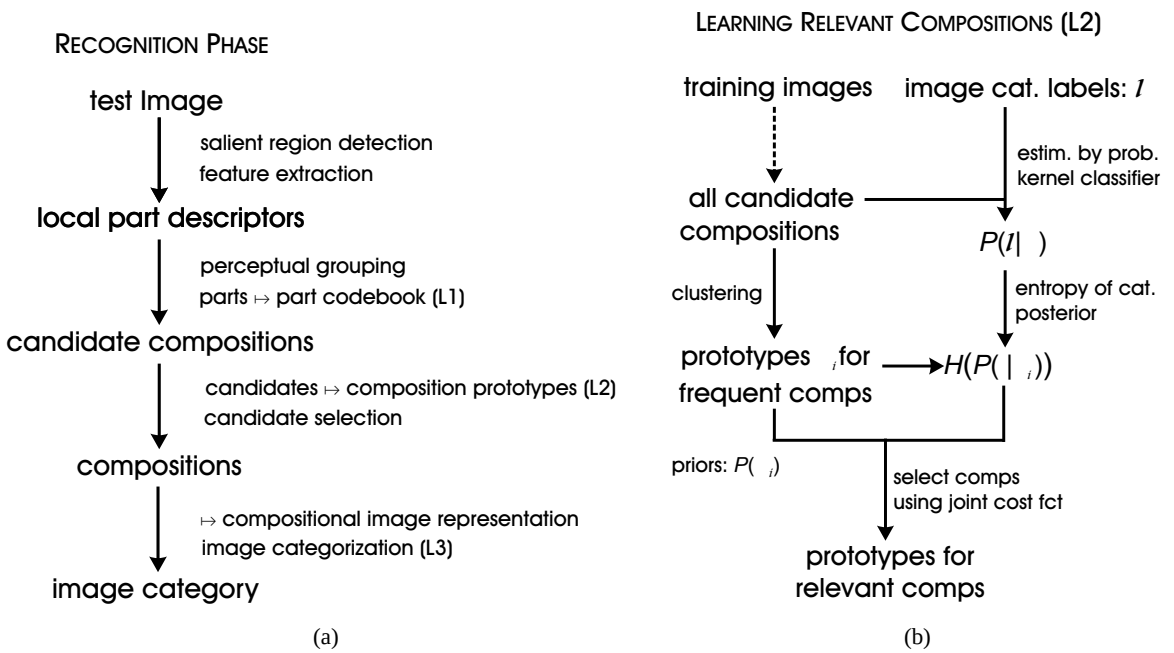



**FIG. 2.** *(a) Recognition based on compositions. (b) Learning relevant compositions (learning stage L2 from (a)), see text for details.*

(probably overlapping) subregions, where each candidate represents an agglomeration of local parts. Consecutively, composition candidates have to be encoded. Therefore all detected local part descriptors are represented as probability distributions over a codebook which is obtained using histogram quantization in the learning stage.

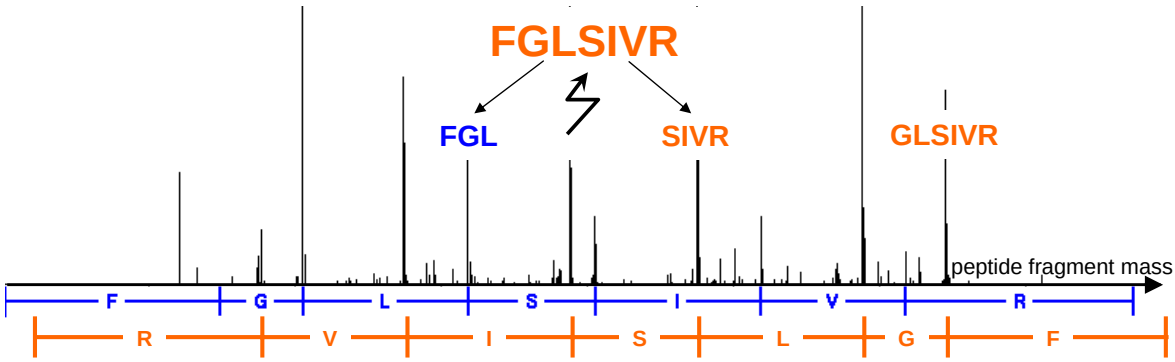



**FIG. 3.** *Mass spectrum of the peptide "FGLSIVR" and the peaks of the a possible peptide split into prefix sequence "FGL" and suffix sequence "SIVR".*

This codebook models locally typical configurations of the categories under consideration. A composition is then represented as a mixture distribution of all its part distributions, i.e. a *bag of parts*. The full processing pipeline is depicted in Figure 2.

In a next stage relevant compositions have to be selected, discarding irrelevant candidates that represent background clutter. The set of relevant compositions has to be computed in the learning phase from the training data in a weakly supervised manner (see Figure 2(b)). As intermediate compositional representations should have limited description length, this learning obeys the following rationale : (i) Firstly, we aim at a set of compositions that occur frequently in the visual world of the categories under consideration. For that purpose all composition candidates found in all the training images are clustered and the prior assignment probabilities of candidates to these prototypes is estimated. (ii) Secondly, relevant compositions have to support the discrimination of sets of categories from another. Clutter that is present in many different categories or configurations has to be discarded to reduce the model complexity. In order to find a relevance measure for discriminating categories, the category posteriors of compositions are learned from the training data. An image is then represented as a *bag of compositions*, i.e. a histogram over the detected groupings. Details can be found in [OMM 06].

## 3. Mass Spectrometry with Chains of Hidden Markov Models

In the process of high-throughput protein identification, mass spectrometry has attained considerable importance during the last decade. Analysis based on mass spectrometry typically starts with a complex protein mixture which is fractionated either by gel electrophoresis or other fractionation methods in order to reduce the complexity of the sample. The proteins are then digested by a specific enzyme like trypsin. The resulting set of peptides is measured by a tandem mass spectrometer coupled with a high performance liquid chromatography device. This preprocessing enables the molecular biologist to single out one particular peptide with a parent mas $M$, e.g., "FGLSIVR" in Figure 3. In a second stage, this peptide (or a mixture of few peptide with almost identical mass) is fragmented by low energy collisions with a noble gas. This fragmentation process yields the MS/MS spectra that ideally contain the masses of all $N$-terminal and $C$-terminal fragments of the unknown peptide.

In the following we will discuss `NovoHMM`, a hidden Markov model that generates mass spectra as a finite automaton over states that correspond to masses (measured in Dalton). The simplest model of an amino acid sequence is a list of random variables with 20 different states, each representing one amino acid in the sequence. The transitions are the conditional probabilities of observing a certain amino acid at position $t$, given the observation at position $t-1$.

A random variable in a one-dimensional Markov model only depends on the preceding variable in the sequence. Since the parent mass constraint has to be fulfilled, we have to know the total mass of all preceding amino acids.

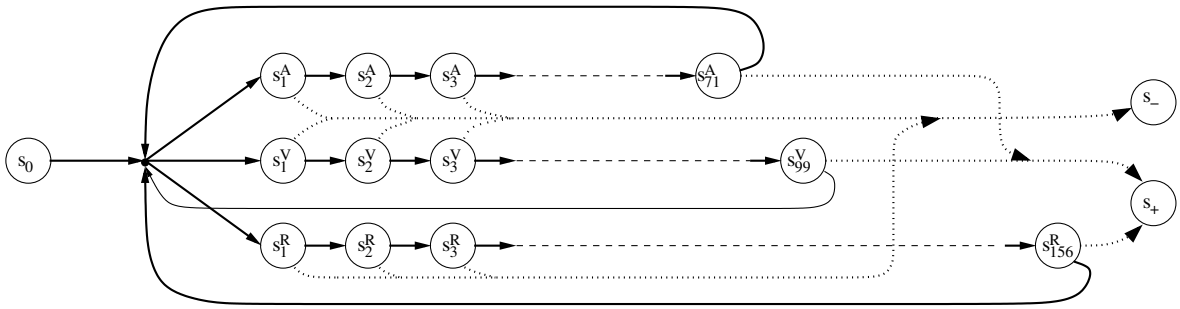



**FIG. 4.** *State transitions in the finite state automaton for tandem mass spectra. Solid arrows denote the possible transitions while generating a peptide. After $M$ steps the automaton is forced to take the positive or negative end state (dotted arrows).*

This information is coded in the model depicted in fig. 4 by the chains with as many states as the respective amino acid weighs in Dalton. The transition graph of the augmented model is depicted by the solid lines in figure 4. A missing edge in this graph denotes vanishing transition probability. Once the first state of an amino acid has been selected, the transitions are constrained to all subsequent states of this amino acid. At the end state of an amino acid, a transition occurs according to the corresponding transition probabilities in the amino-acid-based model.

The resulting model can represent amino acid sequences of arbitrary length with a state sequence of 1 Da step size. In tandem mass spectrometry, however, we actually have additional information in form of the measured parent mass $M$ of the two peptide fragements. This mass constraint enforce the HMM to end in an end state of an amino acid chain (denoted by $s^+$) when the end of the mass scale has been reached. Details of this HMM structure and how the mass peaks in the mass spectrum are generated, can be found in [FIS 05].

## 4. Conclusion

The image categorization problem and the mass spectrometry problem are characterized by a large amount of noisy or incomplete data. The probabilistic modelling approach allows us to adapt to the noise perturbations in the data and it provides a quality score for the learned model. Deterministic models behave in a much more brittle way and produce inferior results in recall experiments on benchmark data sets. The compositional categorization system and `NovoHMM` perform competitively or even outperform alternative approaches on standard benchmark tasks.

**Acknowledgement** : This is joint research with B. Fischer, B. Ommer and V. Roth and project partners.

## 5. Bibliographie

[FIS 05] FISCHER B., ROTH V., ROOS F., GROSSMANN J., BAGINSKY S., WIDMAYER P., GRUISSEM W., BUHMANN J. M., NovoHMM : A Hidden Markov Model for de Novo Peptide Sequencing, *Analytical Chemistry*, vol. 77, n$^o$ 22, 2005, p. 7265 – 7273.

[FIS 06] FISCHER B., GROSSMANN J., BAGINSKY S., ROTH V., GRUISSEM W., BUHMANN J. M., Semi-Supervised LC/MS Alignment for Differential Proteomics, *Bioinformatics*, vol. XXX, 2006, (in press).

[GEM 98] GEMAN S., POTTER D. F., CHI Z., *Composition Systems*, rapport, 1998, Division of Applied Mathematics, Brown University, Providence, RI.

[OMM 06] OMMER B., BUHMANN J. M., Learning Compositional Categorization Models, *ECCV 2006 Proceedings*, vol. III de *Lecture Notes In Computer Science 3753*, Springer Verlag, 2006, p. 316-329.

# Robust clustering of categorical data with the Forward Search

**Andrea Cerioli**[*], **Marco Riani**[*], **Anthony C. Atkinson**[**]

[*] *Dipartimento di Economia – Università di Parma*
*Via Kennedy 6, 43100 Parma, Italy*
*{andrea.cerioli,mriani}@unipr.it*
[**] *Department of Statistics – London School of Economics*
*Houghton Street, London WC2A 2AE, UK*
*a.c.atkinson@lse.ac.uk*

RÉSUMÉ. *Dans ce travail, nous présentons une méthode de classification robuste, appelée Forward Search, pour la classification de tableaux de données de type qualitatif et nous la comparons avec la méthode des k-modes.*

MOTS-CLÉS : *Robust automatic classification, Dissimilarity, Forward Search, Number of clusters, Random starts.*

## 1. Introduction

The Forward Search is a powerful general method for detecting unidentified subsets and multiple masked outliers and for determining their effect on models fitted to the data. The search for multivariate data is given book length treatment by [ATK 04]. The basic idea is to start from a small, robustly chosen, subset of the data and to fit subsets of increasing size, in such a way that outliers and subsets of data not following the general structure are clearly revealed by diagnostic monitoring. The main classification tools in [ATK 04] are forward plots from searches started from subsets of observations in tentatively identified clusters. However, [ATK 06] and [ATK 07] show that for cluster definition, as opposed to outlier identification, several searches are to be preferred, the most informative being those that start in individual clusters and continue to add observations from the cluster until most of them have been used in estimation. This strategy seemingly requires that we know the clusters before running the searches. On the contrary, we use many searches with random starting points to provide evidence on cluster existence and definition. Two major advantages of random starts are a more automatic method of cluster identification and clearer guidance about the number of groups in the data. The purpose of this note is to show how the random-start Forward Search can be used for detecting clusters of multivariate categorical observations.

## 2. A few approaches to clustering categorical data

In spite of its practical relevance, clustering of discrete multivariate observations has received relatively little attention. A commonly used approach is to compute suitable measures of pairwise dissimilarity, such as the simple matching coefficient, and then to use these measures as input for hierarchical clustering algorithms. Hierarchical agglomeration plays an important role also in the recent clustering algorithm of [FRI 04], which can be used with categorical information. One major problem with hierarchical algorithms is that, as the number of objects grows, they rapidly become computationally unacceptable and provide results that are difficult to represent. They are, therefore, inappropriate for the analysis of large or even moderate data sets, like those frequently encountered, e.g., in marketing research. The *k*-modes algorithm of [HUA 98] and [CHA 01] is a notable exception which tries to combine the efficiency of the standard *k*-means paradigm for interval data with the need to take categorical information into account. This is accomplished by running a *k*-means type algorithm with simple matching dissi-

milarities instead of Euclidean distances and cluster modes instead of means. However, as with *k*-means, the results from *k*-modes can be very sensitive to the choice of the starting solution and even to the order of the observations in the data set. An additional shortcoming is that cluster modes may not be uniquely defined at some steps of the iterative procedure, thus leading to indeterminacy in the clustering solution. Also the detection of the appropriate number of groups may be problematic in the *k*-modes algorithm, and standard *k*-means heuristics for this purpose are not particularly helpful with categorical data.

## 3. The Forward Search

In this section we show how the random-start Forward Search algorithm can overcome most of the shortcomings of standard approaches for clustering categorical data. This algorithm is also robust, in the sense that it can highlight "unusual" observations. In a clustering context, unusual observations can either be multivariate outliers or inliers that do not belong to any of the main groups. The latter typology occurs frequently and is of great concern with categorical data. Let $S = \{u_1, u_2, \ldots, u_n\}$ be a set of $n$ units for which we observe $v$ nominal categorical variables $X_1, X_2, \ldots, X_v$. Denote with $x_{ij} \in \mathcal{C}^{(j)}$ the observed class of $X_j$ in unit $u_i$, and with $\mathcal{C}^{(j)}$ the set of possible classes for $X_j$. The number of such classes is $c_j$. We use dummy coding to process this categorical information. Let $x_{ij}^{(c)} = 1$ if $x_{ij} = c$ and $x_{ij}^{(c)} = 0$ otherwise. Unit $u_i$ can be represented as $x_i = [x_{i1}^{(1)}, \ldots, x_{i1}^{(c_1)}, \ldots, x_{iv}^{(1)}, \ldots, x_{iv}^{(c_v)}]'$, a vector of dimension $C = \sum_{j=1}^{v} c_j$. The dissimilarity between $u_i$ and $u_l$ is measured as

$$d(u_i, u_l) = \sum_{j=1}^{v} \sum_{c \in \mathcal{C}^{(j)}} (x_{ij}^{(c)} - x_{lj}^{(c)})^2, \qquad i, l = 1, \ldots, n. \tag{1}$$

It is easy to see that $d(u_i, u_l)$ is equivalent to simple matching, since both measures provide the same ordering of the dissimilarities among pairs of units. However, definition (1) has the advantage of being easily generalized to encompass differential weighting and correlation among the classes of the same variable ([KUR 70] ; [FRI 04]).

The next step is to define a measure of closeness between a unit and a population. We make the mild assumption that $x_i$ is a random observation from a population with class probabilities $\pi = [\pi_1^{(1)}, \ldots, \pi_1^{(c_1)}, \ldots, \pi_v^{(1)}, \ldots, \pi_v^{(c_v)}]'$, so that $\mathrm{E}(x_{ij}^{(c)}) = \pi_j^{(c)}$, for $j = 1, \ldots, v$ and $c \in \mathcal{C}^{(j)}$. Following (1), we compute the dissimilarity between $u_i$ and the mean vector $\pi$, or, when $\pi$ is unknown, its sample estimate $\hat{\pi} = [\hat{\pi}_1^{(1)}, \ldots, \hat{\pi}_1^{(c_1)}, \ldots, \hat{\pi}_v^{(1)}, \ldots, \hat{\pi}_v^{(c_v)}]'$, as

$$d_i = d(u_i, \hat{\pi}) = \sum_{j=1}^{v} \sum_{c \in \mathcal{C}^{(j)}} (x_{ij}^{(c)} - \hat{\pi}_j^{(c)})^2.$$

In the forward search the mean estimate $\hat{\pi}$ is repeatedly computed on a subset of $m$ observations, $S(m)$ say, yielding the $C$-dimensional vector $\hat{\pi}(m) = [\hat{\pi}_j^{(c)}(m)]'$. From this subset we obtain $n$ dissimilarities

$$d_i(m) = d(u_i, \hat{\pi}(m)) = \sum_{j=1}^{v} \sum_{c \in \mathcal{C}^{(j)}} (x_{ij}^{(c)} - \hat{\pi}_j^{(c)}(m))^2 \qquad i = 1, \ldots, n. \tag{2}$$

We start with a randomly selected subset of $m_0$ observations which grows in size during the search. When subset $S(m)$ is used in fitting, we order the dissimilarities (2) and take the units corresponding to the $m + 1$ smallest as the new subset $S(m + 1)$. Usually this process augments the subset by one unit, but sometimes two or more observations enter as one or more leave.

To detect potential clusters, we look at forward plots of quantities derived from the dissimilarities (2). One of the most useful plots is that of the minimum dissimilarity amongst units not in the subset

$$d_{\min}(m) = \min d_i(m) \qquad i \notin S(m). \tag{3}$$

Apart from some initial noise, the searches starting in subsets of units with similar features will lead to the same forward plot of $d_{\min}(m)$. We look at bunches of similar trajectories to identify the number of clusters and how

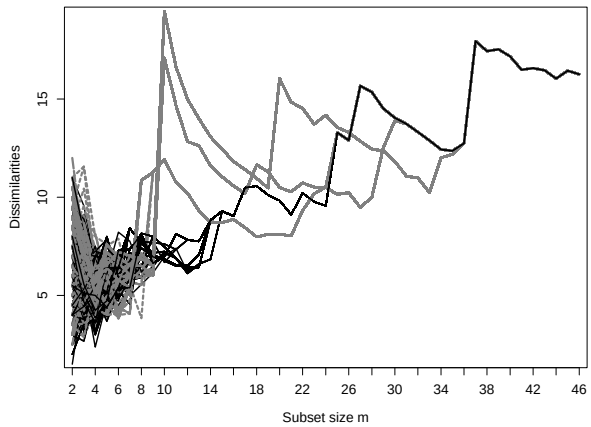



**FIG. 1.** *Soybean data. Forward plots of $d_{\min}(m)$ for 1000 searches with random starting points.*

they originate. We look at peaks in the forward plots of $d_{\min}(m)$ for precisely identifying such clusters. Cluster membership is obtained by looking at the units in the subsets when the peaks occur. Other valuable tools for detecting important cluster features include the forward plots of individual dissimilarities $d_i(m)$, $i = 1, \ldots, n$, the entry plot showing the composition of $S(m)$ at each step of the search, and the forward plots of sample proportions $\hat{\pi}_j^{(c)}(m)$, for $c \in \mathcal{C}^{(j)}$ and $j = 1, \ldots, v$.

## 4. Application

Some simulation evidence of the good performance of the random-start Forward Search algorithm for clustering categorical observations is provided by [CER 06], even when outliers are present. Here we compare the performance of the Forward Search with the *k*-modes algorithm on a real world data set that has frequently been used to test clustering algorithms for categorical data.

We analyze the soybean data set, available from the UCI Machine Learning Repository. It has 47 observations on 35 categorical variables, although 14 of them have only one category. Thus 21 variables are available for analysis. Each observation is labelled as one of four soybean diseases : Diaporthe Stem Canker, Charcoal Rot, Rhizoctonia Root Rot, and Phytophthora Rot. Except for the last disease, which has 17 observations, all other diseases have 10 instances each. We discard the disease information for clustering the soybean data set. Instead, we use this information to assess the recovery performance of different algorithms. [HUA 98] reports the results for the *k*-modes algorithm. This technique is sensitive to the order in which the units are listed, so repeated application to random permutations of the original records is required. Even in the unrealistic situation where $k$ is known, the algorithm is able to attain complete recovery of the four disease classes in only 14% of the permutations of the soybean data set. In 36% of these permutations the result is less than good, with more than 6 misclassified units.

Figure 1 contains the results of 1000 forward searches from randomly selected starting subsets of size $m_0 = 2$ from the soybean data set. For each search the dissimilarity $d_{\min}(m)$ is plotted. The most striking feature is that, from $m = 15$ onwards, the searches follow only four different trajectories, regardless of their starting point. These trajectories then merge towards the end of the algorithm, as units from different clusters enter into $S(m)$. The four distinct trajectories provide clear evidence about the group structure, information that is not available from the *k*-modes algorithm. We can identify the groups by looking at the peaks exhibited by these trajectories. The three peaks at $m = 10$, pictured in grey in Figure 1, are very pronounced and correspond to well separated clusters of the same size. The fourth trajectory, shown in boldface in Figure 1, increases at a more stable pace, an evidence of a cluster without sharp boundaries. The largest peak at $m = 10$ is for searches including units from the second group

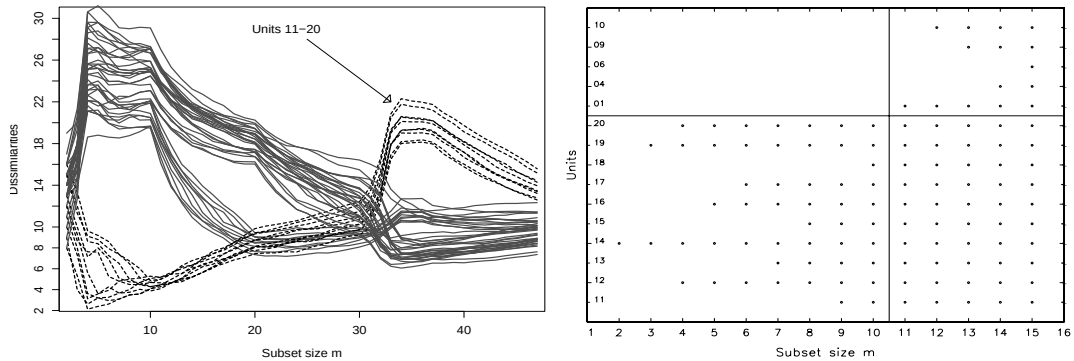



**FIG. 2.** *Soybean data : searches that originate the largest peak in $d_{\min}(m)$ at $m = 10$. Left-hand panel : forward plots of individual dissimilarities. Right-hand panel : entry plot for $m = 2, \ldots, 15$.*

(Charcoal Rot) numbered from 11 to 20 in the data set. In these searches, at $m = 10$ the observations from the other groups are all remote and have large dissimilarities from $S(m)$. Figure 2 depicts the structure emerging from the viewpoint provided by this cluster. The left-hand panel shows the forward plots of individual dissimilarities. The effect of adding units from a different group to a homogeneous subset is evident just after $m = 10$, when the first observation from the Diaporthe Stem Canker group enters into $S(m)$. The right-hand panel is the corresponding entry plot for $m = 2, \ldots, 15$, where the dots at each step show the units belonging to $S(m)$.

The general structure of the data is revealed by the different trajectory shapes in the left-hand panel of Figure 2. However, a more accurate dissection can be obtained by looking at the searches that give rise to the other peaks in Figure 1. Repeating Figure 2 for these searches shows that the second largest peak at $m = 10$ is for the units in the Diaporthe Stem Canker group, while the third largest peak at $m = 10$ identifies the observations with the Rhizoctonia Root Rot disease. The fourth disease corresponds to the remaining stable trajectory in Figure 1. This trajectory has its first peak at $m = 18$, although $d_{\min}(17)$ is only slightly smaller. The corresponding units in $S(18)$ belong to the Phytophthora Rot group, with the exception of unit 28, which is the last to join the subset. Observation 28 also belongs to the subset originating one of the peaks at $m = 10$, so it may be viewed as a "borderline" unit. Such borderline units need not be firmly classified in any specific cluster, at least at an exploratory stage.

## 5. Bibliography

[ATK 04] ATKINSON A. C., RIANI M., CERIOLI A., *Exploring Multivariate Data with the Forward Search*, Springer, New York, 2004.

[ATK 06] ATKINSON A., RIANI M., CERIOLI A., Random Start Forward Searches with Envelopes for Detecting Clusters in Multivariate Data, *Data Analysis, Classification and The Forward Search*, Zani, S., Cerioli, A., Riani, M. and Vichi, M., Springer, Berlin, in press, 2006.

[ATK 07] ATKINSON A. C., RIANI M., Exploratory Tools for Defining Clusters and Detecting Non-normality in Multivariate Data, submitted, 2007.

[CER 06] CERIOLI A., RIANI M., ATKINSON A. C., Robust classification with categorical variables, *COMPSTAT 2006*, Rizzi, A. and Vichi, M., Physica-Verlag, Heidelberg, in press, 2006.

[CHA 01] CHATURVEDI A., GREEN P. E., CARROLL J. D., *K*-modes clustering, *Journal of Classification*, vol. 18, 2001, p. 35–55.

[FRI 04] FRIEDMAN J. H., MEULMAN J. J., Clustering objects on subsets of attributes, *Journal of the Royal Statistical Society B*, vol. 66, 2004, p. 815–849.

[HUA 98] HUANG Z., Extensions to the *k*-means algorithm for clustering large data sets with categorical values, *Data Mining and Knowledge Discovery*, vol. 2, 1998, p. 283–304.

[KUR 70] KURCZYNSKI T. W., Generalized distance and discrete variables, *Biometrics*, vol. 26, 1970, p. 525–534.

# Apprentissage de correspondance entre schémas pour des documents semi-structurés
## Auteurs

**Patrick Gallinari, Ludovic Denoyer, Francis Maes, Guillaume Wisniewski**

*Université Pierre et Marie Curie, CNRS*
*8 rue du Capitaine Scott 75015 Paris*
*Patrick.gallinari@lip6.fr*

*MOTS-CLÉS : Apprentissage statistique, recherche d'information, structuration automatique de documents*

### 1 Résumé

L'apprentissage statistique a traditionnellement développé des méthodes pour le traitement de données représentées sous forme vectorielle ou séquentielle. De nombreux champs d'application demandent aujourd'hui de traiter des données plus complexes représentées sous la forme d'arbres ou de graphes. C'est le cas dans des domaines comme la recherche d'information multimédia, la biologie, le langage naturel. L'apprentissage a commencé à développer des méthodes pour le traitement de ces données structurées. Parmi les problématiques traitées, on retrouve celles classiques de l'analyse des données comme la discrimination ou la classification automatique mais adaptées au cas de données structurées. On voit également apparaître de nouvelles idées comme la production de données structurées (e.g. à partir d'une séquence de mots générer un arbre traduisant des relations entre éléments de cette séquence).
Nous considérons ici un problème spécifique d'apprentissage dans des données structurées qui combine ces deux aspects traitement et production de structure, qui est l'apprentissage de correspondances entre structures arborescentes. La motivation de cette étude provient de travaux en recherche d'information dans les documents textuels semi-structurés (e.g. XML, HTML). Ceux-ci associent à l'information de contenu des structures relationnelles qui permettent de coder des informations bien plus riches que le contenu de base des documents. Un problème clé lié à ces données est celui de l'hétérogénéité : les documents provenant de différentes sources auront des formats ou des schémas différents. L'interrogation de ces données passe par la connaissance de la correspondance entre les schémas des différentes sources. La multiplicité des sources d'information requiert une automatisation du processus permettant d'établir cette correspondance, qui passe par l'apprentissage à partir des données. Nous formulons ici le problème comme l'apprentissage de correspondance entre des arbres étiquetés. Ces arbres qui codent à la fois les structures relationnelles et le contenu sont définis dans des espaces de très grande taille. Après une rapide synthèse des travaux récents dans l'apprentissage de données structurées, nous décrivons un modèle stochastique pour l'apprentissage de ces correspondances. Celui-ci offre un cadre général dont nous développerons différentes instances adaptées à des problèmes spécifiques d'apprentissage de

correspondance entre schémas de documents. Une des difficultés majeures du problème est sa complexité liée à la dimension de l'espace de recherche. La recherche d'une structure cible optimale est un problème combinatoire qui est classiquement traité par des algorithmes de programmation dynamique. La dimension des espaces traités dans le cas de document rend prohibitives ces solutions classiques. Nous présentons des solutions heuristiques qui permettent de développer des algorithmes de complexité raisonnable pour traiter des problèmes de très grande taille. Des résultats expérimentaux seront présentés sur différents types de corpus.

.

## 2 Bibliographie


CHIDLOVSKI B., FUSELIER J., 2005, A probabilistic learning model for xml annotation of documents, IJCAI'05.

DAUME III H., MARCU D., 2005, Learning as search optimisation: approximate large margin methods for structured predictions, ICML'05.

DENOYER L., GALLINARI P., 2004, Bayesian networks model for semi structured document classification. Information processing and management. Volume 40 , Issue 5, 807-827.

DENOYER L., WISNIEWSKI G., GALLINARI P., 2004, Document structure matching for heterogeneous corpora, SIGIR Workshop on semi-structured documents.

DENOYER L., WISNIEWSKI G., GALLINARI P., 2006, Modèle probabiliste pour l'extraction de structures dans les documents semi-structurés --- Application aux documents Web, CORIA'06

TCHOCHANTARIDIS I., JOACHIMS T., HOFFMANN T., ALTUN Y., 2005, Large margin methods for structured and interdependent output variables, JMLR 6, 1453-1484.

# Bayesian Data Integration with Gaussian Process Priors

**Mark Girolami**

*Department of Computing Science*
*University of Glasgow*
*girolami@dcs.gla.ac.uk*

RÉSUMÉ. *The predictive accuracy of non-parametric classification methods, specifically kernel based, can be significantly improved by integrating diverse data sources in a principled manner. Methods for heterogeneous data integration within a classification context have been previously proposed utilising Semi-Definite-Programming and binary Support Vector Machines. In this contribution it is shown that full Bayesian inference can be achieved for integrating multiple datasets in the multi-way classification setting employing Gaussian Process priors and this is successfully illustrated on a protein fold prediction problem.*

MOTS-CLÉS : *Nonparametric Classification, Gaussian Process Prior, Bayesian Inference, Variational Bayesian Approximations.*

## 1. Introduction

Various emerging quantitative measurement technologies are producing genome, transcriptome and proteome-wide data collections which has motivated the development of data integration methods within an inferential framework. It has been demonstrated that for certain prediction tasks within computational biology synergistic improvements in performance can be obtained via integration of a number of (possibly heterogeneous) data sources. In [DIN 01] six different parameter representations of proteins were employed for fold recognition of proteins using Support Vector Machines (SVM). It was observed that certain dataset combinations provided increased accuracy over the use of any single datset. Likewise in [PAV 02] a comprehensive experimental study observed improvements in SVM based gene function prediction when data from both microarray expression and phylogentic profiles were combined. More recently protein network inference was shown to be improved when various genomic data sources were integrated [YAM 04]. In [BEN 05] it was shown that superior prediction accuracy of protein-protein interactions was obtainable when a number of diverse data types were combined in an SVM.

Whilst all of these papers exploited the kernel method [SHA 04] in providing a means of data fusion within SVM based classifiers it was only in [LAN 04] that a means of estimating an optimal linear combination of the kernel functions was presented using semi-definite programming. However, the methods developed in [LAN 04] are based on binary SVM's, whilst arguably the majority of classification problems within computational biology are inherently multiclass. It is unclear how this approach could be extended to discrimination over multiple-classes. In addition the SVM is non-probabilistic and whilst *post hoc* methods for obtaining predictive probabilities are available [PLA 99] these are not without problems such as overfitting. On the other hand Gaussian Process (GP) methods [RAS 06] for classification provide a very natural way to both integrate and infer optimal combinations of multiple heterogeneous datasets via composite covariance functions within the Bayesian framework. In this paper it is shown that GP's can be employed on large scale bioinformatics problems where there are multiple data sources and an example of protein fold prediction [DIN 01] is provided.

## 2. Data Fusion with Gaussian Process Priors

Let us denote each of $\mathcal{J}$ independent (possibly heterogeneous) feature representations, $\mathcal{F}_j(X)$, of an object $X$ by $\mathbf{x}_j \; \forall \;\; j = 1 \cdots \mathcal{J}$. For each object there is a corresponding polychotomous response target variable, $t$, so to model this reponse we assume an additive generalised multinomial probit regression model. Each distinct, and possibly heterogeneous, feature representation of $X$, $\mathcal{F}_j(X) = \mathbf{x}_j$, is nonlinearly transformed such that $f_j(\mathbf{x}_j) : \mathcal{F}_j \mapsto \mathbb{R}$ and a linear model is employed in this new space such that the overall nonlinear transformation is $f(X) = \sum_{j=1}^{\mathcal{J}} \beta_j f_j(\mathbf{x}_j)$.

### 2.1. *Composite Covariance Functions*

Rather than specifying a functional form for each of the functions $f_j(\mathbf{x}_j)$ we assume that each nonlinear function corresponds to a Gaussian process (GP) such that $f_j(\mathbf{x}_j) \sim GP(\boldsymbol{\theta}_j)$ where $GP(\boldsymbol{\theta}_j)$ corresponds to a Gaussian process with mean and covariance functions $m_j(\mathbf{x}_j)$ and $C_j(\mathbf{x}_j, \mathbf{x}'_j; \boldsymbol{\theta}_j)$ where $\boldsymbol{\theta}_j$ denotes a set of hyperparameters associated with the covariance function. Due to the assumed independence of the feature representations the overall nonlinear function will also be a realisation of a Gaussian process defined as $f(X) \sim GP\left(\boldsymbol{\theta}_1 \cdots \boldsymbol{\theta}_{\mathcal{J}}, \beta_1 \cdots \beta_{\mathcal{J}}\right)$ where now the overall mean and covariance functions follow as $\sum_{j=1}^{\mathcal{J}} \beta_j m_j(\mathbf{x}_j)$ and $\sum_{j=1}^{\mathcal{J}} \beta_j^2 C_j(\mathbf{x}_j, \mathbf{x}'_j; \boldsymbol{\theta}_j)$.

For target response values, $t \in \{1 \cdots K\}$ (i.e. a multiclass setting) and further assuming zero-mean GP functions then for $N$ object samples, $X_1 \cdots X_N$, each defined by the $\mathcal{J}$ feature representations, $\mathbf{x}_j^1 \cdots \mathbf{x}_j^N$, denoted by $\mathbf{X}_j$, and associated class specific response $\mathbf{f}_k = [f_k(X_1) \cdots f_k(X_N)]^{\mathsf{T}}$ we have the overall GP prior as a multivariate Normal such that

$$\mathbf{f}_k \mid \mathbf{X}_{j=1 \cdots J}, \boldsymbol{\theta}_{1k}, \cdots \boldsymbol{\theta}_{\mathcal{J}k}, \alpha_{1k} \cdots \alpha_{\mathcal{J}k} \sim \mathcal{N}_{\mathbf{f}_k}\left(\mathbf{0}, \sum\nolimits_j \alpha_{jk} \mathbf{C}_{jk}(\boldsymbol{\theta}_{jk})\right)$$

where we employ $\alpha_{jk}$ to denote the positive random variables $\beta_{jk}^2$ and each $\mathbf{C}_{jk}(\boldsymbol{\theta}_{jk})$ is an $N \times N$ matrix with elements $C_j(\mathbf{x}_j^m, \mathbf{x}_j^n; \boldsymbol{\theta}_{jk})$.

A GP functional prior, over all possible responses (classes), is now available where possibly heterogeneous data sources are integrated via the composite covariance function. It is then, in principle, a straightforward matter to perform Bayesian inference with this model and no further recourse to *ad hoc* binary classifier combination methods or ancillary optimisations to obtain the data combination weights is required.

### 2.2. *Bayesian Inference*

The inference methods detailed in [GIR 06] are adopted where the auxiliary variables $y_{nk} = f_k(X_n) + \epsilon_{nk}$, $\epsilon_{nk} \sim \mathcal{N}(0, 1)$ are introduced. The $N \times 1$ dimensional vector of target class values associated with each $X_n$ is given as $\mathbf{t}$ where each element $t_n \in \{1, \cdots, K\}$. The $N \times K$ matrix of GP random variables $f_k(X_n)$ is denoted by $\mathbf{F}$. We represent the $N \times 1$ dimensional columns of $\mathbf{F}$ by $\mathbf{F}_{\cdot,k}$ and the corresponding $K \times 1$ dimensional vectors, $\mathbf{F}_{n,\cdot}$, which are formed by the indexed rows of $\mathbf{F}$ . The $N \times K$ matrix of auxiliary variables $y_{nk}$ is represented as $\mathbf{Y}$, where the $N \times 1$ dimensional columns are denoted by $\mathbf{Y}_{\cdot,k}$ and the corresponding $K \times 1$ dimensional rows as $\mathbf{Y}_{n,\cdot}$ . The multinomial probit likelihood [GIR 06] is adopted which follows as

$$t_n = j \quad \text{if} \quad y_{nj} = \max_{1 \leq k \leq K} \{y_{nk}\}$$

and this has the effect of dividing $\mathbb{R}^K$ into $K$ non-overlapping $K$-dimensional cones $\mathcal{C}_k = \{\mathbf{y} : y_k > y_i, k \neq i\}$ where $\mathbb{R}^K = \cup_k \mathcal{C}_k$ and so each $P(t_n = i | \mathbf{Y}_{n,\cdot})$ can be represented as $\delta(y_{ni} > y_{nk} \; \forall \; k \neq i)$. Independent Gamma priors, with parameters $\boldsymbol{\varphi}_k$, are placed on each $\alpha_{kj}$ and the individual components of $\boldsymbol{\theta}_{jk}$ (denote $\boldsymbol{\Theta}_k = \{\boldsymbol{\theta}_{jk}\}_{j=1 \cdots \mathcal{J}}$), so this defines the full model likelihood and associated priors.

### 2.3. *MCMC Procedure*

Samples from the full posterior $P(\mathbf{Y}, \mathbf{F}, \boldsymbol{\Theta}_{1\cdots K}, \boldsymbol{\alpha}_{1\cdots K}, \boldsymbol{\varphi}_{1\cdots K}|X_{1\cdots N}, \mathbf{t}, \mathbf{a}, \mathbf{b})$ (where $\mathbf{a}$ & $\mathbf{b}$ are hyper-parameters associated with the gamma priors) can be obtained from the following Metropolis-within-Blocked-Gibbs Sampling scheme indexing over all $n = 1 \cdots N$ and $k = 1 \cdots K$.

$$\begin{aligned}
\mathbf{Y}_{n,\cdot}^{(i+1)}|\mathbf{F}_{n,\cdot}^{(i)}, t_n &\sim \mathcal{TN}(\mathbf{F}_{n,\cdot}^{(i)}, \mathbf{I}, t_n) \\
\mathbf{F}_{\cdot,k}^{(i+1)}|\mathbf{Y}_{\cdot,k}^{(i+1)}, \boldsymbol{\Theta}_k^{(i)}, \boldsymbol{\alpha}_k^{(i)}, X_{1,\cdots,N} &\sim \mathcal{N}(\boldsymbol{\Sigma}_k^{(i)}\mathbf{Y}_{\cdot,k}^{(i+1)}, \boldsymbol{\Sigma}_k^{(i)}) \\
\boldsymbol{\Theta}_1^{(i+1)}, \boldsymbol{\alpha}_1^{(i+1)}|\mathbf{F}_{\cdot,1}^{(i+1)}, \boldsymbol{\varphi}_1^{(i)}, X_{1,\cdots,N} &\sim P(\boldsymbol{\Theta}_k^{(i+1)}, \boldsymbol{\alpha}_k^{(i+1)}) \\
\boldsymbol{\varphi}_k^{(i+1)}|\boldsymbol{\Theta}_k^{(i+1)}, \boldsymbol{\alpha}_k^{(i+1)}, a_k, b_k &\sim P(\boldsymbol{\varphi}_k^{(i+1)})
\end{aligned}$$

where $\mathcal{TN}(\mathbf{F}_{n,\cdot}, \mathbf{I}, t_n)$ denotes a conic truncation of a multivariate Gaussian. An accept-reject strategy can be employed in sampling from the conic truncated Gaussian however this will very quickly become inefficient for problems with moderately large numbers of classes and as such a further Gibbs sampling scheme may be required.

Each $\boldsymbol{\Sigma}_k^{(i)} = \mathbf{C}_k^{(i)}(\mathbf{I} + \mathbf{C}_k^{(i)})^{-1}$ and $\mathbf{C}_k^{(i)} = \sum_{j=1} \alpha_{jk}^{(i)} \mathbf{C}_{jk}(\boldsymbol{\theta}_{jk}^{(i)})$ with the elements of $\mathbf{C}_{jk}(\boldsymbol{\theta}_{jk}^{(i)})$ defined as $C_j(\mathbf{x}_j^m, \mathbf{x}_j^n; \boldsymbol{\theta}_{jk}^{(i)})$. A Metropolis sub-sampler is required to obtain samples for the conditional $P(\boldsymbol{\Theta}_k^{(i+1)}, \boldsymbol{\alpha}_k^{(i+1)})$. Finally $P(\boldsymbol{\varphi}_k^{(i+1)})$ is a simple product of Gamma distributions.

### 2.4. *Obtaining Predictive Posteriors*

The predictive likelihood of a test sample $X_*$ is $P(t_* = k|X_*, X_{1\cdots N}, \mathbf{t}, \mathbf{a}, \mathbf{b})$ which can be obtained by integrating over the posterior and predictive prior such that

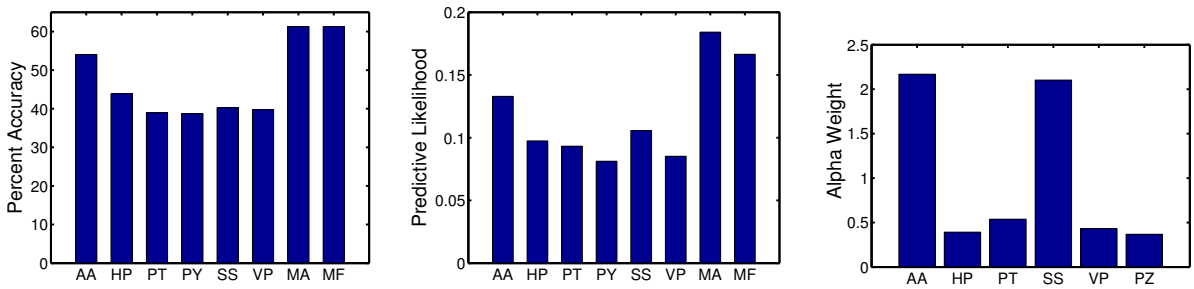
$$\int P(t_* = k|\mathbf{f}_*)p(\mathbf{f}_*|\boldsymbol{\Omega}, X_*, X_{1\cdots N})p(\boldsymbol{\Omega}|X_{1\cdots N}, \mathbf{t}, \mathbf{a}, \mathbf{b})d\mathbf{f}_* d\boldsymbol{\Omega}$$

where $\boldsymbol{\Omega} = \mathbf{Y}, \boldsymbol{\Theta}_{1\cdots K}, \boldsymbol{\alpha}_{1\cdots K}$. A Monte-Carlo estimate is obtained by using samples drawn from the full posterior $\frac{1}{S}\sum_{s=1}^{S} \int P(t_* = k|\mathbf{f}_*)p(\mathbf{f}_*|\boldsymbol{\Omega}^{(s)}, X_*, X_{1\cdots N})d\mathbf{f}_*$and the integral over the predictive prior requires further conditional samples to be drawn from each $p(\mathbf{f}_*|\boldsymbol{\Omega}^{(s)}, X_{1\cdots N})$ finally yielding an estimate of $P(t_* = k|X_*, X_{1\cdots N}, \mathbf{t}, \mathbf{a}, \mathbf{b})$

$$\frac{1}{LS}\sum_{l=1}^{L}\sum_{s=1}^{S} P\left(t_* = k|\mathbf{f}_*^{(l|s)}\right) = \frac{1}{LS}\sum_{l=1}^{L}\sum_{s=1}^{S} E_{p(u)}\left\{\prod_{j\neq k}\Phi\left(u + f_{*,k}^{(l|s)} - f_{*,j}^{(l|s)}\right)\right\}$$

## 3. Protein Fold Prediction with GP Based Data Fusion

To illustrate the proposed GP based method of data integration a protein fold classification problem originally studied in [DIN 01] is considered. The task is to devise a predictor of 27 SCOP classes from a set of low homology protein sequences. Six different feature sets are available characterizing (1) Amino Acid composition (AA) ; (2) Hydrophobicity profile (HP) ; (3) Polarity (PT) ; (4) Polarizability (PY) ; (5) Secondary Structure (SS) ; (6) Van der Waals volume. In [DIN 01] a number of combination strategies were employed in devising a multiway classifier from a series of binary SVM's. The best predictive accuracy obtained on an independent set of low sequence similarity proteins was 53%. It was noted after extensive careful experimentation by the authors that a combination of Gaussian kernels each composed of the (AA), (SS) and (HP) datasets improved predictive accuracy. We employ a variational approximation of GP based method detailed in [GIR 06] to devise a classifier for this task where now we employ a composite covariance function, a linear combination of RBF functions for each data set. Figure (1) shows the predictive performance of the GP classifier in terms of percentage prediction accuracy (a) and predictive likelihood on the test set (b). We note a significant synergistic increase in performance when all data sets are




(a) (b) (c)

**FIG. 1.** *(a) The prediction accuracy for each individual data set and the corresponding combinations, (MA) employing inferred weights and (MF) employing a fixed weighting scheme (b) The predictive likelihood achieved for each individual data set and with the integrated data (c) The posterior mean values of the covariance function weights* $\alpha_1 \cdots \alpha_6$.

combined and weighted (MA). Although the test error is the same for an equal weighting of the data sets (MF) and that obtained using the proposed inference procedure (MA) for (MA) there is a small increase in predictive likelihood i.e. more confident correct predictions being made. It is interesting to note that the weighting obtained (posterior mean for $\boldsymbol{\alpha}$) Figure (1.c) weights the (AA) & (SS) with equal importance whilst other data sets play less of a role in performance improvement. The overall performance accuracy achieved is 62%.

**Références**

[BEN 05] BEN-HUR A., NOBLE W., Kernel methods for predicting protein-protein interactions, *Bioinformatics*, vol. 21, Suppl. 1, 2005, p. 38-46.

[DIN 01] DING C., DUBCHAK I., Multi-class protein fold recognition using support vector machines and neural networks, *Bioinformatics*, vol. 17, 2001, p. 349-358.

[GIR 06] GIROLAMI M., ROGERS S., Variational Bayesian Multinomial Probit Regression with Gaussian Process Priors, *Neural Computation*, vol. 18, n$^o$ 8, 2006, p. 1790–1817.

[LAN 04] LANCKRIET G. R. G., BIE T. D., CRISTIANINI N., JORDAN M. I., NOBLE W. S., A statistical framework for genomic data fusion, *Bioinformatics*, vol. 20, 2004, p. 2626-2635.

[PAV 02] PAVLIDIS P., WESTON J., CAI J., NOBLE W. S., Learning gene functional classifications from multiple data types, *Journal of Computational Biology*, vol. 9, n$^o$ 2, 2002, p. 401-411.

[PLA 99] PLATT J., Probabilities for support vector machines, SMOLA A., BARTLETT P., SCHÖLKOPF B., SCHUURMANS D., Eds., *Advances in Large Margin Classifiers*, MIT Press, 1999, p. 61-74.

[RAS 06] RASMUSSEN C. E., WILLIAMS C. K. I., *Gaussian Processes for Machine Learning*, MIT Press, 2006.

[SHA 04] SHAWE-TAYLOR J., CRISTIANINI N., *Kernel methods for Pattern Analysis*, Cambridge University Press, Cambridge, 2004.

[YAM 04] YAMANISHI Y., VERT J. P., KANEHISA M., Protein network inference from multiple genomic data : a supervised approach, *Bioinformatics*, vol. 20, Suppl. 1, 2004, p. 363-370.

# Détection de Données Aberrantes par Boosting Itéré


## Jean-Michel Poggi

*Laboratoire de Mathématique - U.M.R. C8628, « Probabilités, Statistique et Modélisation »,*
*Université Paris-Sud, Bât.425, 91405 Orsay Cedex, France*
*Jean-Michel.Poggi@math.u-psud.fr*



*RÉSUMÉ.*

*Une procédure de détection de valeurs aberrantes dans les problèmes de régression, basée sur l'information fournie par le boosting d'arbres de régression CART, est proposée.*

*L'idée maîtresse consiste à sélectionner l'observation la plus fréquemment rééchantillonnée au cours des itérations du boosting puis de recommencer après l'avoir retirée. Le critère de sélection est basé sur l'application de l'inégalité de Tchebychev au maximum, au cours des itérations du boosting, du nombre moyen d'apparitions dans les échantillons bootstrap.*

*La procédure ne fait donc pas d'hypothèse sur la distribution du bruit et sélectionne les valeurs aberrantes comme des observations particulièrement difficiles à prévoir. On considère un grand nombre de jeux de données réelles ou artificielles et une étude comparative avec des méthodes éprouvées en montre l'intérêt.*

*Cet exposé est issu de : Cheze N., Poggi J-M., "Outlier Detection by Boosting Regression Trees", Prépublications mathématiques Orsay, 2005-17, 23 p.*



*MOTS-CLÉS :* *Boosting; CART; Régression; Valeurs aberrantes.*


### Bibliographie


Borra, S., Di Ciaccio A. (2002), "Improving nonparametric regression methods by bagging and boosting", *Computational Statistics and Data Analysis*, vol.38, n.4, pp. 407-420.

Cheze N., Poggi J-M. (2005), "Outlier Detection by Boosting Regression Trees", *Prépub. Math. Orsay*, 2005-17, 23 p.

Breiman L., Friedman J. H., Olshen R. A., Stone C. J. (1984), *Classification and Regression Trees*. Chapman & Hall.

Breiman L. (1998), "Arcing classifiers", *The Annals of Statistics* 26, 801-849.

Drucker, H. (1997), "Improving Regressors using Boosting Techniques", *Proc. of the 14th Int. Conf. on Machine Learning*, 107-115, Morgan Kaufmann.

Freund Y., Schapire R. E. (1997), "A decision-theoretic generalization of on-line learning and an application to boosting", *Journal of Computer and System Sciences*, 55 (1), 119-139.

Gey, S., Poggi, J.-M. (2006), "Boosting and instability for regression trees", *Computational Statistics & Data Analysis*, vol. 50(2), 533-550.

Rousseeuw, P.J. and Leroy, A. (1987), *Robust regression and outlier detection*, Wiley.

Rousseeuw P., Driessen V. (1999), "A Fast Algorithm for the Minimum Covariance Determinant Estimator", *Technometrics*, 41, pp. 212-223.

Verboven, S., Hubert, M. (2005), LIBRA: a MATLAB Library for Robust Analysis, *Chemometrics and Intelligent Laboratory Systems*, 75, 127-136.

# The family of hierarchical classes models: State of the art


## Iven Van Mechelen

*Département de Psychologie,*
*Université de Leuven,*
*Tiensestraat 102*
*3000 LEUVEN, Belgique*



*ABSTRACT. Hierarchical classes models consitute a distinct family of models for two- or multiway data. The models include an approximate reconstruction of the data, along with clusterings of each of the modes involved in the data and a representation of the structure of (generalized) implication-type relations that can be naturally defined on these modes. A special feature of the models is that they go with a comprehensive graphical representation. In this paper, I will present a state-of-the-art overview of research on the hierarchical classes family. The overview will start on a deterministic level with an introduction into the original (disjunctive) hierarchical classes model for binary two-way data as developed by De Boeck and Rosenberg (1988); this model will further be reconceptualized as a particular type of biclustering model. Subsequently, we will move on to a stochastic level with a discussion of several possible extensions of the deterministic model with distributional assumptions. Then, the estimation of deterministic as well as stochastic hierarchical classes models will be considered, along with issues to be dealt with in the analysis of data on the basis of the models. We conclude with an overview of recent research on several generalizations of the original hierarchical classes model, including (a) conjunctive models, (b) models for rating- and real-valued data, (c) models for multiway data, and (d) models for multiblock data or data fusion.*



*KEY WORDS : Hierarchical classes models, simultaneous clustering*



*RÉSUMÉ. Les modèles à classes hiérarchiques constituent une famille distincte de modèles pour des tableaux à deux ou plusieurs entrées. Les modèles comprennent une reconstitution approximative des données, aussi bien que des classifications de chacun des ensembles inclus dans les données et une représentation des relations d'implication (généralisées) qui peuvent être définies d'une façon naturelle sur ces ensembles. Une characteristique particulière de ces modèles est leur association avec une représentation graphique compréhensive. Dans ce papier, je propose une revue des recherches sur la famille des modèles à classes hiérarchiques. Cette revue commence, sur un plan déterministe, avec une introduction sur le modèle original (disjonctive) à classes hiérarchiques pour des données binaires à deux entrées comme proposé par De Boeck et Rosenberg (1988); ce modèle sera aussi reconceptualisé comme une sorte particulière de modèle de classification simultanée. Ensuite, sur un plan stochastique, on discutera plusieurs extensions du modèle déterministe avec des hypothèses de distribution. Puis, l'estimation des modèles à classes hiérarchiques deterministes ainsi que stochastiques sera considérée, aussi bien que plusieurs questions concernant l'analyse des données à la base des modèles. On conclut par une revue de recherches récentes sur plusieurs généralisations du modèle à classes hiérarchiques original, comprenant: (a) des modèles conjonctives, (b) des modèles pour des données réelles, (c) des modèles pour des données à plusieurs entrées, et (d) des modéles pour des données multi-tableaux ou la fusion de données.*



*MOTS-CLÉS : Modèles à classes hiérarchiques, classification simultanée*

**Communications**

# Analyse en Composantes Principales Mixte

**Rafik Abdesselam**

*CREM UMR CNRS 6154 - Université de Caen, Basse-Normandie*
*Esplanade de la paix,*
*14032 CAEN Cedex*
*rafik.abdesselam@unicaen.fr*

RÉSUMÉ. *Le traitement simultané de données mixtes (quantitatives et qualitatives) ne peut pas se réaliser directement par les méthodes classiques de la statistique exploratoire multidimensionnelle. Dans ce travail, l'analyse factorielle sur données mixtes proposée est une analyse en composantes principales normée après transformation des indicatrices des variables qualitatives en variables quantitatives au travers de projections de nuages de points dans l'espace des individus correspondant à des analyses de la variance multivariée. La méthode est évaluée sur la base d'une application sur données réelles mixtes.*

MOTS-CLÉS : *Analyse en Composantes Principales, Analyse de la variance multivariée, Rapport de corrélation.*

## 1. Introduction

Dans le cadre d'un traitement de données multidimensionnelles, il est très fréquent que le thème homogène de variables à analyser soit constitué de données mixtes (variables quantitatives et qualitatives). La méthodologie usuelle de traitement consiste soit à mettre les variables qualitatives [resp. quantitatives] en éléments illustratifs dans une Analyse en Composantes Principales (ACP) [resp. Analyse des Correspondances Multiples (ACM)], soit encore, de discrétiser les variables quantitatives du thème en variables qualitatives en vue d'une ACM ; ce qui introduit très souvent un biais dû au choix du nombre de classes et de leurs amplitudes égales ou différentes, et qui occasionne une perte d'information. De nombreux chercheurs se sont interéssés à cette problématique et ont proposé des méthodes qui traitent simultanément les deux types de variables en éléments actifs : l'ACP avec indicatrices introduite par M. Tenenhaus [TEN 77], [ESC 79], [SAP 90], [YOU 81], [PAG 02] et plus récemment l'Analyse Factorielle de Données Mixtes (AFDM) proposée par J. Pagès ([PAG 04]).

L'Analyse en Composantes Principales Mixte (ACPM) proposée est une ACP normée des variables quantitatives et des indicatrices des variables qualitatives transformées en variables combinaisons linéaires des variables quantitatives, à partir de projections orthogonales de nuages de points dans l'espace des individus. L'ACPM est formellement proche de l'AFDM en ce sens qu'elles consistent toutes les deux à "transformer" les variables qualitatives en quantitatives en vue d'une ACP, mais elles se différencient toutefois par le type d'ACP et le choix de la tranformation.

## 2. Transformation des données qualitatives en quantitatives

On dispose de $p$ variables quantitatives centrées $\{x^j; j = 1, p\}$ et de $m$ variables qualitatives $(y_1, \ldots, y_l, \ldots, y_m)$ auxquelles sont associées au total $q = \sum_{l=1}^{m} q_l$ variables indicatrices non centrées $\{y_l^k; k = 1, q_l\}_{\{l=1,m\}}$.

On utilisera les notations suivantes pour construire la matrice $M$ associée au produit scalaire de référence dans l'espace des individus $E = E_x \oplus E_{y_1} \oplus \ldots \oplus E_{y_l} \oplus \ldots \oplus E_{y_m} = \mathbb{R}^{p+q}$.

$E_x = \mathbb{R}^p$ étant le sous-espace des individus associé par dualité aux $p$ variables quantitatives centrées,
$E_{y_l} = \mathbb{R}^{q_l}$ étant le sous-espace des individus associé par dualité aux $q_l$ variables indicatrices non centrées des modalités de $y_l$,
$\oplus\{E_{y_l}\}_{\{l=1,m\}} = \mathbb{R}^q$ étant les $m$ sous-espaces des individus associés aux $m$ variables qualitatives,
$X$ est la matrice d'ordre $(n,p)$ des données quantitatives associée aux variables $\{x^j; j=1,p\}$,
$Y_l$ est la matrice des données qualitatives d'ordre $(n,q_l)$ associée aux $q_l$ indicatrices non centrées $\{y_l^k; k=1,q_l\}$ de la lème variable qualitative $y_l$,
$M_{y_l}$ [ resp. $M_x$ ] est la matrice du produit scalaire dans l'espace $E_{y_l}$ [ resp. $E_x$ ], isomorphe du sous-espace de même nom, via l'injection canonique,
$D_{y_l}$ est la matrice diagonale des poids définie par $[D_{y_l}]_{kk} = n_k/n$ pour tout $k = 1, q_l$, où $n_k$ est le nombre d'individus possédant la modalité $k$ de $y_l$,
$N_x = \{x_i \in E_x\ ;\ i = 1,n\}$ est le nuage des individus associé au tableau de données quantitatives $X$,
$N_{y_l} = \{y_i \in E_{y_l}\ ;\ i = 1,n\}$ est le nuage des individus associé au tableau de données qualitatives $Y_l$,
$P^M_{E_{y_l}}$ est l'opérateur de projection $M$-orthogonale sur $E_{y_l}$.

La transformation des données qualitatives en quantitatives se fait à l'aide de la construction statistique et géométrique de $m$ nuages $\hat{N}_x^{y_l} = \{P^M_{E_{y_l}}(x_i); x_i \in N_x\} \subset E_{y_l} \subset E$. Pour tout $l = 1, m$, le sous-espace $E_{y_l}$ est considéré comme sous-espace explicatif sur lequel est projeté M-orthogonalement le nuage $N_x$ des données quantitatives du sous-espace à expliquer $E_x$.

Le produit scalaire $M$ de référence dans l'espace des individus $E = E_x \oplus \{E_{y_l}\}_{\{l=1,m\}} = \mathbb{R}^{p+q}$ joue un rôle fondamental dans notre approche pour réaliser les $m$ projections. A priori, pour tout $l = 1, m$, le produit scalaire $M_{y_l}$ intra le sous-espace $E_{y_l}$ sur lequel on projette, pourrait être quelconque ; le choix $M_{y_l} = \chi^2_{y_l} = D^{-1}_{y_l}$ (distance du khi-deux) simplifie les calculs. Quant au produit scalaire $M_x$, intra le sous-espace $E_x$, on choisit $M_x = V_x^+$ (distance de Mahalanobis) afin de maximiser le critère d'inertie expliquée.

La matrice $M$ d'ordre $(p+q)$ associée au produit scalaire partitionné et équilibré dans $E$, relativement à l'ensemble des couples de variables $\{x^j; j=1,p\}$ et $\{{y_l}^k; k=1,q_l\}_{l=1,m}$, est telle que :

$$\left\{\begin{array}{ll} M_x = V_x^+ \quad ; \quad M_{y_l} = \chi^2_{y_l} & pour\ l = 1,m \\ M_{xy_l} = M_x[(V_x M_x)^{\frac{1}{2}}]^+ V_{xy_l} M_{y_l}[(V_{y_l} M_{y_l})^{\frac{1}{2}}]^+ = V_x^+ V_{xy_l} \chi^2_{y_l} & pour\ l = 1,m \\ M_{y_l y_{l'}} = M_{y_l}[(V_{y_l} M_{y_l})^{\frac{1}{2}}]^+ V_{y_l y_{l'}} M_{y_{l'}}[(V_{y_{l'}} M_{y_{l'}})^{\frac{1}{2}}]^+ = \chi^2_{y_l} V_{y_l y_{l'}} \chi^2_{y_{l'}} & pour\ l \neq l' \end{array}\right.$$

où $V_{y_l} = {}^tY_l D Y_l$ , $V_x = {}^tX D X$ et $V_{xy_l} = {}^tX D Y_l$ désignent les matrices de variances-covariances, $D = (1/n) I_n$ est la matrice diagonale des poids des n individus et $I_n$ la matrice unité d'ordre $n$. $[(V_x M_x)^{\frac{1}{2}}]^+]$ est l'inverse généralisée de Moore-Penrose de $(V_x M_x)^{\frac{1}{2}}]$. L'introduction d'inverses généralisées est une conséquence de la singularité des matrices $V_x$ et $V_{y_l}$ , lorsque rang $[V_x] < p$ et puisque rang $[V_{y_l}] = q_l - 1$.

Le produit scalaire $M$ positionne les sous-espaces des individus $E_x$ et $\{E_{y_l}\}_{\{l=1,m\}}$ tel que l'on puisse traduire en terme d'inertie dans l'espace des individus $E$, la structure des associations observées entre les sous-espaces des variables $F_x$ et $F_y$ associés par dualité dans l'espace des variables $F = \mathbb{R}^n$ muni de la métrique diagonale des poids $D$.

Pour tout $l = 1$ à $m$, on note $\hat{X}^{y_l} = X V_x^+ V_{xy_l}$ la matrice des données d'ordre $(n, q_l)$ associée au nuage projeté $\hat{N}_x^{y_l} \subset E_{y_l}$ et $G_l = {}^t\hat{X}^{y_l} D\, 1_n$ le vecteur coordonnées de son centre de gavité, où $1_n$ désigne le vecteur unité d'ordre $n$.

REMARQUE. — L'ACP du triplet ( $\hat{X}^{y_l}; \chi^2_{y_l}; D$ ) est équivalente à l'analyse de la variance multivariée (MANOVA) entre les $p$ variables quantitatives et les $q_l$ indicatrices associées aux niveaux du facteur explicatif $y_l$, et dont l'inertie expliquée, $I(\hat{N}_x^{y_l}) = trace(V_{y_l x} V_x^+ V_{xy_l} \chi^2_{y_l})$, est égale à la trace de Pillai.

On note alors, $Z_l = Y_l - {}^tG_l$ la matrice des données quantitatives d'ordre $(n, q_l)$ associée au nuage $N_{z_l} \subset E_{y_l}$ : le nuage $N_{y_l}$ associé au tableau de données qualitatives est centré relativement au nuage $\hat{N}_x^{y_l}$, et dont l'inertie $I(N_{z_l}) = q_l - 1$.

Ainsi, les $m$ tableaux d'indicatrices non centrées $[\, Y_1, \ldots, Y_l, \ldots, Y_m \,]$ associés aux $m$ variables qualitatives sont centrés relativement à $[\, G_1, \ldots, G_l, \ldots, G_m \,]$ puis remplacés par les $m$ tableaux $[\, Z_1, \ldots, Z_l, \ldots, Z_m \,]$ de variables quantitatives via les $m$ tableaux $[\, \hat{X}^{y_1}, \ldots, \hat{X}^{y_l}, \ldots, \hat{X}^{y_m} \,]$ des $m$ MANOVA séparées.

**Définition**

L'ACPM du tableau de données mixtes $[\, X,\, Y_1, \ldots, Y_l, \ldots,\, Y_m \,]$ d'ordre (n ; p+q) consiste à effectuer l'ACP normée du tableau de données (quantitatives) $[\, X,\, Z_1,\, \ldots,\, Z_l, \ldots,\, Z_m \,]$.

REMARQUE. — D'un point de vue méthodologique, l'ACPM se présente comme un enchaînement d'analyses factorielles : - MANOVA ( pour transformer les variables qualitatives en quantitatives ; on tient ainsi compte des rapports de corrélation) - ACP normée ( pour synthétiser les corrélations de l'ensemble des variables quantitatives et qualitatives transformées ). D'un point de vue pratique, il suffit simplement de centrer chaque tableau d'indicatrices $Y_l$ par rapport au centre de gravité correspondant $G_l = V_{y_l x} V_x^{+}\, {}^tXD1_n$, puis d'exécuter un programme classique d'ACP.

REMARQUE. — L'AFDM proposée par J. Pagès dans [PAG 04] du même tableau de données mixtes consiste à effectuer l'ACP usuelle du tableau de données $[X^{cr}, Y_1/\sqrt{D_{y_l}}, \ldots, Y_l/\sqrt{D_{y_l}}, \ldots, Y_m/\sqrt{D_{y_m}}]$, où les variables quantitatives $X^{cr}$ sont normées (centrées et réduites) et les indicatrices des variables qualitatives sont affectées d'une pondération. Pour tout $l = 1, m$, cela revient à diviser les valeurs de l'indicatrice $y_l^k$ de la variable $y_l$ par $\sqrt{\frac{n_k}{n}}$ : codage ACP de l'indicatrice $y_l^k$.

NOTE. — L'objectif de l'ACPM est le même que celui de l'AFDM, à savoir rechercher les facteurs principaux $F^s$ qui maximisent le critère mixte proposé par G. Saporta [SAP 90] et B. Escofier [ESC 79] :

$$\sum_{j=1}^{p} r^2(x^j, F^s) + \sum_{l=1}^{m} \eta^2(y_l\,, F^s) = \sum_{j=1}^{p} \cos^2\theta_{js} + \sum_{l=1}^{m} \cos^2\theta_{ls},$$

où $r^2$ et $\eta^2$ sont respectivement le carré du coefficient de corrélation linéaire des variables quantitatives et le rapport de corrélation des variables qualitatives avec le facteur, et $\theta$ l'angle entre les vecteurs correspondant.

## 3. Exemple d'application

On s'intéresse à un échantillon de 27 petites voitures du marché belge [LAM 90]. On dispose d'un thème homogène de 9 variables mixtes dont $p = 6$ caractétristiques continues : la cylindrée, la consommation urbaine, la vitesse maximum, le volume du coffre, le rapport poids/puissance et la longueur, et $m = 3$ caractéristiques nominales : la puissance fiscale (4CV/5CV/6CV), la marque du constructeur (Française/Etrangère) et quatre classes de prix (CP1/CP2/CP3/CP4) totalisant $q = 9$ modalités. L'objectif est de synthétiser au sens des corrélations l'ensemble de ces caractéristiques mixtes.

Les principaux résultats de l'ACPM sont présentés dans les tableaux et graphiques suivants, ils s'intéprètent avec les règles classiques d'une ACP.

**TAB. 1.** *Statistiques sommaires - Rapports de corrélation*

| Libellé | Moyenne | Ecart-type | Minimum | Maximum | P. FISCALE | MARQUE | C.PRIX |
|---|---|---|---|---|---|---|---|
| CONS | 7.14 | 1.12 | 5.60 | 9.30 | 0.809 | 0.009 | 0.636 |
| CYLI | 1165.63 | 204.17 | 903.00 | 1597.00 | 0.843 | 0.002 | 0.846 |
| VITE | 154.26 | 21.94 | 115.00 | 200.00 | 0.690 | 0.010 | 0.826 |
| VOLU | 901.41 | 301.67 | 202.00 | 1200.00 | 0.136 | 0.168 | 0.029 |
| RP/P | 18.65 | 5.42 | 10.20 | 33.10 | 0.562 | 0.042 | 0.660 |
| LONG | 3.62 | 0.07 | 3.40 | 3.70 | 0.094 | 0.022 | 0.163 |

Le tableau 1 récapitule les statistiques élémentaires des variables quantitatives ainsi que leurs rapports de corrélation avec les variables qualitatives considérées.

**TAB. 2.** *Moyennes pour le centrage des indicatrices*

| $G_1$ | | | $G_2$ | | $G_3$ | | | |
|---|---|---|---|---|---|---|---|---|
| 4CV | 5CV | 6CV | FRAN | ETRA | CP1 | CP2 | CP3 | CP4 |
| -4.39 | 1.53 | 2.85 | 6.49 | -6.49 | -8.60 | 4.24 | 0.52 | 3.84 |

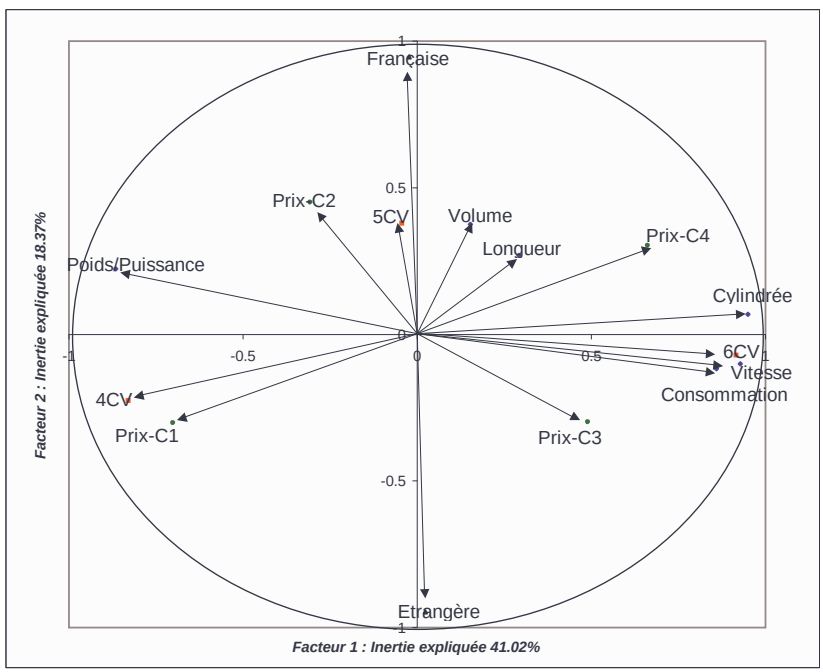
Le tableau 2 présente les moyennes ayant servi à la transformation des indicatrices des trois variables qualitatives en variables quantitatives, étape préalable à l'application de l'ACP normée sur l'ensemble des variables. Bien que l'inertie totale soit égale au nombre de variables $p + q = 15$ (ACP centrée-réduite), elle est résumée et synthétisée par $p + q - m = 12$ facteurs dans le tableau 3.

**TAB. 3.** *Valeurs propres issues de l'ACPM*

| Axe | 1 | 2 | 3 | 4 | 5 | 6 | 7 | ... | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|
| Valeur propre | 6.154 | 2.756 | 2.349 | 1.210 | 0.937 | 0.808 | 0.308 | ... | 0.072 | 0.023 |
| % Inertie | 41.02 | 18.37 | 15.66 | 8.06 | 6.25 | 5.39 | 2.05 | ... | 0.48 | 0.15 |
| % Cumulé | 41.02 | 59.40 | 75.06 | 83.12 | 89.37 | 94.75 | 96.81 | ... | 99.85 | 100.00 |

Outre les résultats graphiques et numériques classiques d'aide à l'interprétation d'une ACP, le tableau 4 donne, les carrés des corrélations linéaires des variables quantitatives et les rapports de corrélation des variables qualitatives avec les premières composantes principales. Il permet de juger de la qualité de la représentation de l'ensemble des variables mixtes sur les plans principaux de l'ACPM (cf. Note).

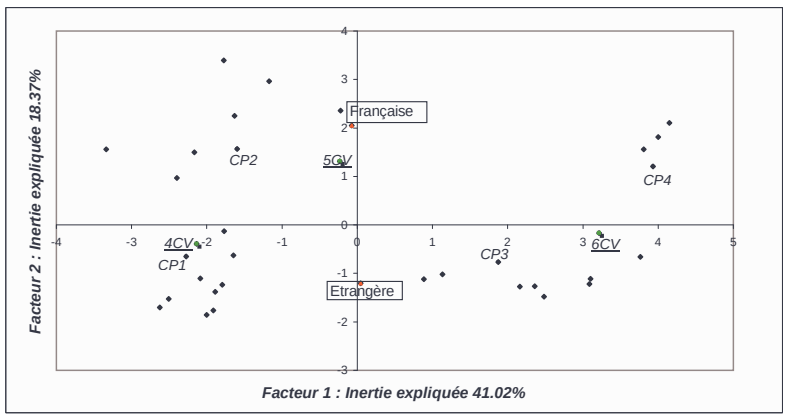
**FIG. 1.** *Représentation des variables mixtes sur le premier plan principal*



**TAB. 4.** *Critère : Qualité de la représentation des variables mixtes*

| Axe | 1 | 2 | 3 | 4 | 5 | 6 | ... | 11 | 12 | Somme |
|---|---|---|---|---|---|---|---|---|---|---|
| CONS | **.738** | .014 | .120 | .001 | .028 | .001 | ... | .029 | .000 | 1.00 |
| CYLI | **.903** | .005 | .004 | .001 | .005 | .002 | ... | .035 | .000 | 1.00 |
| VITE | **.862** | .010 | .011 | .051 | .010 | .002 | ... | .001 | .014 | 1.00 |
| VOLU | .023 | .141 | .227 | **.382** | .078 | .076 | ... | .001 | .000 | 1.00 |
| RP/P | **.748** | .050 | .000 | .055 | .028 | .004 | ... | .000 | .004 | 1.00 |
| LONG | .085 | .073 | .112 | .005 | .256 | .456 | ... | .000 | .000 | 1.00 |
| P.FISCALE | **.918** | .146 | **.628** | .029 | .163 | .007 | ... | .001 | .000 | 2.00 |
| MARQUE | .001 | **.895** | .038 | .009 | .020 | .006 | ... | .001 | .000 | 1.00 |
| C.PRIX | **.929** | .364 | **.705** | **.498** | .219 | .186 | ... | .001 | .003 | 3.00 |
| Somme | 5.206 | 1.697 | 1.844 | 1.032 | .808 | .741 | ... | .070 | .022 | 12.00 |

**FIG. 2.** *Représentation des individus et des modalités (centres de gravité) sur le premier plan principal*




## 4. Conclusion

L'ACPM proposée semble bien prendre en compte l'équilibre des structures des deux types de variables : les corrélations entre les variables quantitatives, les associations entre les modalités des variables qualitatives ainsi que leurs rapports de corrélation. Les résultats ainsi obtenus sont identiques à ceux de l'AFDM. Enfin, il serait intéressant de voir ce que donnerait l'analyse canonique généralisée avec les $p + m$ groupes de variables mixtes.

## 5. Bibliographie


[ABD 05] ABDESSELAM R., Dissymmetrical Multivariate ANalysis Of VAriance, *Classification and Data Analysis*, In Book of Short Papers, Group of the Italian Statistical Society, Editors S. Zani and A. Cerioli, Italy, 2005, p. 189–192.

[ESC 79] ESCOFIER B., PAGÈS J., Traitement simultané de variables quantitatives et qualitatives en analyse factorielle, *Cahier de l'analyse des données*, vol. 4(2), 1979, p. 137–146.

[LAM 90] LAMBIN J., *La recherche marketing, Analyser - Mesurer - Prévoir*, Edt McGraw-Hill, Paris, 1990.

[PAG 02] PAGÈS J., Analyse factorielle multiple appliquée aux variables qualitatives et aux données mixtes, *Revue de Statistique Appliquée*, vol. L(4), 2002, p. 5–37.

[PAG 04] PAGÈS J., Analyse factorielle de données mixtes, *Revue de Statistique Appliquée*, vol. LII(4), 2004, p. 93–111.

[SAP 90] SAPORTA G., Simultaneous analysis of qualitative and quantitative data, *Atti della XXXV riunione scientifica*, Società italiana di statistica, 1990, p. 63–72.

[TEN 77] TENENHAUS M., Analyse en composantes principales d'un ensemble de variables nominales ou numériques, *Revue de Statistique Appliquée*, vol. XXV(2), 1977, p. 39–56.

[YOU 81] YOUNG F., Quantitative analysis of qualitative data, *Psychometrika*, vol. 46(4), 1981, p. 357–388.

# Une approche de caractérisation des contextes appelants et appelés des liens hypertextes

**Moustafa Al-Hajj, Gilles Verley, Hubert Cardot**

*Université François-Rabelais de Tours*
*Laboratoire d'Informatique (EA 2101),*
*64, Avenue Jean Portalis,*
*37200 TOURS - France*
{*moustafa.al-hajj,gilles.verley,hubert.cardot*}*@univ-tours.fr*

RÉSUMÉ. *Nous nous intéressons à la sémantique des liens hypertextes, en termes d'extraction et d'exploitation, dans le but de faciliter le partage des connaissances sur le Web. Dans cet article, nous nous concentrons sur l'élaboration d'outils d'aide à l'analyse de la sémantique des liens hypertextes, nous proposons une automatisation de la reconnaissance des formes littéraires des contextes appelants des liens et des contextes appelés par des liens.*

MOTS-CLÉS : *analyse sémantique de liens hypertextes, treillis de Galois, réseau de neurones, k-plus proches voisins, arbre de décision, Web sémantique.*

## 1. Introduction

Nous cherchons à faire l'analyse sémantique des liens hypertextes. Pour ce faire, nous avons construit notre propre corpus avec, comme domaine, les biographies d'hommes célèbres. Pour effectuer l'analyse sémantique manuelle d'un lien hypertexte, nous proposons une méthode qui consiste à faire l'analyse sémantique des deux contextes, contexte appelant du lien et contexte appelé par le lien, et à trouver la relation sémantique entre le contexte appelant et le contexte appelé.

L'analyse sémantique des deux contextes, contexte appelant du lien et contexte appelé par le lien, consiste à les décrire dans une phrase composée de trois parties :

- La première pour dire qu'il s'agit d'un contexte du lien ou d'un contexte appelé par le lien.
- La deuxième pour décrire la *forme littéraire* du contexte - appelant ou appelé - qu'on analyse.
- La troisième pour décrire, par quelques mots clés reliés naturellement, le contexte appelant (resp. appelé) en cours d'analyse.

La relation sémantique entre le contexte appelant et le contexte appelé consiste à relier un ou plusieurs mots clés du contexte appelant à un ou plusieurs mots clés du contexte appelé.

Dans cet article, nous nous intéressons à l'extraction de la deuxième partie de la sémantique des deux contextes, contextes appelant du lien et contexte appelé par le lien, à savoir la caractérisation des *formes littéraires* des contextes, contextes appelants des liens et contextes appelés par les liens.

On nomme "contexte appelant d'un lien" l'ensemble minimal de textes, caractères et objets, autour du lien qui constituent une seule idée, concept ou sujet. De même, on nomme "contexte appelé par un lien" l'ensemble minimal de textes, caractères et objets de la page ciblée par le lien qui constituent un sujet en rapport avec le "contexte appelant du lien".

Pour des raisons de simplicité et d'automatisation, nous considérons le contexte appelant d'un lien comme la partie de la page comprise entre la première balise "$a\ name''$ qui précède le lien et la première balise "$a\ name''$ qui suit le lien. De même, nous considérons le contexte appelé par le lien comme la partie de la page cible du lien comprise entre le début de la cible du lien et la première balise "$a\ name''$ qui suit.

Dans la section 2 nous définissons les formes littéraires, les paramètres des contextes sont définis en section 3, le choix de la base de contextes pour l'expérimentation sera présenté en section 4, les sections 5, 6 et 7 sont consacrés aux essais de classement par des outils de reconnaissance de formes, on termine par une conclusion en section 8.

## 2. Classes des contextes des liens

Dans cet article, nous nous intéressons à la caractérisation des *formes littéraires* des contextes, nous nous sommes inspirés des travaux de [PAPY 03] pour définir nos classes, nous en avons retenu quelques classes et en avons rajouté d'autres spécifiques au domaine des biographies d'hommes célèbres. Après une observation des *formes littéraires* des différents contextes de notre corpus, nous avons opté pour les classes suivantes :

- **Classe *sommaire*** : Le contenu du contexte est un résumé qui comporte les titres des parties des sites, c'est la même chose que la "page carrefour interne" définie par [PAPY 03]. On les reconnaîtra principalement grâce à l'adjacence des liens.
- **Classe *illustration graphique*** : Le contenu du contexte est une illustration graphique par une image, c'est la même chose que la "page informative avec texte illustré" défini par [PAPY 03]. On les reconnaîtra principalement grâce à la présence d'images de taille importante dans le contexte.
- **Classe *récit*** : Le contenu du contextes est en majorité du texte, on les reconnaîtra principalement grâce à la présence de texte en grand quantité dans le contexte.
- **Classe *citation*** : Le contenu du contexte est un texte qui fait référence directe à une oeuvre dans sa totalité ou en partie. On les reconnaîtra principalement grâce à la présence de texte en quantité moyenne et sans liens hypertextes.
- **Classe *liste*** : Le contenu du contexte est une suite d'articles inscrits les uns à la suite des autres. On les reconnaîtra principalement grâce à la présence des puces ou numéros aux débuts des articles.

## 3. Extraits des données

En partant des caractéristiques citées auparavant, il est possible d'établir le profil d'un contexte en constituant un vecteur d'informations. Le profil est construit par une analyse et un traitement statistique de balises $HTML$. Les paramètres les plus significatifs obtenus à partir de notre échantillon documentaire initial sont : *nbHref* : nombre de liens ; *nbImg* : nombre d'images ; *TGimg* : taille de la plus grande image ; *SMoyImg* : surface moyenne des images ; *nbMot* : nombre de mots hors balise ; *nbLEH* : nombre de lignes entre balises "a href" ; *nbLigne* : nombre de lignes hors balise ; *nbBListe* : nombre de balises qui définissent des listes et/ou listes avec puces et/ou les énumérations ; *nbBPg* : nombre des balises qui définissent les paragraphes ; *nbBSLigne* : nombre de balises de saut de lignes ; *cit* : prend 1 si des mots tels que "citation" figurent en balise 'méta name' et 0 sinon ; *def* : prend 1 si des mots tels que "définition" figurent en balise méta name et 0 sinon ; *desc* : prend 1 si des mots tels que "description" figurent en balise 'méta name' et 0 sinon ; *sommaire* : prend 1 si des mots tels que "sommaire, résumé" figurent en balise méta name et 0 sinon.

L'agent Web recueille les indicateurs quantitatifs, et les stocke sous forme d'une matrice, chaque ligne correspond à un contexte et chaque colonne correspond à l'un des paramètres cités précédemment (tableau 1).

| nbHref | nbImg | TGimg | SMoyImg | nbMot | nbLEH | nbLigne | nbBListe | nbBPg | nbBSLigne | cit | def | Desc | Sommaire |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 10 | 1 | 4628 | 4628 | 2770 | 23 | 239 | 40 | 47 | 0 | 0 | 0 | 0 | 0 |
| 9 | 2 | 0 | 0 | 308 | 0 | 40 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

**TAB. 1.** *Deux lignes de la matrice documents / paramètres*

## 4. Découpage de la base de données

Le corpus de documents sur lequel on travaille est composé de biographies d'hommes célèbres.

Pour la phase d'expérimentation, nous avons choisi 1029 contextes parmi les contextes appelants de liens hypertextes et des contextes appelés par les liens hypertextes de notre corpus. Ensuite nous avons annoté ces contextes manuellement par leurs formes littéraires. A partir de cet ensemble de contextes, nous avons tiré au hasard 852 contextes pour la base d'apprentissage et ce qui reste (177 contextes) sera pour la base de test. Le tableau 2 est un récapitulatif des effectifs des formes littéraires dans les deux bases.

| | Citation | Illustration | Liste | Sommaire | Récit |
|---|---|---|---|---|---|
| Base d'apprentissage | 376 | 13 | 59 | 130 | 274 |
| Base de test | 80 | 3 | 14 | 18 | 62 |
| % de classes dans les 2 bases | 44,3 | 1,6 | 7,1 | 14,4 | 32,6 |

**TAB. 2.** *Effectif de formes littéraires dans les bases*

La classe citation est fortement représentée du fait du domaine d'application de biographies d'hommes célèbres.

Ensuite nous avons mené quatre expériences de classification : avec les treillis de Galois, les k-plus proches voisins, les réseaux de neurones et les arbres de décision.

## 5. Classification avec les treillis de Galois

La construction du treillis de Galois [NGU 02] se fait à partir d'un tableau binaire, et les techniques de classification se basant sur les treillis de Galois traitent avec des objets d'attributs binaires. Donc un passage de paramètres quantitatifs aux paramètres qualitatifs doit être fait. Pour ce passage, nous avons défini pour chaque paramètre quantitatif quatre intervalles, le premier correspond à des valeurs très petites du paramètre, le deuxième à des valeurs petites, le troisième à des valeurs grandes, le quatrième à des valeurs très grandes.

Les attributs binaires de chaque contexte sont obtenus de la façon suivante :
On obtient les premiers attributs par échantillonnage de chaque valeur des paramètres du contexte (§3) dans les quatre intervalles qui lui sont définis. A ces attributs s'ajoutent cinq attributs binaires dont chacun correspond à une de nos classes et prend 1 si le contexte est de la classe qui correspond à l'attribut et 0 sinon. Le problème de classification d'un nouveau contexte revient alors à lui inférer un attribut de classe.

Nous avons utilisé les deux techniques de classification se basant sur le treillis de Galois que nous avons déjà utilisées dans [HAJ 03] : "Validation Globale" et "Validation Locale". L'application de la méthode "Validation Globale" sur les 177 contextes de la base de test a permis de classer 108 contextes et ils sont tous correctement classés, et l'application de la méthode "Validation Locale" sur l'ensemble de test a permis de classer 154 contextes dont 139 sont correctement classés (tableau 3).

| | | Total | Citation | Illustration | Liste | Sommaire | Récit |
|---|---|---|---|---|---|---|---|
| | Effectifs | 177 | 80 | 3 | 14 | 18 | 62 |
| Valiadtion Globale | Classés | 108 | 57 | 1 | 3 | 12 | 35 |
| | Correctement classés | 108 | 57 | 1 | 3 | 12 | 35 |
| Validation Locale | Classés | 154 | 70 | 2 | 8 | 18 | 56 |
| | Correctement classés | 139 | 67 | 2 | 6 | 13 | 51 |

**TAB. 3.** *Résultats obtenus avec les treillis de Galois*

## 6. Classification avec les k-plus proches voisins et les arbres de décision

Nous avons appliqué les deux méthodes de reconnaissance de formes $k-ppv$ [FIX 51] et les arbres de décisions (C4.5 [QUI 93]) pour classer les contextes de la base de test, les contextes des deux bases d'apprentissage et de test étant représentés par leurs valeurs de paramètres quantitatifs (cf §3).

Avec les valeurs suivantes du paramètre $k$ de la méthode $k-ppv$ $\{5, 10, 15, 20, 25, 30, 40\}$, nous avons obtenu un meilleur classement pour $k = 20$. Le nombre de correctement classés avec le $k-ppv$ est de 83, et celui avec les arbres de décision est de 73. Le tableau 4 récapitule les résultats obtenus par les deux méthodes de reconnaissance.

| | Total | Citation | Illustration | Liste | Sommaire | Récit |
|---|---|---|---|---|---|---|
| Effectifs | 177 | 80 | 3 | 14 | 18 | 62 |
| Correctement Classés par le $k-ppv$, $k = 20$ | 83 | 62 | 0 | 0 | 0 | 21 |
| Correctement Classés par les arbres de décisons | 73 | 60 | 0 | 1 | 1 | 11 |

**TAB. 4.** *Résultats obtenus avec les k-ppv pour k = 20 et avec les arbres de décision*

## 7. Classification avec les réseaux de neurones

Nous avons aussi appliqué un réseau de neurones [BON 98] pour classer les contextes de la base de test, étant donnés les contextes des deux bases représentés par leurs valeurs de paramètres quantitatifs (cf §3).

Nous avons mené plusieurs expériences avec les réseaux de neurones de type :

- Réseaux récurrents contenant dix neurones d'entrée, sept neurones de sortie, un neurone de biais et une couche cachée entièrement récurrente composée de neurones avec des fonctions de transfert tangente hyperbolique et avec l'algorithme d'apprentissage $BPTT$ (BackPropagation Through Time)[RUM 86].
- Réseaux à couches contenant le même nombre de neurones d'entrée et de sortie, un neurone de biais et une couche cachée composée de neurones avec des fonctions de transfert sigmoïde et avec l'algorithme d'apprentissage $BP$(BackPropagation)[RUM 86]. Avec les deux types, nous avons varié le nombre de neurones en couche cachée entre quatre, six, douze, vingt-quatre.

Le meilleur résultat est obtenu avec le réseau à couches de six neurones dans la couche cachée, avec ce réseau, 107 contextes, parmi les 177 contextes de la base de test, ont été correctement classés (tableau 5).

| | Total | Citation | Illustration | Liste | Sommaire | Récit |
|---|---|---|---|---|---|---|
| Effectifs | 177 | 80 | 3 | 14 | 18 | 62 |
| Correctement Classés | 107 | 76 | 1 | 4 | 0 | 26 |

**TAB. 5.** *Résultats obtenus avec le réseau de neurones*

## 8. Conclusion

Ce travail se situe dans un projet plus vaste d'analyse de la sémantique de liens hypertextes [VER 00]. Nous avons présenté une expérience d'extraction, par des outils de reconnaissance de formes, de la partie de la sémantique qui correspond aux formes littéraires des contextes appelants des liens et des contextes appelés par des liens. Une autre expérience est en cours concerne le changement des supports informatiques des contextes des liens.

## 9. Bibliographie


[PAPY 03] PAPY F., BOUNAI N., Navigation et recherche par catégorisation floue des pages HTML, *Actes des JET'2003*, 2003.

[NGU 02] NGUIFO M., NJIWOUA, Treillis de concepts et classification supervisée : un état de l'art, rapport, 2002, CRIL rapport de recherche.

[HAJ 03] AL-HAJJ M., BERTET K., GAY J., OGIER J. -M., Aide à la reconnaissance d'objets détériorés avec un treillis de Galois, *In Atelier Treillis, AFIA 2003, Laval, France*, Juin 2003.

[FIX 51] FIX E., HODGES J. L., Dicriminatory analysis, nonparametric discrimination : Consistency properties, rapport, 1951, USAF School of Aviation Medecine, Randolph Field, TX.

[QUI 93] QUINLAN J. R. , *C4.5 : Programs for Machine Learning*, Morgan Kaufmann, 1993.

[BON 98] BONÉ R., CRUCIANU M., MAKRIS P., ASSELIN DE BEAUVILLE J. -P., A Web Oriented Recurrent Neural Network Simulator, *International conference on neural information processing, Kitakyushu, Jabon*, 1998, p. 97-100.

[RUM 86] RUMELHART D. E., HINTON G. E., WILLIAMS, R. J., Learning internal representations by error propagation, *In Parallel Distributed Processing : Explorations in the Microstructure of Cognition, D.E. Rumelhart, J. McClelland (Eds.) MIT Press, Cambridge*, 1986, p. 318-362.

[VER 00] VERLEY G., ROUSSELLE J. J., An evolved link-specification language for creating and sharing documents on the web, *CRIS 2000 Current Research Information Systems, Helsinky*, 25-27 mai 2000.

# Deux distances locales pour graphes pondérés

**Jean-Baptiste Angelelli, Alain Guénoche**

*IML 163 Av. de Luminy, 13288 Marseille cedex 9*
*guenoche@iml.univ-mrs.fr*

RÉSUMÉ. *Nous étendons la distance de Czekanovski-Dice aux graphes pondérés pour partitioner les sommets des graphes d'interaction directe entre protéines. La première distance correspond au cas où les pondérations sont des probabilités d'interaction et la seconde au cas où ce sont des intensités. Si les poids sont tous égaux à 1, elles redonnent les valeurs de Czekanovski-Dice.*

MOTS-CLÉS : *Graph distances, graph partitionning*

## 1. Introduction

Nous utilisons la distance de Czekanovski-Dice (Dice, 1945) sur les graphes d'interaction entre protéine pour appliquer des méthodes de partitionnement (Guénoche, 2004). Ce n'est pas une *vraie* distance mais un indice de dissimilarité ; deux sommets adjacents qui n'ont que des sommets adjacents communs sont à distance 0 et elle ne vérifie pas l'inégalité triangulaire (Fichet et Le Calvé 1984). Nous continuerons néanmoins (comme tout le monde) à la qualifier de distance. Elle prend en compte la simple présence ou absence d'arêtes entre deux sommets. Les graphes dont nous disposons maintenant sont pondérés et nous étendons la distance de Czekanovski-Dice à ce type de graphe.

Soit $X$ un ensemble de $n$ sommets, $E$ l'ensemble des $m$ arêtes et $\Gamma = (X, E)$ le graphe correspondant. Nous admettrons qu'il est connexe sinon ses différentes composantes connexes seront traitées séparément. Pour toute partie $Y$ de $X$, soit $\Gamma(Y)$ l'ensemble des sommets hors de $Y$ qui sont adjacents à $Y$,

$$\Gamma(Y) = \{x \in X \setminus Y \text{ tels que } \exists y \in Y, (x, y) \in E\},$$

et $\overline{\Gamma}(Y) = Y \bigcup \Gamma(Y)$. Le voisinage de $x$ est donc $\Gamma(x)$. Le degré de $x$ est noté $Dg(x) = |\Gamma(x)|$ et $\delta$ désigne le degré maximum du graphe. Soit $E(Y)$ l'ensemble des arêtes internes à la classe $Y \subset X$ :

$$E(Y) = \{(x, y) \in E \text{ telles que } x \in Y \text{ et } y \in Y\}.$$

## 2. La distance de Dice

La distance de Czekanovski-Dice entre sommets est définie par

$$D(x, y) = \frac{|\Delta(\overline{\Gamma}(x), \overline{\Gamma}(y))|}{|\overline{\Gamma}(x)| + |\overline{\Gamma}(y)|},$$

dans laquelle $\Delta$ est la différence symétrique entre deux ensembles. On remarquera que ce sont les $\overline{\Gamma}$ qui sont utilisés, ce qui revient à ajouter une boucle en chaque sommet du graphe, puisque $x$ est adjacent à $x$. Nous utilisons cette distance pour plusieurs raisons :

– elle s'est avérée très efficace dans les méthodes de partitionnement de graphes (Kuntz 1992, Jouve et al. 2002, Guénoche 2005), nettement plus que le nombre d'arêtes d'un plus court chemin entre $x$ et $y$ ;

– C'est une mesure *locale*, puisque $D(x, y)$ est calculable à partir des seuls sommets adjacents à $x$ ou à $y$ ;
– Toute paire de sommets séparés par plus de deux arêtes a une distance maximum égale à 1 ;
– De ce fait, cette distance est calculable en $O(n\delta^3)$ comme nous le montrerons dans la suite.

La valeur de $D(x, y)$ est le quotient de deux quantités, la part spécifique à $x$ et à $y$, notée $P_{spe}(x, y)$, et la part totale, notée $P_{tot}(x, y)$. Si $x$ et $y$ ne sont pas adjacents, ces deux sommets sont comptabilisés dans les deux parts. Par contre, s'ils sont adjacents, ils n'interviennent que dans la part totale. Ainsi, cette distance donne deux valeurs nettement différentes, suivant que $x$ et $y$ sont adjacents ou non. On notera également que l'intersection $\Gamma(x) \bigcap \Gamma(y)$ est comptée deux fois dans la part totale.

La distance de Dice comptabilise des sommets, adjacents à $x$ ou à $y$. Si l'on veut utiliser des pondérations, il faut considérer les arêtes et non plus les sommets.

## 3. Deux distances pondérées

Dans les graphes d'interaction, une arête $(x, y)$ signifie que les protéines $x$ et $y$ ont une interaction directe, c'est à dire qu'elles sont en contact à un moment donné, et pas seulement qu'elles appartiennent à un *complexe protéique*, c'est à dire un ensemble de protéines réalisant une fonction cellulaire particulière. Cette information peut être obtenue de diverses manières, en particulier à l'aide d'expériences à grande échelle, comme les cribles double hybride, qui donnent de nombreux faux positifs. Les biologistes s'ingénient à verifier et confirmer l'existence de ces interactions à l'aide d'autres expériences ou de références dans la littérature. Ainsi, le poids attribué à une arête peut être considéré comme une probabilité associée à cette interaction.

Mais une autre interprétation est possible ; ce peut être une mesure de l'intensité de l'interaction, suivant le type de cellules considéré, ou suivant qu'elle est essentielle ou secondaire à la réalisation de la fonction, qu'elle la rende plus facile ou qu'elle l'accélère. C'est également une interprétation courante en économie, où la pondération peut quantifier des volumes d'échanges entre acteurs. Ce type de pondération peut aussi être obtenu à partir d'une distance, en donnant des poids forts aux petites valeurs.

Dans un cas comme dans l'autre, la fonction de poids, notée $p : X \times X \longrightarrow [0, 1]$, est d'autant plus grande que l'interaction est probable ou forte. La valeur 0 correspond à l'absence d'arête et la valeur 1 à une information sûre ou un contact essentiel.

### 3.1. *Les pondérations sont des probabilités*

Si les pondérations sont des probabilités, la distance notée $D_p$ entre toute paire $(x, y)$ est liée au poids de l'éventuelle arête $(x, y)$. C'est une somme pondérée des distances correspondant aux graphes $\Gamma_-(x, y)$ dans lesquels l'arête (x,y) n'existe pas (elle a un poids 0) et aux graphes $\Gamma_+(x, y)$ dans lesquelles elle existe sûrement (avec un poids 1).

$$D_p(x, y) = (1 - p(x, y)) \times \Gamma_-(x, y) + p(x, y) \times \Gamma_+(x, y) \quad (1)$$

Mais les graphes $\Gamma_+$ et $\Gamma_-$ sont eux même fonction des probabilités des arêtes de $E(Y)$, avec $Y = \Gamma(x) \cup \Gamma(y)$. Le calcul de $D_p(x, y)$ devrait se faire par sommation sur les parties de $Y$, chacune étant pondérée par le produit des poids des arêtes présentes dans cette partie. Pour des raisons de complexité (énumération des parties d'un ensemble), nous avons retenu une formule plus simple.

1. Dans le cas des configurations $\Gamma_-(x, y)$ la part spécifique à $x$ et à $y$ est composée :
– des poids des arêtes correspondant aux sommets rattachés exclusivement à $x$ ou à $y$,
– pour un sommet $s$ rattaché à $x$ et à $y$, de la différence de poids $|p(x, s) - p(y, s)|$, qui correspond aux situations où $s$ n'est plus rattaché qu'à l'un des sommets,
– des deux boucles en $x$ et en $y$, implicitement de poids 1, nécessaires pour rester compatible avec la formule de Czekanovski-Dice.

Et donc la partie spécifique est égale à $P_{spe}(x,y) = 2 + \sum_{s\in Y} |p(x,s) - p(y,s)|$. Pour la partie totale, il faut compter $P_{tot}(x,y) = 2 - 2 \times p(x,y) + \sum_{s\in Y} p(x,s) + p(y,s)$, la valeur 2 correspondant encore aux boucles et les $2 \times p(x,y)$ à l'arête $(x,y)$ comptée 2 fois dans la somme bien qu'absente de ces configurations.

2. Dans le cas des configurations $\Gamma_+(x,y)$ la part spécifique à $x$ et à $y$ est égale à $P_{spe}(x,y) = \sum_{s\in Y} |p(x,s) - p(y,s)|$ ; les boucles n'y sont plus, puisque $x$ et $y$ sont tous les deux adjacents à $x$ et $y$. Et la partie totale est $P_{tot}(x,y) = \sum_{s\in Y} p(x,s) + p(y,s) + 2 + 2 \times (1 - p(x,y))$, le dernier terme correspondant au poids de l'arête $(x,y)$ égal à 1 dans les graphes $\Gamma_+$.

Donc pour reprendre la définition (1), nous noterons

$$S(x,y) = \sum_{s\in Y} |p(x,s) - p(y,s)| \text{ et } T(x,y) = \sum_{s\in Y} p(x,s) + p(y,s)$$

pour aboutir à la distance pondérée par des probabilités :

$$D_p(x,y) = (1 - p(x,y)) \times \frac{S(x,y) + 2}{T(x,y) + 2 - 2p(x,y)} + p(x,y) \times \frac{S(x,y)}{T(x,y) + 4 - 2p(x,y)}.$$

### 3.2. *Les pondérations sont des intensités*

Si les pondérations sont considérées comme des intensités d'interaction, les interactions spécifiques à $x$ ou à $y$ tendent à les éloigner et les interacteurs communs à les rapprocher. Ainsi, les liens entre $x$ et $y$ peuvent être rangés en deux catégories :

- les forces d'attraction correspondant à l'arête éventuelle $(x,y)$ et pour tout sommet $s$ adjacent à $x$ et à $y$ la somme des poids des arêtes $(x,s)$ et $(y,s)$ ;
- les forces de répulsion correspondant aux sommets rattachés à un seul des $x$ ou $y$, ainsi que les boucles, dans le cas où les sommets $x$ et $y$ ne sont pas adjacents.

Notons $R(x,y)$ la partie répulsive et $A(x,y)$ la partie attractive correspondant à la paire $(x,y)$. On a

$$R(x,y) = \sum_{s\in\Gamma(x)\setminus\Gamma(y)} p(x,s) + \sum_{s\in\Gamma(y)\setminus\Gamma(x)} p(y,s),$$

$$A(x,y) = \sum_{s\in\Gamma(x)\cap\Gamma(y)} p(x,s) + \sum_{s\in\Gamma(y)\cap\Gamma(x)} p(y,s).$$

$R(x,y)$ correspond à la différence symétrique, donc au numérateur dans la formule de Dice et $A(x,y)$ à l'union $\Gamma(x) \bigcup \Gamma(y)$. On remarquera que si $(x,y) \in E$, le poids de l'arête $(x,y)$ bien est compté 2 fois dans $A(x,y)$.

La distance $D_i$ basée sur les intensités d'interaction est donc la version quantifiée de la distance de Czekanovski-Dice ; elle est définie par :

$$D_i(x,y) = \frac{R(x,y) + 2}{R(x,y) + A(x,y) + 2} \text{ si } (x,y) \notin E$$

$$D_i(x,y) = \frac{R(x,y)}{R(x,y) + A(x,y) + 2} \text{ si } (x,y) \in E$$

On notera que si les sommets $x$ et $y$ sont séparés par plus de deux arêtes, c'est la première formule qui s'applique et la distance est toujours égale à 1, quel que soit ce nombre d'arêtes.

## 4. Quelques propriétés

**Proposition 1** *Si les arêtes ont toutes un poids égal à 1, les valeurs obtenues par $D_p$ et $D_i$ sont identiques à la distance de Czekanovski-Dice.*

*Preuve :*

– Quand les pondérations sont des probabilités, la formule pour $D_p$ se réduit à :

$$\frac{S(x,y)+2}{T(x,y)+2} \text{ quand } (x,y) \notin E \text{ et } \frac{S(x,y)}{T(x,y)+2} \text{ quand } (x,y) \in E,$$

formules dans lesquelles $S(x,y)$ est le nombre d'arêtes spécifiques à $x$ ou à $y$ et $T(x,y)$ le nombre d'arêtes dans $\Gamma(x)$ plus celle dans $\Gamma(y)$, ce qui correspond à la formule de Czekanovski-Dice puisque l'arête $(x,y)$ est bien comptée deux fois ;

– Quand les pondérations sont des intensités,

$$R(x,y) = S(x,y) \text{ et } R(x,y) + A(x,y) = T(x,y)$$

ce qui redonne les mêmes formules que précédemment.

**Proposition 2** *Ces distances pondérées sont calculables en $O(n\delta^3)$.*

*Preuve :* Pour évaluer la distance de Czekanovski-Dice de façon efficace, il faut coder le graphe par ses listes d'adjacence. Ainsi pour chaque sommet $x$, on doit calculer au plus $\delta^2$ valeurs de distance. Pour déterminer les parties spécifiques et totales dans l'évaluation de $D(x,y)$, on examine au plus $2\delta$ sommets adjacents à $x$ ou à $y$. Donc toutes les valeurs de distance strictement plus petites que 1 sont calculées en $O(n\delta^3)$. Pour les graphes pondérés, on adjoindra aux listes d'adjacence les listes de poids des arêtes, de façon à sommer ces poids dans l'une ou l'autre part. La complexité du calcul est donc identique.

A titre d'exemple, précisons que la distance se calcule en quelques secondes pour un graphe peu dense à 8000 sommets, sur un ordinateur de bureau standard.

La suite de l'exposé portera sur des procédures de simulation qui tendent à prouver que la prise en compte de pondérations permet de mieux retrouver des classes initiales, si les arêtes de poids fort lient des sommets appartenant à une même classe ou si les arêtes de poids faibles connectent des sommets de classes distinctes. Une illustration sur le graphe des pays européens ayant une frontière commune, pondéré par le trafic routier, permettra de mesurer l'influence des pondérations dans la construction de classes.

Nous montrerons ainsi que la cohérence des classes est renforcée par l'usage de pondérations rationnelles.

**Bibliographie**


L.R. Dice (1945) Measures of the amount of ecologic association between species, *Ecology*, 26, 297-302.

B. Fichet, G. Le Calvé (1984) Structure géométrique des principaux indices de dissimilarité sur signes de présence-absence, *Statistique et Analyse de Données*, 3, 11-44.

A. Guénoche (2004) Clustering by vertex density in a graph, Proceedings of IFCS congress *Classification, Clustering and Data Mining Applications*, D. Banks et al. (Eds.), Springer, 15-24.

A. Guénoche. (2005) Comparison of algorithms in graph partitioning, ALIO/EURO conference on Combinatorial Optimization, Paris, submitted.

B. Jouve, P. Kuntz, F. Velin (2002) Extraction de structure macroscopiques dans des grands graphes par une approche spectrale, *Extraction des Connaissances et Apprentissage*, 1, 4, 173-184.

P. Kuntz (1992) *Représentation euclidienne d'un graphe abstrait en vue de sa segmentation*, Thèse de l'Ecole des Hautes Etudes en Sciences Sociales, Paris.

# Analyse statistique des puces à ADN :

***Généralisation de la méthode dchip***

**Mounir Aout**

*STID Lisieux – IUT de Caen*
*11 Bd Jules Ferry,*
*14100 Lisieux*
*{m.aout}@lisieux.iutcaen.unicaen.fr*

RÉSUMÉ. *Les puces à ADN sont devenues un outil incontournable pour l'analyse de l'expression de milliers de gènes simultanément. Cependant, des défis importants concernant l'estimation de l'expression des gènes restent à relever. Dans ce papier, on propose d'estimer de telles expressions. Notre méthode généralise celle de Li et Wong, plus précisément, on montre comment le modèle de Li et Wong peut être exploité dans sa généralité pour obtenir de nouvelles estimations. Finalement, on compare les modèles obtenus.*

MOTS-CLÉS : *Puces à ADN, Expression des gènes, Maximum de Vraisemblance.*

## 1. Introduction

Les puces à ADN sont un outil considérable qui permet de visualiser l'expression de milliers de gènes simultanément. Cette technologie récente trouve de nombreuses applications dans des domaines aussi divers que la médecine, la biologie fondamentale ou la microbiologie. Ainsi, l'utilisation des techniques d'analyse du transcriptome ont permis des avancées considérables dans la génomique fonctionnelle, la recherche pharmaceutique, le contrôle des produits de l'industrie agro-alimentaire etc... Notre intérêt se porte particulièrement sur les puces à haute densité, commercialisées par la société Affymetrix, qui ont une structure très particulière. Pour chaque gène, une série de 10 à 20 sondes, réparties sur toute la séquence du gène, est représentée sur la lame. A chacune de ces sondes $PM$ (Perfect Match) est associée une sonde $MM$ (MisMatch) dont la séquence est identique à la séquence $PM$ mais avec une mutation ponctuelle située en position centrale. La sonde $MM$ permet de quantifier la part du signal non-spécifique (bruit de fond) associé à la sonde $PM$. Le calcul du niveau d'expression d'un gène est relativement complexe, et il existe plusieurs méthodes pour l'effectuer. Cependant, on se limitera aux modèles Affymetrix *MAS4.0* [AFF 99] ou *MAS5.0* [AFF 01], et la méthode de Li et Wong [LI 01b] connue sous le nom 'dchip'[1]. *MAS4.0* considère une moyenne pondérée des différences ($PM_{ij} - MM_{ij}$) de chaque paire de sondes associées à ce gène, où $j = 1, ..., J$ est le nombre de paires de sondes. Cette différence moyenne est basée sur le modèle statistique suivant, pour chaque gène $i$ : $PM_{ij} - MM_{ij} = \theta_i + \epsilon_{ij}, j = 1...J$. L'indice de l'expression est représenté par $\theta_i$. Les inconvénients de cette méthode ne seront pas évoqués ici, ni ceux de *MAS5.0* [IRI 03b]. [LI 01b] propose d'utiliser le modèle suivant : Pour un gène donné sur la puce $i = 1, ...I$ : $PM_{ij} - MM_{ij} = \theta_i\phi_j + \epsilon_{ij}, i = 1, ...I, j = 1, ...J, \epsilon \cong N(0, \sigma^2)$. Ce modèle sera étudié plus en détail dans la suite de ce travail.

1. http ://www.dchip.org

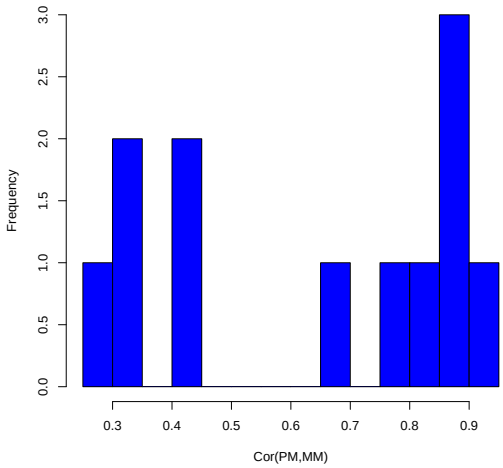



**FIG. 1.** *Correlation entre PM et MM.*

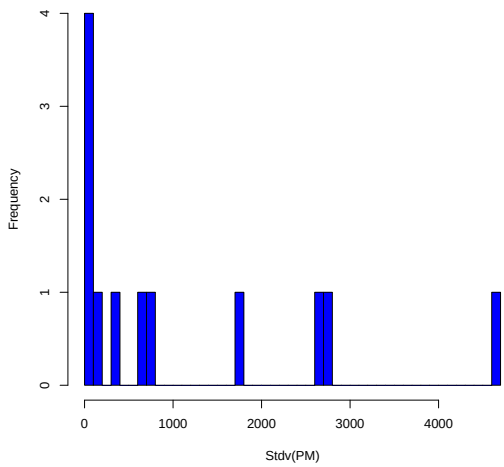



**FIG. 2.** *Variance de PM.*

## 2. Le modèle de Li et Wong généralisé

Selon Li et Wong, les intensités des $PM$ et $MM$ sont données par :

$$PM_{ij} = \nu_{ij} + \theta_i \alpha_j + \theta_i \phi_j + \epsilon_{ij}^P \quad [1]$$

$$MM_{ij} = \nu_{ij} + \theta_i \alpha_j + \epsilon_{ij}^M \quad [2]$$

où $I$ est le nombre d'échantillons (puces) et $J$ le nombre de paires de sondes du gène étudié. $\theta$ est l'indice d'expression, $\nu$ est une intensité due à l'hybridation non-spécifique, $\alpha$ est le taux de croissance des $MM$ et $\phi$ est le taux additionnel de croissance des $PM$. Toutes ces quantités sont positives. Afin d'avoir un modèle identifiable, on ajoute la condition $\sum_j \phi_j^2 = J$.
Ce modèle n'a été utilisé que dans un cadre réduit que nous appellerons *RLW* :

$$PM_{ij} - MM_{ij} = \theta_i \phi_j + \epsilon_{ij}, \epsilon \cong N(0, \sigma^2)$$

[LEM 02] ont utilisé les équations 1 et 2, mais en supposant que les $PM$ et $MM$ sont indépendantes, ce qui revient à utiliser les distributions marginales. [TAI 04] a récemment introduit un modèle dans lequel il a supposé que les erreurs sont corrélées mais avec une corrélation constante. En général, cette hypothèse ne concorde pas avec les observations comme le montre les figures 1 et 2. Pour plus de détails : [AOU 07].
Afin de tenir compte de variations biologiques et expérimentales, nous proposons d'augmenter ce modèle pour traduire les variations des corrélations empiriques entre $PM$ et $MM$ en fonction des puces ainsi que leurs variances respectives. En effet, les erreurs sont supposées suivre une distribution normale bi-variée :

$$\begin{pmatrix} \epsilon_{ij}^P \\ \epsilon_{ij}^M \end{pmatrix} \cong N\left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_i^2 & \rho_i \sigma_i^2 \\ \rho_i \sigma_i^2 & \sigma_i^2 \end{pmatrix} \right)$$

où $\sigma_i^2$ est la variance et $\rho_i$ est le coefficient de corrélation. Ce modèle sera appelé *FLW*.

En considérant les données $(PM_{ij}, MM_{ij})$, nous pouvons estimer les paramètres avec la méthode de maximum de vraisemblance.
La log-vraisemblance est donnée par :

$$l = \sum_{i,j} log(K_i) - \sum_{i,j} \frac{1}{2\sigma_i^2(1-\rho_i^2)} \left[ X_1^2 - 2\rho_i X_1 X_2 + X_2^2 \right]$$

où $K_i = 1/2\pi\sigma_i^2\sqrt{1-\rho_i^2}$, $X_1 = PM_{ij} - \nu_{ij} - \theta_i\alpha_j - \theta_i\phi_j$ et $X_2 = MM_{ij} - \nu_{ij} - \theta_i\alpha_j$.
On obtient :

$$\hat{\theta_i} = \frac{\sum_j \phi_j \left[PM_{ij} - \rho_i MM_{ij} - (1-\rho_i)\nu_{ij}\right] + (1-\rho_i)\sum_j \alpha_j \left[PM_{ij} + MM_{ij} - 2\nu_{ij}\right]}{\sum_j \phi_j^2 + 2(1-\rho_i)\alpha_j^2 + 2(1-\rho_i)\alpha_j\phi_j}$$

Les autres formules peuvent être obtenues de la même manière. Ces formules doivent être vues comme des étapes d'itérations qui mèneront à une solution finale. Dans ce travail, la résolution explicite de ce système d'equations n'est pas une fin en soi, néanmoins, elles nous sont utiles pour en extraire certaines propriétés.
En supposant les autres paramètres connus, il est facile de voir que $\hat{\theta_i}$ est sans biais de $\theta_i$ puisque $E\left[\hat{\theta_i}\right] = \theta_i$.
Pour la variance, on a :

$$Var(\hat{\theta_i}) = \frac{\sigma_i^2(1-\rho_i^2)}{\sum_j \phi_j^2 + 2(1-\rho_i)\alpha_j^2 + 2(1-\rho_i)\alpha_j\phi_j} \quad [3]$$

## 3. Comparaisons entre *FLW* et *RLW*

La définition du modèle réduit de Li et Wong *RLW* est rappelée ci-dessous :

$$Y_{ij} := PM_{ij} - MM_{ij} = \theta_i\phi_j + \epsilon_{ij}, \sum_j \phi_j^2 = J, \epsilon_{ij} \cong N(0,\sigma^2)$$

L'estimation de l'indice de l'expression $\theta_i$ peut être obtenue en utilisant le maximum de vraisemblance ou la méthode des moindres carrés :

$$\hat{\theta_i} = \frac{\sum_j Y_{ij}\phi_j}{\sum_j \phi_j^2}$$

En supposant que les $\phi$ sont connus, et en se basant sur les hypothèses du modèle *RLW*, on a

$$Var(\hat{\theta_i}) = \frac{\sigma^2}{J}$$

D'autre part, en se basant sur les hypothèses du *FLW*, on peut facilement montrer que

$$Var(\hat{\theta_i}) = \frac{2\sigma_i^2(1-\rho_i)}{\sum_j \phi_j^2} \quad [4]$$

et il est facile de voir que $[3] \leq [4]$.
Ainsi, le $\hat{\theta_i}$ obtenu par *FLW* est sans biais et a une variance plus petite ce qui consolide le choix de notre modèle vis à vis du modèle réduit.

## 4. Le modèle basé sur les PM

Comme il a été discuté dans [LI 01a], les $MM$ montrent une réponse insuffisante au changement de niveau d'expression des gènes. Ce phénomène a généré plusieurs questions quand à l'efficacité de la prise en compte des $MM$ dans le calcul de l'expression des gènes. Le but de cette section est d'étudier un modèle basé uniquement sur les $PM$ et de le comparer au modèle réduit. Le *FLW* modifié devient :

$$PM_{ij} = \nu_j + \theta_i\phi_j + \epsilon_{ij}$$

où $\epsilon_{ij} \cong N(0,\sigma_i^2)$. La même procédure donne :

$$\hat{\phi_j} = \frac{\sum_i \frac{\theta_i}{\sigma_i^2}(PM_{ij} - \nu_{ij})}{\sum_i \frac{\theta_i^2}{\sigma_i^2}}$$

$$\hat{\nu_j} = \frac{\sum_i \frac{1}{\sigma_i^2}(PM_{ij} - \theta_i\phi_j)}{\sum_i \frac{1}{\sigma_i^2}}$$

$$\hat{\theta_i} = \frac{\sum_j \phi_j(PM_{ij} - \nu_j)}{\sum_j \phi_j^2}$$

$$\hat{\sigma_i^2} = \frac{\sum_j (PM_{ij} - \theta_i\phi_j - \nu_j)^2}{J}$$

Pour évaluer ce modèle, nous utilisons les données issues de 'spike-in study' *HGU95A*[2] désignée par Affymetrix. Ces données ont été développées pour valider l'algorithme *MAS5.0*.

## 5. Résultats numériques et conclusions

La résolution du système d'equations ci-dessus est faite en utilisant un algorithme itératif. Nous avons écrit un programme sous l'environement **R** [IHA 96]. En outre, nous avons utilisé l'ensemble des programmes décrits dans [IRI 03a], qui fait partie du projet 'Bioconductor' [3], ainsi que les outils d'évaluation discutés dans [COP 03] via l'interface [4] dont le but est d'évaluer et comparer les différentes méthodes employées pour le calcul d'expression des gènes issues des puces à ADN. Pour plus de détails concernant les critères de comparaisons : [COP 03].
Les résultats sont donnés dans le tableau 1. En comparaison avec le modèle réduit, les valeurs obtenues avec notre modèle se rapprochent de la perfection pour 7 des 10 critères de comparaison dont la médiane des écart-types (1ère entrée du tableau). D'autres comparaisons sont données dans [AOU 07].

| | FLW-PMonly | RLW-PMonly | Perfection |
|---|---|---|---|
| Median SD | 0.066 | 0.132 | 0 |
| Signal detect slope | 0.480 | 0.533 | 1 |
| Signal detect R2 | 0.852 | 0.846 | 1 |
| AUC ($FP < 100$) | 0.783 | 0.674 | 1 |
| AFP, call if $fc > 2$ | 7.331 | 36.907 | 0 |
| ATP, call if $fc > 2$ | 10.728 | 11.427 | 16 |
| IQR | 0.211 | 0.446 | 0 |
| FC =2, AUC ($FP < 100$) | 0.460 | 0.167 | 1 |
| FC=2, AFP, call if $fc > 2$ | 6.821 | 28.642 | 0 |
| FC=2, ATP, call if $fc > 2$ | 1.000 | 1.250 | 16 |

**TAB. 1.** *Résultats de comparaisons entre RLW et FLW basés sur les PM*

### 5.1. *Conclusions*

On a présenté une comparaison entre les modèles réduits et complets de Li et Wong en utilisant l'approche bivariée ou uniquement les $PM$. Pour analyser la différence dans la performance générée par ces deux modèles, on a utilisé des critères théoriques et numériques. Afin de faire un choix, la précision(variance) et l'exactitude(biais) ont été comparées entre les différents modèles. Nos estimations sont sans biais avec une variance inférieure à celle de *RLW*. En outre, l'utilisation des $PM$ uniquement permet d'obtenir des augmentations considérables en comparaison avec le modèle réduit. Finalement, l'emploi d'algorithmes plus pertinents pour résoudre numériquement le système d'équations du modèle *FLW-PMonly* devrait permettre l'obtention de meilleurs résultats.

2. http ://www.affymetrix. com/analysis/downloadcenter2.affx
3. http ://www.bioconductor.org
4. http ://affycomp.biostat.jhsph.edu

## 6. Bibliographie


[AFF 99] AFFYMETRIX, Microarray Suite User Guide, Version 4, 1999.

[AFF 01] AFFYMETRIX, Microarray Suite User Guide, Version 5, 2001.

[AOU 07] AOUT M., Comparisons of gene expression indexes for oligonucleotide arrays, *JDS*, vol. 5-3, 2007.

[COP 03] COPE L., IRIZARRY R., JAFFEE H., WU Z., SPEED T., A benchmark for affymetrix geneChip expression measures, *Bioinformatics*, vol. 20, 2003, p. 323–331.

[IHA 96] IHAKA R., GENTLEMAN R., R : a language for data analysis and graphics, *J. Comput. Graph. Stat*, vol. 5, 1996, p. 299–314.

[IRI 03a] IRIZARRY R., GAUTIER L., COPE L., An R package for analyses of Affymetrix oligonucleotide arrays, *In Parmigiani, G.,Garrett, E.S.,Irizarry, R.A. and Zeger, S.L. (eds)*, Springer, 2003.

[IRI 03b] IRIZARRY R., HOBBS B., COLLIN F., BEAZER-BARCLAY Y., ANTONELLIS K., SCHERF U., SPEED T., Exploration, normalization, and summaries of high density oligonucleotide array probe level data, *Biostatistics*, vol. 4(2), 2003, p. 249–264.

[LEM 02] LEMON W., PALATINI J., KRAHE R., WRIGHT F., Theoretical and experimental comparisons of gene expression indexes for oligonucleotide arrays, *Bioinformatics*, vol. 18(11), 2002, p. 1470–6.

[LI 01a] LI C., WONG W., Model-based analysis of oligonucleotide arrays : model validation, design issues and standard error application, *Genome Biology*, vol. 2(8), 2001, p. research0032.1-0032.11.

[LI 01b] LI C., WONG W., Model based analysis of oligonucleotide arrays :Expression index computation and outliers detection, *PNAS*, vol. 98, 2001, p. 31–36.

[TAI 04] TAIB Z., Statistical analysis of oligonucleotide microarray data, *CRAS*, vol. 237(3), 2004, p. 175–180.

# Booster les réseaux de neurones récurrents pour la prévision multipas

**Mohammad Assaad, Romuald Boné, Hubert Cardot**

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

*RÉSUMÉ. Nous nous intéressons au problème de la prévision multipas par réseaux de neurones. Après une présentation des différentes approches existantes pour aborder ce problème, nous évaluons les performances d'un algorithme de boosting pour réseaux de neurones récurrents (RNRs). Le boosting est un méta-algorithme qui combine un grand nombre de RNRs, chacun d'entre eux étant généré en apprenant sur une version différente de l'ensemble d'apprentissage d'origine, version obtenue en se concentrant sur les exemples difficiles. Nous comparons les résultats obtenus sur trois séries temporelles de référence avec ceux de la littérature.*

*MOTS-CLÉS : Séries temporelles, prévision multipas, réseaux de neurones récurrents, boosting.*

## 1 Introduction

La dégradation des performances des systèmes de prévision de séries temporelles dès lors que l'horizon de prévision augmente est un problème bien connu. L'estimation de la valeur future d'une variable peut être raisonnablement fiable pour une prévision à court terme mais se dégrade lorsqu'on aborde le problème de la prévision à long terme. Toutefois la prévision multipas de séries temporelles est indispensable dans de nombreux domaines d'application, qui s'étendent de la modélisation des phénomènes naturels au contrôle de systèmes dynamiques en passant par la finance, le marketing, la météorologie, etc. La majeure partie de la littérature publiée traite du problème de la prévision de séries temporelles à un pas de temps. En effet, la prévision multipas demeure un problème difficile pour lequel les résultats obtenus par des extensions simples des méthodes développées pour la prévision à un pas sont souvent décevants. Par ailleurs, si beaucoup de méthodes obtiennent des résultats proches sur des problèmes de prévision à un pas, des différences significatives se font jour quand des extensions de ces méthodes sont employées sur des problèmes multipas.

Pour améliorer les performances obtenues, nous pouvons adapter des procédures générales qui ont démontré leur capacité à améliorer les performances de divers modèles de base. Une de ces procédures est connue sous le nom de boosting [SCH 90]. Le gain qu'un modèle apporte par rapport à une estimation aléatoire est amplifié dans l'algorithme de boosting par la construction séquentielle de plusieurs de ces modèles, qui se concentrent progressivement sur les exemples qui s'avèrent difficiles dans l'ensemble d'apprentissage d'origine. Nous utilisons une version du boosting adaptée au traitement des données séquentielles par réseaux de neurones récurrents [ASS 05], qui travaille sur des versions pondérées des données initiales d'apprentissage et combine les résultats par médiane pondérée.

## 2 La prévision multipas par réseaux de neurones

Le plus souvent les caractéristiques du phénomène qui a généré la série sont inconnues. L'approche habituellement adoptée pour estimer les valeurs futures $\hat{x}(t+1)$ repose sur l'utilisation d'une fonction $f$ qui prend comme entrée une fenêtre temporelle constituée d'un nombre fixe $p$ de valeurs passées de la série temporelle : $\hat{x}(t) = f\big(x(t-1), x(t-2), \ldots, x(t-p)\big)$. De par leurs propriétés d'approximation universelle, les réseaux de neurones sont souvent utilisés pour la modélisation des fonctions non-linéaires $f$. Parmi les nombreuses architectures neuronales employées pour la prévision des séries temporelles, nous pouvons mentionner les perceptrons multicouches (PMCs) avec une fenêtre temporelle en entrée [RUM 86], les PMCs avec connexions Finite Impulse Response (FIR) [BAC 92], les réseaux récurrents obtenus en utilisant des connexions Infinite Impulse Response (IIR) dans les PMCs [WAN 94]. Employer une fenêtre temporelle de taille fixe s'avère limité pour beaucoup d'applications : si la fenêtre temporelle est trop étroite, des informations importantes peuvent en être exclues, alors que si la fenêtre est trop large, des entrées inutiles peuvent agir en tant que bruit. Idéalement, pour un problème donné, la taille de la fenêtre temporelle devrait pouvoir s'adapter au contexte. Ceci peut être accompli en utilisant des réseaux de neurones récurrents [RUM 86] ou des réseaux récurrents avec des connexions FIR [BON 02].

Indépendamment de la nature des modèles utilisés, plusieurs méthodes permettent de traiter les problèmes de prévision multipas.

La première méthode, la plus répandue, consiste à mettre au point un prédicteur pour le problème à un pas de temps et à l'utiliser récursivement pour le problème multipas correspondant. Les estimations fournies par le modèle pour le pas de temps suivant sont représentées en entrée du modèle jusqu'à ce que l'horizon de la prévision désirée soit atteint. Cette méthode simple est pénalisée par l'accumulation des erreurs durant les pas de temps successifs ; le modèle diverge rapidement du comportement désiré.

Une deuxième et meilleure méthode consiste à apprendre le prédicteur sur le problème à un pas de temps et, en même temps, à utiliser la rétropropagation des pénalités à travers les pas de temps afin de punir le prédicteur pour les erreurs cumulées. Quand les modèles sont des PMCs ou des RNRs, une telle procédure est directement inspirée de l'algorithme *BackPropagation Through Time* (BPTT) [RUM 86], qui effectue une descente de gradient sur l'erreur cumulée.

Dans la troisième méthode, appelée méthode directe, le prédicteur traite directement le problème de prévision multipas. Des résultats expérimentaux et des considérations théoriques [ATI 99] tendent à prouver que la méthode directe se conduit toujours mieux que la première méthode et au moins aussi bien que la seconde. Une amélioration des performances est constatée si l'on utilise des RNRs que l'on entraîne en augmentant progressivement l'horizon de prévision [SUY 95]. Une méthode proche [DUH 01] enchaîne plusieurs réseaux. Le premier apprend à prédire à $t$+1, un second apprend à prédire à $t$+2 en utilisant comme entrée supplémentaire la prévision du premier réseau et ainsi de suite jusqu'à l'horizon de prévision souhaité.

Récemment, deux méthodes de prévision sensiblement différentes ont été proposées. Dans [JAE 04] un réseau de neurones récurrent de grande taille, appelé réservoir, est produit aléatoirement pour que ses neurones présentent un ensemble varié de comportements. En apprenant les poids sortants, ces comportements sont combinés afin d'obtenir la prévision désirée. Dans [SCH 05], une approche évolutionniste est employée pour obtenir les poids des neurones cachés de RNRs et des méthodes comme la régression linéaire ou la programmation quadratique sont appliquées pour calculer les relations linéaires optimales entre la couche cachée et la sortie.

Nous n'avons cité ici qu'une partie de la littérature récente sur la prévision neuronale mais les méthodes que nous avons mentionnées caractérisent bien les approches existantes. Néanmoins, il est particulièrement difficile d'identifier une typologie des problèmes de prévision multipas et d'évaluer quelles méthodes sont les plus appropriées pour un type donné de problème.

L'approche que nous présentons pour la prévision multipas se base sur la troisième méthode et applique un algorithme de boosting adapté aux RNRs, qui possèdent une mémoire interne et ne nécessitent pas de fenêtre temporelle.

## 3 Boosting pour les réseaux de neurones récurrents

L'algorithme de boosting utilisé doit être conforme aux restrictions imposées par le contexte général de l'application. Dans notre cas, il doit pouvoir travailler quand une quantité limitée de données est disponible et accepter comme régresseur les RNRs. Nos mises à jour sont basées sur la technique de [DRU 97], mais nous appliquons une transformation affine aux poids avant de les utiliser afin d'empêcher les RNRs d'ignorer purement et simplement les exemples plus faciles sur des problèmes comme celui des taches solaires. Puis, au lieu d'effectuer un échantillonnage avec remplacement selon la distribution mise à jour, nous préférons pondérer l'erreur calculée pour chaque exemple d'apprentissage (en utilisant toutes les données) à la sortie du RNR avec la valeur de la distribution correspondant à l'exemple.
La décision d'utiliser toutes les données de l'ensemble d'apprentissage pour la mise au point de chaque régresseur permet de respecter la condition selon laquelle l'algorithme doit être capable de traiter des séries temporelles constituées d'une quantité limitée de données. Notre idée est que les modifications des poids permettent de donner plus d'importance aux exemples difficiles pour l'apprentissage du régresseur suivant. L'algorithme de boosting que nous proposons peut alors être décrit comme suit :

(1) Initialiser les poids : $D_1(q) = 1/Q$ avec $Q$ nombre d'exemples d'apprentissage. $n = 0$
(2) Répéter
  (a) Incrémenter $n$. Développer un modèle (RNR) en employant l'ensemble d'apprentissage au complet et en pondérant l'erreur quadratique pour chaque exemple $q$ par $D_n(q)$
  (b) Mettre à jour les poids des exemples :
    (i) calculer $L_n(q)$ pour chaque exemple $q$, $q = 1, \ldots, Q$, avec une des trois fonctions :

$$L_n^{linéaire}(q) = \left| y_q^{(n)}(x_q) - y_q \right| / S_n \,,\quad L_n^{quadratique}(q) = \left| y_q^{(n)}(x_q) - y_q \right|^2 / S_n^2 \,,$$

$$L_n^{saturée}(q) = 1 - \exp\left( - \left| y_q^{(n)}(x_q) - y_q \right| / S_n \right), \text{ où } S_n = \sup_q \left| y_q^{(n)}(x_q) - y_q \right| \,;$$

    (ii) calculer la moyenne de l'erreur $\varepsilon_n = \sum_{q=1}^{Q} D_n(q) L_n(q)$ et la confiance $\alpha_n = (1 - \varepsilon_n)/\varepsilon_n$

    (iii) les poids des exemples deviennent $D_{n+1}(q) = \dfrac{1 + k \cdot p_{n+1}(q)}{Q + k}$ avec $p_{n+1}(q) = \dfrac{D_n(q) \alpha_n^{(L_n(q)-1)}}{Z_n}$

    où $Z_n$ est une constante de normalisation
  jusqu'à ce que le critère d'arrêt soit atteint : performance inférieure au hasard
(3) Combiner les RNRs en utilisant la médiane pondérée ou la moyenne pondérée.

## 4 Résultats et conclusion

Nous avons appliqué notre algorithme à la prévision multipas de trois séries temporelles de référence, une naturelle à dépendances temporelles trouvées expérimentalement à court terme et deux artificielles dont les dépendances à moyen et long termes sont contrôlées explicitement.
Les taches solaires sont des taches sombres liées à l'activité du champ magnétique du soleil. La série donne le nombre moyen annuel de taches apparues à la surface du soleil de 1700 à 1979. Il est d'usage d'employer les données de 1700 à 1920 pour l'apprentissage et les suivantes pour l'ensemble de test.
Les séries de Mackey-Glass sont des séries de référence utilisées pour l'évaluation de nombreux systèmes de prévision. Ces séries sont générées par l'équation différentielle non linéaire $dx(t)/dt = -0{,}1x(t) + 0{,}2x(t-\tau)/\left(1 + x^{10}(t-\tau)\right)$. Pour $\tau > 16{,}8$, la dynamique du modèle correspond au chaos déterministe. Le choix de $\tau = 17$ et de $\tau = 30$, valeurs habituellement retenues, permet de générer les séries MG17 et MG30 respectivement. Les 500 premières valeurs sont dédiées à l'ensemble d'apprentissage et les 100 valeurs suivantes à l'ensemble de test.

Les réseaux récurrents employés contiennent un neurone d'entrée, un neurone de sortie, un neurone de biais et une couche cachée entièrement récurrente composée de neurones avec des fonctions de transfert tangente hyperbolique. Cette couche est composée de 12 neurones pour les taches solaires et de 7 neurones pour les données de Mackey-Glass 17 et 30, ce qui correspond aux meilleurs résultats obtenus par BPTT sans boosting et pour des prévisions à un pas de temps.
Nous avons réalisé 5 expériences pour chaque architecture. La moyenne des résultats a été déterminée après ces 5 essais. Nous avons fixé à 50 le nombre maximal $n$ de RNRs pour chaque expérience. Nous comparons nos résultats avec l'algorithme BPTT d'origine et avec deux algorithmes constructifs, CBPTT et EBPTT, qui ajoutent des connexions à délais aux RNRs [BON 02].

| | Modèle | | | | | |
|---|---|---|---|---|---|---|
| *h* | BPTT | CBPTT | EBPTT | Lin. 10 | Qua. 20 | Sat. 20 |
| 1 | 0,24 | **0,17** | 0,19 | 0,18 | **0,17** | 0,18 |
| 2 | 0,88 | 0,69 | 0,53 | 0,43 | **0,40** | 0,42 |
| 3 | 1,14 | 0,99 | 0,79 | **0,54** | **0,54** | 0,67 |
| 4 | 1,22 | 1,17 | 0,80 | **0,67** | 0,73 | 0,76 |
| 5 | 1,01 | 0,99 | 0,88 | 0,74 | **0,69** | 0,77 |
| 6 | 1,02 | 1,01 | 0,84 | 0,73 | **0,68** | 0,74 |

**Tableau 1. Taches solaires : Moyenne des résultats de l'EQMN sur l'ensemble de test en fonction de l'horizon *h* (agrégation par médiane pondérée, qui donne les meilleurs résultats).**

| | MG17 | | | | | | MG30 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *h* | BPTT | CBPTT | EBPTT | Lin. 150 | Qua. 100 | Sat. 100 | BPTT | CBPTT | EBPTT | Lin. 300 | Qua. 200 | Sat. 150 |
| 1 | 21,9 | 12,9 | 0,9 | 0,17 | **0,16** | 0,17 | 11,7 | 2,5 | 1,8 | **0,45** | **0,45** | 0,47 |
| 2 | 179 | 124 | 101 | **0,24** | 0,28 | 0,25 | 19,9 | 9,7 | 3,3 | 0,49 | **0,48** | 0,59 |
| 3 | 145 | 124 | 16,3 | 0,57 | 0,57 | **0,52** | 3,9 | 2,2 | 1,6 | 0,56 | **0,55** | 0,64 |
| 4 | 8,6 | 7,6 | 4,5 | 0,57 | 0,54 | **0,52** | 2,2 | 2,1 | 1,5 | 0,47 | **0,43** | 0,48 |
| 5 | 266 | 253 | 181 | **0,98** | 1,26 | 1,27 | 2,6 | 2,3 | 0,9 | 0,85 | **0,67** | 0,72 |
| 6 | 321 | 321 | 232 | **2,11** | 15,2 | 4,66 | 8,9 | 8,3 | 6,4 | **1,75** | 1,92 | 1,80 |
| 7 | 320 | 320 | 297 | 2,65 | **2,48** | 3,69 | 70,1 | 65,6 | 64,3 | 2,98 | 4,56 | **2,27** |
| 8 | 310 | 308 | 285 | **9,97** | 12,3 | 24,8 | 336 | 203 | 112 | **5,08** | 109 | 57 |
| 9 | 312 | 309 | 302 | 8,19 | **7,18** | 8,89 | 801 | 379 | 257 | 84 | 275 | **3,71** |
| 10 | 336 | 331 | 219 | 14,1 | **12,2** | 15 | 892 | 383 | 73,7 | 2,79 | 204 | **2,63** |
| 11 | 289 | 218 | 252 | **9,80** | 12 | 16,8 | 411 | 230 | 285 | **6,34** | 21,3 | 8,05 |

**Tableau 2. Les séries MG17 et MG30 : Moyenne des résultats de l'EQMN (*$10^3$) sur l'ensemble de test en fonction de l'horizon *h* (agrégation par médiane pondérée).**

Les tableaux 1 et 2 comparent la moyenne des résultats de l'EQMN (Erreur Quadratique Moyenne Normalisée qui est l'erreur quadratique moyenne divisée par la variance de la série temporelle) avec ceux de la littérature. Nos modèles ont deux paramètres, la fonction de coût qui peut être linéaire, quadratique ou saturée, et $k$ le paramètre de notre algorithme de boosting. La valeur de $k$ choisie ici correspond aux meilleurs résultats obtenus pour une prévision à $t+1$. Par exemple, lin. 10 correspond à une fonction de coût linéaire avec $k=10$. Remarquons que nos résultats pourraient être améliorés par un réglage des paramètres directement sur le problème de prévision multipas.
Pour la série des taches solaires, notre algorithme développe environ 9 réseaux avec les deux fonctions de coût linéaire et quadratique et 30 réseaux avec la fonction saturée, comme pour la prévision à un pas de temps. Le nombre moyen de réseaux reste pratiquement constant quand l'horizon augmente.
Avec les deux séries MG17 et MG30, pour la fonction linéaire, notre algorithme développe en moyenne entre 30 et 40 réseaux pour MG17 et entre 20 et 45 pour MG30, pour la fonction quadratique, sur les deux

séries, le nombre moyen est légèrement supérieur et pour la fonction saturée, le nombre maximum de réseaux fixé par l'utilisateur (50) est atteint pour tous les horizons sur les deux séries.
Les résultats expérimentaux qui ont été obtenus sur les trois séries de références montrent que notre algorithme de boosting pour les réseaux de neurones récurrents améliore fortement les prévisions multipas. Le boosting semble renforcer la prise en compte des dépendances temporelles entre les données et les sorties désirées situées plusieurs pas de temps dans le futur. En effet, l'influence de l'algorithme de boosting se révèle moins notable pour les prévisions multipas avec la série des taches solaires alors que les dépendances à court terme sont connues pour être essentielles dans cette série. Le fait que pour les ensembles de données de Mackey-Glass les résultats soient meilleurs peut être expliqué en notant que les dépendances à long terme jouent un rôle plus important pour MG17 et MG30 que pour les taches solaires. L'évaluation des capacités de prévision multipas de notre algorithme sur d'autres séries chaotiques est un de nos objectifs à court terme.

## 5 Bibliographie


[ASS 05] ASSAAD M., BONÉ R., CARDOT H., "Study of the Behavior of a New Boosting Algorithm for Recurrent Neural Network", *International Conference on Artificial Neural Networks*, p. 169-174, 2005.

[ATI 99] ATIYA A. F., EL-SHOURA S. M., SHAHEEN S. I., EL-SHERIF M. S., "A Comparison Between Neural Network Forecasting Techniques - Case Study: River Flow Forecasting", *IEEE Transactions on Neural Networks*, vol. 10, n°2, p. 402-409, 1999.

[BAC 92] BACK A. D., TSOI A. C., "Stabilization Properties of Multilayer Feedforward Networks with Time-Delays Synapses", *Artificial Neural Networks*, vol. 2, p. 1113-1116, 1992.

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

[DRU 97] DRUKER H., "Improving Regressors using Boosting Techniques", *Fourteenth International Conference on Machine Learning*, p. 107-115, 1997.

[DUH 01] DUHOUX M., SUYKENS J., DE MOOR B., VANDEWALLE J., "Improved Long-Term Temperature Prediction by Chaining of Neural Networks", *International Journal of Neural Systems*, vol. 11, n°1, p. 1-10, 2001.

[JAE 04] JAEGER H., "Harnessing Nonlinearity: Predicting Chaotic Systems and Saving Energy in Wireless Communication", *Science*, vol. 304, p. 78-80, 2004.

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

[SCH 90] SCHAPIRE R. E., "The Strength of Weak Learnability", *Machine Learning*, vol. 5, p. 197-227, 1990.

[SCH 05] SCHMIDHUBER J., WIERSTRA D., GOMEZ F. J., Evolino: "Hybrid Neuroevolution / Optimal Linear Search for Sequence Learning", *Int. Joint Conference on Artificial Intelligence*, 2005.

[SUY 95] SUYKENS J. A. K., VANDEWALLE J., "Learning a Simple Recurrent Neural State Space Model to Behave Like Chua's Double Scroll", *IEEE Transactions on Circuits and Systems*, vol. 42, p. 499-502, 1995.

[WAN 94] WAN E. A., "Time Series Prediction by Using a Connection Network with Internal Delay Lines", *Time Series Prediction: Forecasting the Future and Understanding the Past*, Addison-Wesley, vol. XV, p. 195-217, 1994.

# Note sur le point d'égalité

**Jean Beney**

*Département Informatique, INSA de Lyon - Bât. Blaise Pascal*
*69621 Villeurbanne Cedex*
*jean.beney@insa-lyon.fr*

RÉSUMÉ. *De la définition du point d'égalité entre la précision et le rappel, nous montrons qu'il peut être calculé directement. Nous examinons les liens entre ce point d'égalité et le point où F(1) est maximum. Les expériences menées sur 3 ensembles de documents montrent que ces deux points sont toujours très proches, ce qui permet de gagner du temps dans la recherche du maximum pour F(1). De plus, si on remplace ce point maximum par le point d'égalité, les résultats obtenus sur un jeu de test indépendant ne sont pas significativement différents.*

MOTS-CLÉS : *classification supervisée, précision, rappel, fonction F*

## 1. Introduction

Parmi les divers moyens d'évaluer les résultats d'une classification supervisée, la précision et le rappel sont très populaires car ils peuvent être considérés comme représentant l'effet du résultat sur un utilisateur potentiel. Ces mesures sont décrites dans chaque article qui les utilise. Leurs formules, et parfois leur interprétation sont données, mais rarement les propriétés mathématiques de leurs formules.

Van Rijsbergen [RIJ 79] a fait état d'études des fondements statistiques de l'évaluation des classifieurs, notamment des rapports entre précision et retombées lorsque les scores sont distribués normalement. Pour une discussion (assez complète) des mesures utilisées, on se reportera à [SEB 02] [1].

Quand un score (généralement linéaire) est calculé pour chaque exemple dans chaque classe, un seuil donne la limite des scores au-dessus de laquelle les objets seront mis dans la classe. Il faut alors trouver une valeur du seuil qui soit un compromis car la précision et le rappel varient en sens contraire : quand on accepte plus d'exemples, le rappel augmente et la précision généralement diminue. Entre autres, deux stratégies (dont on peut facilement interpréter les résultats) sont employées pour trouver ce compromis : le point d'égalité où la précision et le rappel sont égaux et le maximum d'une fonction F des deux mesures (habituellement F(1), la moyenne harmonique).

Y-a-t-il un lien entre ces deux stratégies ? [YAN 99] a remarqué que les seuils obtenus dans les 2 cas sont souvent très proches mais pas identiques. Dans ces conditions, quelle est la perte si l'on remplace un des seuils par l'autre ?

Nous allons étudier quelques propriétés mathématiques du point d'égalité et du maximum de F(1), puis nous présenterons les résultats d'expérimentations menées en classification de documents.

1. où nous pouvons lire au sujet du point d'égalité :"a plot of $\pi$ [the precision] as a function of $\rho$ [the recall] is computed by repeatedly varying the threshold ; breakeven is the value of $\rho$ (or $\pi$) for which the plot intersects the $\rho = \pi$ line.". C'est un travail qui semble inutile, comme nous allons le montrer.

## 2. Le point d'égalité

Étant donné le résultat brut d'une classification automatique (une classe et son complément), les objets sont séparés en 4 groupes suivant deux propriétés : ils sont pertinents ou non, ils ont été sélectionnés ou non. On compte alors le nombre d'objets dans chaque groupe.

| | pertinent | non pertinent | total |
|---|---|---|---|
| sélectionné | $X$ | $Y$ | $S$ |
| non sélectionné | $Z$ | $W$ | $N-S$ |
| total | $R_e$ | $N-R_e$ | $N$ |

$P = \frac{X}{X+Y} = \frac{X}{S}$ $\qquad$ $R = \frac{X}{X+Z} = \frac{X}{R_e}$

La précision $P$ est la proportion d'objets pertinents parmi les sélectionnés. Le rappel $R$ est la proportion d'objets sélectionnés parmi les pertinents. Remarquez que $R_e$ est constant alors que $S$ varie avec les paramètres de la méthode utilisées, entre autre le seuil.

Comme [BLO 04] l'a noté, il s'ensuit immédiatement que le point d'égalité est obtenu, dans le cas général, quand le nombre d'objets sélectionnés est égal au nombre d'objets pertinents. Une autre solution de l'équation est le cas où il n'y a pas d'objets pertinents sélectionnés ($X = 0$) : c'est le cas du rejeteur universel, ou bien le cas où un ou plusieurs objets non pertinents ont obtenu des scores supérieurs à tous ceux des objets pertinents.

$$P = R \Leftrightarrow X = 0 \vee S = R_e \Leftrightarrow X = 0 \vee Y = Z \qquad [1]$$

Par la suite nous supposerons $X > 0$ et le cas qui nous intéresse est le cas général d'un classifieur pas trop mauvais : $S = R_e \Leftrightarrow Y = Z$ [2]

## 3. La fonction $F(\beta)$

La fonction F est utilisée pour trouver un compromis entre la précision et le rappel, en accordant éventuellement plus d'importance à l'un qu'à l'autre.

$$F(\beta) = \frac{(\beta^2+1)PR}{\beta^2 P + R} \qquad [2]$$

Propriétés : $\quad F(0) = P \quad F(\infty) = R \quad F(1) = \frac{2PR}{P+R} \quad 0 \leq F(\beta) \leq 1$

C'est pourquoi on choisit $\beta = 1$ quand il n'y a pas de raison de privilégier la précision ni le rappel.

$$F(1) = \frac{2PR}{P+R} = \frac{2X}{R_e+S} \qquad [3]$$

*Le maximum de F(1)*

Généralement, le nombre d'objets sélectionnés dépend d'un seuil $\theta$ dont il s'agit de trouver la valeur maximisant F(1). Ce maximum est obtenu quand la dérivée $F' = \frac{\delta F}{\delta \theta}$ est nulle :

$F' = 2\frac{X'(R_e+S)-XS'}{(Re+S)^2}$ Et comme $S = X + Y$ alors $F' = 2\frac{(R_e+Y)X'-XY'}{(Re+S)^2}$

Le dénominateur étant toujours positif si nous supposons qu'il existe au moins un exemple ($R_e > 0$) :

$$F' = 0 \;\Leftrightarrow\; (R_e+Y)X' = XY' \;\Leftrightarrow\; \frac{Y'}{X'} = \frac{X+Z+Y}{X} = \frac{Z+Y}{X} + 1$$

2. En cas d'égalité des scores, il se peut qu'on ne trouve pas de seuil qui fournisse exactement un point d'égalité. On se contentera alors d'un point approché $S = R_e \pm 1$ et la différence obtenue sera faible si le nombre d'objets est élevé.

Pour $X = 0$, $F$ est nulle ou indéterminée, donc n'admet pas maximum. Il reste : $F' = 0 \Leftrightarrow Y' \geq X'$

De plus, l'égalité $Y' = X'$ ne sera vérifiée que si $Y = Z = 0$ (classifieur parfait $P = R = 1$).

*La fonction F au point d'égalité*

La valeur de $F(\beta)$ au point d'égalité est ([YAN 99]) : $F(\beta) = \frac{(\beta^2+1)PP}{\beta^2 P+P} = P = R$

$P$ et $R$ varient en sens contraire. Ils sont liés par une relation qui dépend de la distribution des scores. En conséquence, le maximum de F(1) n'est pas nécessairement obtenu lorsque $P = R$. De la définition 2, nous déduisons :

$$F' = 2\frac{(P'R + PR')(P + R) - (P' + R')PR}{(P + R)^2} = 2\frac{P'R^2 + 2PR'}{(P + R)^2}$$

Est-il possible que $F' = 0$ quand $P = R$ ?

$$F' = 0 \Leftrightarrow P'R^2 + P^2R' = 0 \qquad\qquad F' = 0 \wedge P = R \Rightarrow P' + R' = 0$$

Donc $P = R \wedge \frac{P'}{R'} = -1 \Rightarrow F' = 0$, ce qui semble être vérifié lorsque l'on trace les courbes P/R/F :

au point d'égalité (intersection avec la bissectrice), la dérivée de la précision par rapport au rappel est -1.

## 4. Expérimentation

*Les ensembles de documents*

Les expérimentations ont portés sur trois ensembles de demandes de brevets. Deux d'entre eux, en anglais, proviennent de l'Office Européen des Brevets : EPO1F est composé d'une sélection de documents n'appartenant qu'à une des 16 classes (1000 documents par classe) ; EPO2F est un flot normal de demandes où beaucoup de documents appartiennent à plusieurs des 44 *directoires* (68 000 documents, au moins 2000 documents par classe). Le troisième ensemble contient des résumés en français de demandes à l'Office Mondial de la propriété intellectuelle (WIPO-F) : 680 000 documents inégalement répartis entre 119 classes.

*Conditions de l'expérimentation*

Le système utilisé est LCS [BEN 03, KOS 03] avec la méthode Winnow symétrique [LIT 88, DAG 97], ce qui explique que certains scores sont négatifs. Des expériences précédentes nous ont fait utiliser les paramètres suivants : promotion 1.03, démotion 0.97, 3 itérations sur les documents, seuil épais [0.5, 2], force des termes LTC, sélection des termes par l'incertitude [PET 02].

*Résultats*

Pour comparer le point d'égalité et le maximum de F(1), on peut d'abord comparer leur position, en terme du nombre (relatif) de documents sélectionnés ou selon la valeur du seuil :

$$\delta S = \frac{R_e - S_M}{R_e}, \qquad \delta\theta = \theta\, au\, BEP \; - \; \theta\, au\, F(1)max$$

Comme $0.5 < \theta < 0.9$, la différence n'est jamais supérieure à 2% de la valeur du seuil.

Ensuite, la différence de qualité est mesurée par $\delta F = F(1)max - F(1)au\, BEP$, d'abord sur l'ensemble d'apprentissage : la qualité au point d'égalité est parfois sensiblement moindre. Quand la classification est appliquée sur un ensemble indépendant de test, la variation moyenne de F(1) est bien plus petite. Cette différence $\delta F$ est souvent négative, ce qui signifie que le point d'égalité calculé sur l'ensemble d'apprentissage donne sur un ensemble de test une valeur de F(1) supérieure à celle obtenue avec le point donnant le maximum de F(1) sur l'ensemble d'apprentissage. Nous avons aussi remarqué que les plus grand écarts, positifs ou négatifs, sont obtenus pour des classes contenant peu de documents.

| | EPO1F (16 classes) | EPO2F (44 classes) | WIPO-F (119 classes) |
|---|---|---|---|
| $\delta S$min | -1.90% | -3.97% | -8.49% |
| $\delta S$max | 1.47% | 28.32% | 25.00% |
| $\delta S$moy | -0.10% | 12.11% | 1.97% |
| $\delta\theta$ min | -0.094 | -0.015 | -0.200 |
| $\delta\theta$ max | 0.044 | 0.070 | 0.224 |
| $\delta\theta$ moy | 0.007 | 0.004 | 0.007 |
| train :$\delta F$ min | 0% | 0.15% | 0.% |
| $\delta F$ max | 0.35% | 2.76% | 6.42% |
| $\delta F$ moy | 0.06% | 1.27% | 0.41% |
| test :$\delta F$ min | -1.27% | -1.71% | -6.89% |
| $\delta F$ max | 0.64% | 1.87% | 4.14% |
| $\delta F$ moy | -0.09% | 0.10% | 0.27% |

**TAB. 1.** *Différences entre le maximum de F(1) et le point d'égalité avec leurs macro-moyennes.*

## 5. Conclusion

Comme il est impossible de trouver un relation formelle reliant le point d'égalité entre précision et rappel et le maximum de F(1), l'expérimentation est nécessaire et montre qu'ils sont obtenus pour des valeurs de seuils très similaires. Comme le point d'égalité est obtenu directement en sélectionnant un nombre d'objets égal au nombre d'objets pertinents, le maximum de F(1) doit être cherché autour du point d'égalité, ce qui économise entre 9% et 14% du temps d'apprentissage, soit 37min pour les 119 classes de WIPO-F.

De plus, si l'on remplace le seuil où F(1) est maximum par le point d'égalité, la qualité obtenue sur un jeu de test indépendant n'est pas modifiée significativement, la variation moyenne restant bien en dessous de 1%.

## 6. Bibliographie


[BEN 03] BENEY J., KOSTER C., Classification supervisée de brevets : d'un jeu d'essai au cas réel, *Actes du XXIème congrès Inforsid*, 2003, p. 50–59.

[BLO 04] BLOEHDORN S., HOTHO A., Boosting for Text Classification with Semantic Features, *Proceedings of the Workshop on Text-Based Information Retrieval (TIR-04) at the 27th German Conference on Artificial Intelligence (KI 2004)*, 2004, p. 25–41.

[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 03] KOSTER C., SEUTTER M., BENEY J., Multi-Classification of Patent Applications with Winnow, *Proceedings of PSI 2003, LNCS 2890*, Springer-Verlag, 2003, p. 545–554.

[LIT 88] LITTLESTONE N., Learning quickly when irrelevant attributes abound : A new linear-threshold algorithm, *Machine Learning*, vol. 2, 1988, p. 285–318.

[PET 02] PETERS C., KOSTER C. A., Uncertainty-Based Noise Reduction and Term Selection in Text Categorization, *Proceedings of the 24th BCS-IRSG European Colloquium on IR Research : Advances in Information Retrieval*, 2002, p. 248–267.

[RIJ 79] VAN RIJSBERGEN C. J., *Information retrieval*, Butterworths, 1979.

[SEB 02] SEBASTIANI F., Machine learning in automated text categorisation, *ACM Computing Surveys*, vol. 34(1), 2002, p. 1–47.

[YAN 99] YANG Y., An evaluation of statistical approaches to text categorization, *Information Retrieval*, vol. 1(1), 1999, p. 69–90.

# Concept et Inférence pour la Statistique Structurale

**Frédéric Chateau**

*Laboratoire ERIC,*
*Université Lumière – Lyon2*
*5 av. Pierre Mendès-France*
*69 676 Bron Cedex, France*

*RÉSUMÉ. Les méthodes de l'analyse des données saisissent les individus dans leur réalité multidimensionnelle et relative. Mais les connaissances qu'elles induisent n'entrent pas dans le cadre inférentiel de la statistique fishérienne, au double sens de la généralisation et de la prédictibilité. Une réflexion sur les fondements logiques de l'inférence et de la décision nous conduit à distinguer les catégories de la statistique individu-propriété et de la statistique structurale. Pour celle-ci, nous développons des méthodes analogues à celles de la théorie des sondages et à la discrimination statistique. Nous proposons une application dans le domaine de l'environnement sonore urbain*[1].

*MOTS-CLÉS : Analyse des Données, Inférence, Logique*

## 1 Perspective logique de la Statistique Structurale

La pratique moderne de la logique des prédicats se fonde sur le concept aristotélicien de la substance et des attributs, qui trouve son expression moderne dans la théorie des types de Russel. La relation d'appartenance implique des attributs d'un type supérieur à celui de l'individu qui en est l'objet. Un concept est défini en intention à l'aide de propositions logiques qui permettent d'en établir l'extension.

Cette théorie fait écho à l'utilisation pour l'analyse discriminante des groupes définis *a priori*, mais elle sous-tend également la pratique des probabilités d'appartenance à une classe-attribut. On relie ainsi les développements de la statistique fishérienne à l'épistémologie du concept de la logique formelle. Nous appellerons donc ici ces développements, **statistique individu-propriété**.

La logique de la relation individu-propriété ne s'applique pas aux repérages spatiaux construits par l'analyse des données : typologies, coordonnées ou régions de l'espace factoriel. Ainsi, on ne peut utiliser sans contradiction [SUT 95] les typologies construites par classification automatique, car elles ne délimitent pas des classes–concepts définis en intention. Au plan formel, les objets de l'analyse des données ne présentent pas la consistance du concept-attribut.

[1] *INRETS - Projet de recherche PIE, «Prospective et Indicateurs des impacts des transports sur l'Environnement, outil d'évaluation et d'aide à la décision». Les données sont les niveaux sonores moyens sur 1" mesurés pour 200 sites. Ils sont recodés en indices macro-acoustiques (indices énergétiques et comptages d'événements [BEA 05]) pour décrire les profils temporels des sites : les situations sonores urbaines.*

En effet, la logique de la relation individu-propriété interdit de définir une propriété à partir de la collection des individus car elle doit lui être préalable. Or, caractériser un individu à partir de sa position dans l'espace des facteurs d'une analyse multivariée revient à faire dépendre ses propriétés des autres individus actifs de l'analyse et la situation est pire encore si cette position est résumée par une classification automatique dont les frontières dépendent de la distribution empirique.

Observons également que la place de l'interprétation affaiblit encore le statut des objets de l'analyse des données comme attributs d'une proposition prédicative. En effet, l'opérateur y définirait une mesure selon la signification qu'il donne à ses *observations* – interprétation des axes, finesse des partitions.

Les constructions de l'analyse des données et des propositions qu'elles permettent d'énoncer, relèvent de l'*existence statistique* et leur statut épistémologique est précisé dans un article fondateur [BEN 80]. Pour autant, ces objets ne relèvent pas du cadre de la logique des prédicats qui sous-tend le langage et les développements probabilistes. La question de leur usage inférentiel conduit à les inscrire dans un cadre méthodologique et conceptuel spécifique; ce que nous appelons **statistique structurale**.

L'épistémologie du concept est traversée d'une autre ligne de pensée qui s'origine chez Spinoza[2] et trouve avec Wittgenstein son développement moderne. Elle contourne la distinction scolastique entre substance et attribut dans une visée méthodologique. Spinoza contre Thomas d'Aquin ou Wittgenstein contre Russell, objectent à la logique formelle de ne pas rendre compte du monde et cherchent les conditions de production d'énoncés rationnels hors la logique des prédicats. Ainsi, Wittgenstein, admet une vision du concept comme constitution de familles de ressemblance ou structure qui associe forme et sens.

Nous définissons le concept de la statistique structurale comme l'ensemble des interrelations entre les individus d'une collection selon le point de vue des axes résumant un ensemble de variables-propriétés qui le représentent sans l'épuiser. Ce concept se constitue à la fois des systèmes d'oppositions qui structurent l'univers et de la place de chacun des individus. Les méthodes factorielles le dégagent de la partie contingente des mesures et en réduisent la dimensionnalité pour le rendre opératoire.

Nous déployons ce concept dans les deux directions principales développées avec succès dans le cadre de la statistique individu-propriété : l'inférence d'un échantillon vers une population mère et la prédictibilité de la position d'un individu anonyme dans une structure :

- l'inférence - la théorie de l'estimation a du sens dans des contextes et des usages limités en analyse des données. Pourtant, il est fréquent d'utiliser ces outils sur des échantillons aléatoires.
- l'aide à la décision – comme évoqué *supra*, le calcul des probabilités *a posteriori* est inapproprié hors le cadre individu-propriété. Pourtant, les projections d'espace-cible analogue à une variable endogène sont utilisées pour des raisonnements prédictifs. Mais elles n'offrent pas la rigueur de l'analyse discriminante et ne présentent pas la même proximité au langage de la décision.

## 2 Méthodes inférentielles pour la Statistique Structurale

La problématique de l'inférence en analyse des données, a donné lieu à des travaux de statistique mathématique qui ont établi un certain nombre de distributions liées à la décomposition spectrale de matrice empiriques. Mais ces travaux ont une portée pratique réduite, d'une part à cause de conditions distributionnelles restrictives et, d'autre part, parce qu'ils concernent surtout les qualités des éléments, chacun considéré indépendamment, et non globalement les uns par rapport aux autres - comme il serait pertinent dans le cadre conceptuel de la statistique structurale.

[2] *En particulier, on peut observer l'usage du terme* essence *au lieu de la* substance *scolastique dans la proposition suivante : « une affection quelconque d'un individu quelconque montre avec l'affection d'un autre d'autant plus de discordance que l'essence de l'un diffère de l'essence de l'autre ». Proposition 57 du livre III de l'éthique.*

Les fluctuations d'échantillonnage sont abordées à partir des techniques de simulation, en particulier de ré-échantillonnage *bootstrap*; l'expression «stabilité externe» [GRE 84], a été introduite pour ces travaux. L'ACP normée pose un problème de dimension de l'espace des individus, traité par projection des matrices de corrélation *bootstrap* comme éléments supplémentaires [CHA 96]. Ces méthodes permettent de tracer des zones de confiances pour les positions des variables.

Par rapport au cadre de la théorie de l'estimation, il s'agit d'un usage non standard du *bootstrap,* en ceci que les calculs s'appuient sur les axes construits d'après l'échantillon originel. C'est une inférence *post data* ou *bootstrap* partiel. La vision des fluctuations offerte par le *bootstrap* partiel apparaît conservatrice par rapport au *bootstrap* complet qui recalcule la structure - de manière dégradée - sur chaque échantillon.

Pour approfondir ces méthodes dans le cadre de la statistique structurale, nous concentrons l'attention sur les positions des $n$ individus de l'échantillon originel $X_{np}$. Ils sont imparfaitement représentés dans l'échantillon *bootstrap* $X^*_{np}$ mais les relations pseudo-barycentriques permettent de dériver leurs coordonnées *bootstrap* $\varphi^{**}_{iq}$ des coordonnées *bootstrap* $\psi^{*}_{jq}$ des $p$ variables, établies par simulation.

Nous introduisons l'indice de consistance de la structure *Cs* comme le rapport de l'écart-type d'échantillonnage des distances entre individus, et la moyenne de cette distance – variance empirique de la collection. On la calcule pour chaque couple d'individus entre les $B$ échantillons *bootstrap*.

$$Cs = 1 - \left[V(d)\right]^{1/2} \Big/ E(d)$$

$$\text{avec} \qquad V(d) = \frac{1}{2n(n-1)} \sum_{i,i'} \frac{1}{B-1} \sum_{b} \left[ d(i,i') - d^{**}_{b}(i,i') \right]^2 \qquad E(d) = \frac{1}{2n(n-1)} \sum_{i,i'} d(i,i')$$

On constate sur un jeu de données simulées que l'indice *Cs* croît avec la taille $n$ de l'échantillon, et décroît avec le nombre $q$ d'axes retenus quand celui-ci reste faible devant $p$.

Dans la même perspective, on teste la stabilité d'affectation des individus dans un codage de la structure. Pour les situations sonores urbaines c'est une classification centrée sur les demi-axes d'une ACP – méthode inspirée des *k-means* axiales [LEL 94]. Une hypersphère-centre est définie par un rayon $r$ de façon à contenir une fraction fixée de l'échantillon – ici, *20%*. A l'extérieur, chaque individu est affecté au demi-axe qui lui est le plus proche. En dimension *2*, cette classification se représente comme suit :

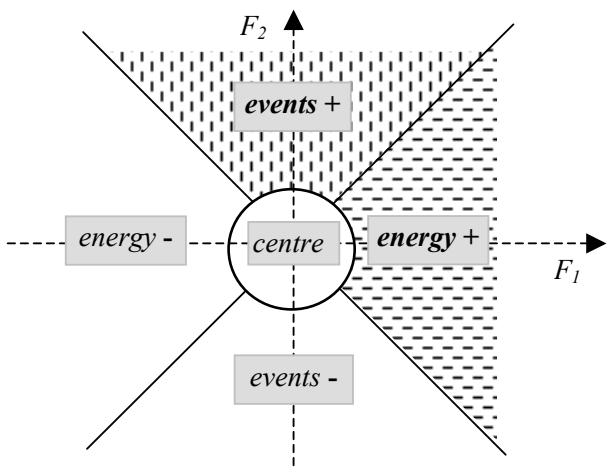



*figure f1 : espace des situations sonores urbaines (plan F1xF2), codage en 5 situations-classes.*

L'axe *F1* correspond à l'énergie acoustique moyenne des sites, et l'axe *F2* à la densité d'événements sonores isolés. Les classes *energy+* et *events+* correspondent à deux types de situations environnementales pathologiques. La stabilité de la structure du point de vue de ce codage est représentée par l'indice de consistance de la classification *Cc*. Pour chaque échantillon *bootstrap*, on dérive comme précédemment les coordonnées *bootstrap* des individus de l'échantillon originel pour calculer leur classe d'affectation *bootstrap*. La comparaison de cette classe d'affectation à sa classe réelle permet de calculer une valeur du Kappa pondéré de Cohen, $\kappa_\omega$ , qui mesure l'agrément entre la classification originelle et la classification *bootstrap*. *Cc* est la moyenne sur *B* simulations des *B* valeurs $\kappa_{\omega b}$ obtenues.

## 3 Théorie de la décision pour la Statistique Structurale

Le paradigme dominant de la statistique décisionnelle repose sur le calcul des probabilités *a posteriori* d'appartenance à la classe souhaitable du point de vue du décisionnaire - son score. Ce score est ensuite muni d'un seuil de décision. Nous avons vu *supra* que cette méthode est inappropriée à la statistique structurale dont les objets sont hétérodoxes au concept de la statistique individu-propriété. Pourtant, il est utile de prédire la position d'un individu dans une structure; c'est le cas pour les greffes d'enquêtes.

Décomposons le temps logique de la décision reposant sur l'analyse discriminante :

- on suppose que les individus anonymes appartiennent à une des classes, puis on leur affecte la classe la plus probable - classement par la règle du ratio de vraisemblance,
- on leur applique une décision conforme à celle qui serait prise en connaissance de leur classe si elle était dans la réalité identique à leur classe affectée.

L'analyse discriminante, par proximité au langage de la décision, confond classe réelle et classe d'affectation et simplifie certains problèmes pour les exprimer dans ce formalisme. Prenons à la suite de [HAN 96] l'exemple du *credit scoring* : le risque bancaire s'exprime en classes de "bon et mauvais client". Sans que l'échantillon d'apprentissage change, le prêteur peut modifier le seuil de décision selon ses objectifs. On voit que l'analyse ne repose pas sur deux classes *a priori*, mais crée deux classes : "accepté ou refusé". Les catégories "bon et mauvais client" sont des notions intermédiaires de calcul, qui n'ont finalement pas plus de consistance que les classes résumant un espace factoriel après une classification.

En d'autres termes, l'aide à la décision délimite des régions de décision dans l'espace des prédicteurs, mais ne requiert pas obligatoirement des classes *a priori* au sens du concept individus-propriétés. Cette formulation rend les objets de la statistique structurale admissibles pour la décision : elle repose sur la prédiction de la position d'un individu dans la structure, en lieu et place de sa classe prédite.

Notre proposition s'articule en deux temps : construire des projections dans l'espace des prédicteurs de la structure à prédire, puis élaborer scores et règles de décision sur ces projections.

Un graphe complet sur les individus d'arêtes valuées selon leur proximité ou un graphe de voisinages, sont appropriés à décrire le concept de structure. L'analyse de la contiguïté [LEB 69] sur ces graphes généralise l'analyse factorielle discriminante à *k* classes, qui se ramène au cas d'un graphe simple formé de *k* cliques disjointes. Les résultats obtenus [CHA 99] avec cette famille de méthodes - contiguïté, analyse locale au graphe complémentaire, barycentrique locale - montrent leurs qualités de robustesse et de précision.

En particulier, dans des situations comparables, on trouve un avantage à l'analyse des contiguïtés face au sur-ajustement pour les problèmes pauvrement posés. Le codage de l'information à prédire sous forme de graphe sur les individus permet également une grande souplesse de mesure des prédicteurs – individus, groupes - pour concentrer l'attention sur la spécificité du problème posé.

Dans la recherche menée sur l'environnement sonore urbain, les prédicteurs sont des données de morphologie urbaine extraites du SIG du Grand Lyon. L'application vise à munir les sites non observés ou anonymes, d'une situation sonore au sens de la structure dégagée sur les sites d'apprentissage.

Les contiguïtés entre sites d'apprentissage sont calculées sur l'espace des situations sonores (figure *f1*). L'analyse des contiguïtés construit un sous-espace de l'espace des prédicteurs optimal pour projeter les distances sonores entre sites.

Les scores sont dérivés de la notion d'agrément du Kappa pondéré de Cohen, étendue au contexte structural par un agrément entre régions de la structure sur les individus d'apprentissages. Un site anonyme $x$ se voit affecter la classe $C_l$ parmi les cinq situations sonores qui maximise l'agrément moyen $Ag_l(x)$ calculé sur les classes $C(x_k)$ de ses plus proches voisins $x_k$ parmi les sites d'apprentissages, soit la règle $\delta_l$ :

$$\delta_l = argMax\left[Ag_l(x)\right]$$

$$avec \qquad Ag_l(x) = \sum_k \omega\left[l, C(x_k)\right] \quad où \quad \omega(l,l') = 1 - \frac{d(Gl,Gl')}{\max\left[d(Gj,Gj')\right]}$$

Cette règle, tout comme l'analyse des contiguïtés, prend en compte la structure des situations sonores. Plus généralement, la notion d'agrément entre classes permet de construire des règles de décision pour des situations où l'information à prédire est plus complexe que la relation individu propriété. C'est le cas pour une variable ordinale ou une classification pyramidale, on peut également intégrer des coûts d'erreur de classement inégaux en jouant sur les agréments avec une classe médiane.

## 4 Bibliographie


[BEA 05] BEAUMONT J. et SEMIDOR C. "Event descriptors for qualifying the urban sound environment", *EAA European Acoustics Association : Forum Acusticum*, Budapest.

[BEN 80] BENZECRI J.P., "L'âme au bout d'un rasoir", *Les Cahiers de l'Analyse des Données,* vol V - 1980, p 229-242

[CHA 96] CHATEAU F. et LEBART L., "Assessing sample variability in SGVD based analysis, from bootstrap and related methods", *Compstat 96,* Un. Polytècnica de Catalunya, Pratt Ed. Springer Verlag.

[CHA 99] CHATEAU F. "Structured Discriminant Analysis", *Communications in Statistics, Theory and Methods* - Volume 28 N°2. Marcel Dekker, New York.

[GRE 84] GREENACRE M. *Theory and application of correspondence analysis*. Academic Press, London

[HAN 96] HAND D.J. OLLIVER J.J. and LUNN A.D., "Discriminant analysis when the classes arise from a continuum" *Pattern Recognition*, 31 . p 641-650.

[LEB 69]. LEBART L., "Analyse statistique de la contiguïté," *Publications de l'Institut de Statistique de l'Université de Paris*, 1969. Vol XVIII, p 81-112.

[LEL 94] LELU A., Clusters and factors: neural algorithms for a representation of huge data sets. *New Approaches in Classification and Data Analysis*, E. Diday, & al. p 241-248, Springer-Verlag.

[SUT 95] SUTCLIFFE J.P., "Logical machinery for deciding what is or is not classification", *The Newsletter of the Classification Society of North America*. vol 41, p 2-6

# Comparaison d'une méthode de classification descendante hiérarchique monothétique avec Ward et les centres mobiles

**Marie Chavent**[1], **Olivier Briant**[2], **Yves Lechevallier**[3]

[1] *Mathématiques Appliquées de Bordeaux, UMR CNRS 5466, Universités Bordeaux1-2*
[2] *GILCO-Institut National Polytechnique de Grenoble*
[3] *Institut National de Recherche en Informatique et en Automatique*

*RÉSUMÉ. DIVCLUS-T est une méthode de classification descendante hiérarchique monothétique qui cherche à optimiser à chaque étape le critère d'inertie intra-classe. Le dendrogramme a la particularité d'être décrit à chaque palier par une question binaire, comme les arbres de décision ou de régression. Notre objectif est de comparer cette méthode avec deux autres méthodes cherchant également à minimiser l'inertie intra-classe : la classification ascendante hiérarchique de Ward et les centres mobiles. Après avoir donné et commenté les complexités de ces trois méthodes dans le cas quantitatif, nous les comparerons empiriquement sur six jeux de données du Machine Learning repository.*

*MOTS-CLÉS : Classification divisive, classe monothétique, hiérarchie indicée, question binaire.*

## 1. Introduction

La méthode DIVCLUS-T est une méthode de classification descendante hiérarchique qui peut s'appliquer à des données quantitatives ou qualitatives. Cette méthode procède comme toute méthode descendante hiérarchique par divisions successives et s'articule autour des 3 points suivants :

- Les divisions s'arrêtent après $k$ étapes. On obtient donc le "haut" du dendrogramme c'est à dire les partitions de 2 à $k+1$ classes.
- A chaque étape cette méthode choisit de diviser la classe telle que la nouvelle partition ainsi obtenue soit d'inertie intra-classe minimum. Pour des données qualitatives, l'inertie est calculée avec la distance du $\chi^2$ sur le tableau disjonctif complet. Le critère d'inertie intra-classe étant additif cela revient à choisir la classe telle que la variation de l'inertie obtenue en la divisant soit maximum. Dans Ward on agrège à chaque étape les deux classes minimisant ce même critère de variation de l'inertie. Dans Ward et dans DIVCLUS-T on utilise donc le même critère pour indicer la hiérarchie et donc valuer la hauteur des paliers dans le dendrogramme. Dans DIVCLUS-T le choix de la classe à diviser est nécessaire puisque l'on ne continue pas nécessairement les divisions jusqu'à l'obtention des singletons.
- L'algorithme de bi-partitionnement d'une classe à $n$ éléments en deux sous-classes n'évalue pas l'inertie intra-classe des $2^{n-1}-1$ bi-partitions possibles pour en retenir la meilleure, mais évalue ce critère sur l'ensemble de toutes les bi-partitions induites par l'ensemble de toutes les questions binaires. On utilise donc ici l'approche monothétique des arbres de décisions et de régression [MOR 63] [BREI 84] mais dans un cadre non supervisé. Les différences sont nombreuses. En particulier il n'y a pas de variable à expliquer et pas d'élagage.

On peut trouver les détails concernant cette méthode et son implémentation de [CHA 06]. Elle avait été définie mais de manière plus succincte dans le cas quantitatif dans [CHA 97] et dans le cas qualitatif dans [CHA 99]. Dans [CHA 99] la méthode, appelée DIVOP à l'époque, était présentée dans le cadre d'une application en dermatologie conjointement avec une autre méthode divisive monothétique appelée DIVAF, basée sur l'analyse des correspon-

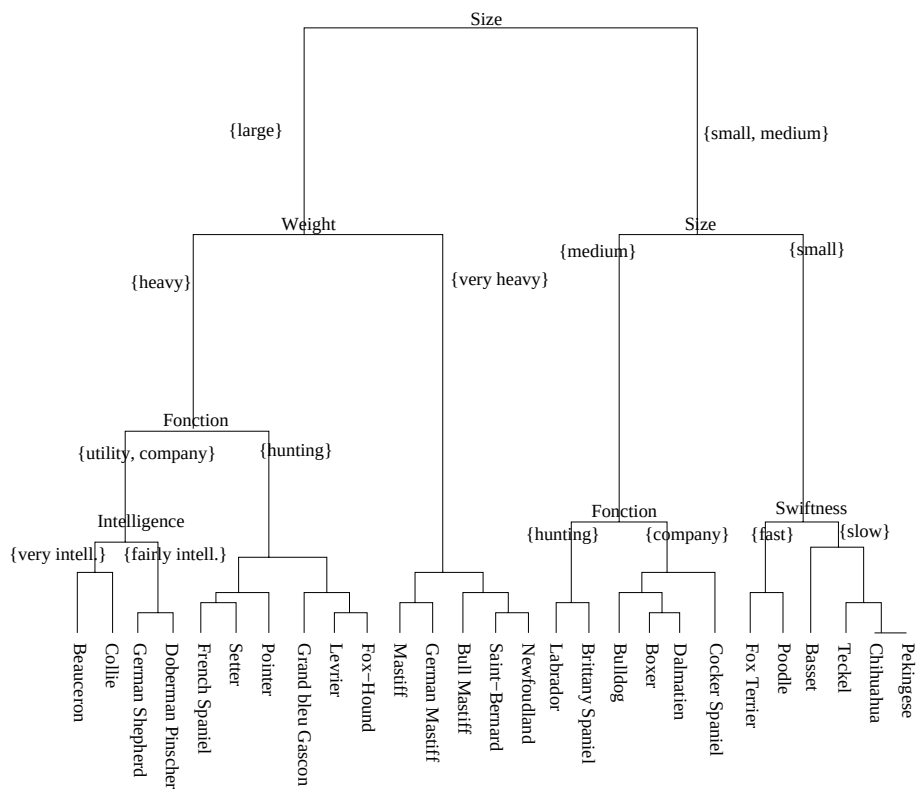



**FIG. 1.** *Dendrogramme obtenu avec DIVCLUS-T pour les données chiens*

dances multiples. Une méthode divisive de type monothétique utilisant le processus de Poisson a également été proposée par [PIR 94].

La méthode DIVCLUS-T est donc une méthode simple dont la principale propriété est de fournir une interprétation immédiate et monothétique du dendrogramme et des classes de la hiérarchie. On voit par exemple sur le dendrogramme obtenu avec DIVCLUS-T pour les données qualitatives sur les chiens [SAP 90] (Figure 1) les sept questions binaires associées aux sept premières divisions. Dans cet exemple on a continué les divisions jusqu'à l'obtention des singletons afin de comparer cette hiérarchie avec celle obtenue avec Ward sur toutes les coordonnées factorielles de l'AFCM. Sur cet exemple les hiérarchies et les dendrogrammes obtenus avec Ward et DIVCLUS-T sont identiques.

## 2. La complexité

La complexité de DIVCLUS-T pour des données quantitatives est $o(kpn(log(n) + p))$ avec $k$ le nombre de classe de la partition la plus fine, $p$ le nombre de variables et $n$ le nombre d'objets.

La classification ascendant hiérarchique de Ward avec la méthode des voisins réciproques est en $o(n^2)$. Le calcul préalable des $n(n-1)/2$ distances Euclidiennes est quant à lui en $o(pn^2)$. On peut donc considérer que l'algorithme de Ward est en $o(pn^2)$. DIVCLUS-T est donc plus performant en terme de complexité que Ward pour de petites valeurs de $k$ et $p < n$.

L'algorithme des centres mobiles est en $o(kpnT)$ où $k$ est le nombre de classes de la partition et $T$ le nombre maximum d'iterations. DIVCLUS-T est alors plus performant en terme de complexité que les centres mobiles lorsque $log(n) + p < T$.

Pour des données qualitatives, la complexité de DIVCLUS-T est exponentielle avec le nombre de modalités des variables. Une solution pour réduire cette complexité est de définir un ordre sur les modalités en prenant par exemple les ordres des modalités sur les axes principaux de l'AFCM.

## 3. Comparaison empirique

Nous avons voulu répondre à la question suivante : l'aspect rigide et simpliste du processus monothétique de DIVCLUS-T implique-t-il des partitions beaucoup moins bonnes en terme d'inertie intra-classe ? Pour donner un premier élément de réponse à cette question, nous avons comparés empiriquement le pourcentage d'inertie expliquée des partitions de 2 à 15 classes obtenues avec DIVCLUS-T, Ward et les centres mobiles, sur 6 jeux de données du Machine Learning repository [HET 98]. Ces six bases, trois quantitatives et trois qualitatives, sont décrites dans le tableau 1.

| Nom | Type | Nb objets | Nb variables(nb categories) |
|---|---|---|---|
| Glass | quantitatives | 214 | 8 |
| Pima Indians diabete | quantitatives | 768 | 8 |
| Abalone | quantitatives | 4177 | 7 |
| Zoo | qualitatives | 101 | 15(2) + 1(6 ) |
| Solar Flare | qualitatives | 323 | 2(6) + 1(4) + 1(3) + 6(2) |
| Contraceptive Method Choice (CMC) | qualitatives | 1473 | 9(4) |

**TAB. 1.** *Description des 6 jeux de données*

Le tableau 2 donne les résultats pour les trois bases quantitatives et les trois méthodes (colonne DIV pour DIVCLUS-T, colonne WARD pour Ward et colonne W+km pour les centres mobiles sur la partition de Ward). Pour les données GLASS, on note que DIVCLUS-T est parfois meilleur que Ward (pour 4 classes), parfois moins bon (pour 2, 3 classes et de 12 à 15 classes) ou encore parfois équivalent (de 5 à 11 classes). Pour les données PIMA, DIVCLUS-T est meilleur ou équivalent à Ward jusqu'à 4 classes puis Ward prend le dessus à partir de 5 classes. Pour les données ABALONE qui est la plus grande base (4177 objets), DIVCLUS-T est meilleur que Ward jusqu'à 4 classes et fournit des résultats proches ensuite. Finalement sur ces trois jeux de données DIVCLUS-T semble plus performant en terme d'inertie expliquée pour les partitions en peu de classes (ce qui n'est pas surprenant puisque DIVCLUS-T descend et que Ward monte) et pour les bases plus volumineuses (ce qui n'est pas non plus surprenant puisque lorsque le nombre d'objet augmente, le nombre de bi-partitions évaluées à chaque étape augmente également).

| | Glass | | | Pima | | | Abalone | | |
|---|---|---|---|---|---|---|---|---|---|
| K | DIV | WARD | W+km | DIV | WARD | W+km | DIV | WARD | W+km |
| 2 | 21.5 | 22.5 | 22.8 | 14.8 | 13.3 | 16.4 | 60.2 | 57.7 | 60.9 |
| 3 | 33.6 | 34.1 | 34.4 | 23.2 | 21.6 | 24.5 | 72.5 | 74.8 | 76.0 |
| 4 | 45.2 | 43.3 | 46.6 | 29.4 | 29.4 | 36.2 | 81.7 | 80.0 | 82.5 |
| 5 | 53.4 | 53.0 | 54.8 | 34.6 | 34.9 | 40.9 | 84.2 | 85.0 | 86.0 |
| 6 | 58.2 | 58.4 | 60.0 | 38.2 | 40.0 | 45.3 | 86.3 | 86.8 | 87.8 |
| 7 | 63.1 | 63.5 | 65.7 | 40.9 | 44.4 | 48.8 | 88.3 | 88.4 | 89.6 |
| 8 | 66.3 | 66.8 | 68.9 | 43.2 | 47.0 | 51.1 | 89.8 | 89.9 | 90.7 |
| 9 | 69.2 | 69.2 | 71.6 | 45.2 | 49.1 | 52.4 | 91.0 | 90.9 | 91.7 |
| 10 | 71.4 | 71.5 | 73.9 | 47.2 | 50.7 | 54.1 | 91.7 | 91.6 | 92.4 |
| 11 | 73.2 | 73.8 | 75.6 | 48.8 | 52.4 | 56.0 | 92.0 | 92.1 | 92.8 |
| 12 | 74.7 | 76.0 | 77.0 | 50.4 | 53.9 | 58.0 | 92.3 | 92.4 | 93.0 |
| 13 | 76.2 | 77.6 | 78.7 | 52.0 | 55.2 | 58.8 | 92.6 | 92.7 | 93.3 |
| 14 | 77.4 | 79.1 | 80.2 | 53.4 | 56.5 | 60.0 | 92.8 | 93.0 | 93.7 |
| 15 | 78.5 | 80.4 | 81.0 | 54.6 | 57.7 | 61.0 | 93.0 | 93.2 | 93.9 |

**TAB. 2.** *Données quantitatives*

Pour les trois bases qualitatives (tableau 3) on obtient le même type de résultats. Pour les données Solar Flare et CMC, DIVCLUS-T est meilleur que Ward jusqu'à respectivement 10 et 8 classes. Pour les données Zoo,

DIVCLUS-T reste toujours en dessous de Ward. C'est peut-être du aux fait que les variables sont binaires et que pour des données qualitatives, le nombre de bi-partitions évaluées à chaque étape augmente avec le nombre de modalités.

| | Zoo | | | Solar Flare | | | CMC | | |
|---|---|---|---|---|---|---|---|---|---|
| K | DIV | WARD | W+km | DIV | WARD | W+km | DIV | WARD | W+km |
| 2 | 23.7 | 24.7 | 26.2 | 12.7 | 12.6 | 12.7 | 8.4 | 8.2 | 8.5 |
| 3 | 38.2 | 40.8 | 41.8 | 23.8 | 22.4 | 23.8 | 14.0 | 13.1 | 14.8 |
| 4 | 50.1 | 53.7 | 54.9 | 32.8 | 29.3 | 33.1 | 18.9 | 17.3 | 20.5 |
| 5 | 55.6 | 60.4 | 61.0 | 38.2 | 35.1 | 38.4 | 23.0 | 21.3 | 24.0 |
| 6 | 60.9 | 64.3 | 65.1 | 43.0 | 40.0 | 42.7 | 26.3 | 24.9 | 27.7 |
| 7 | 65.6 | 67.5 | 68.4 | 47.7 | 45.0 | 47.6 | 28.4 | 28.1 | 29.8 |
| 8 | 68.9 | 70.6 | 71.3 | 51.6 | 49.8 | 52.1 | 30.3 | 30.7 | 32.7 |
| 9 | 71.8 | 73.7 | 73.7 | 54.3 | 53.5 | 54.6 | 32.1 | 33.4 | 35.2 |
| 10 | 74.7 | 75.9 | 75.9 | 57.0 | 57.1 | 58.3 | 33.8 | 35.5 | 37.7 |
| 11 | 76.7 | 77.5 | 77.5 | 59.3 | 60.4 | 61.7 | 35.5 | 37.5 | 40.1 |
| 12 | 78.4 | 79.1 | 79.1 | 61.3 | 62.9 | 64.4 | 36.9 | 39.4 | 41.5 |
| 13 | 80.1 | 80.6 | 80.6 | 63.1 | 65.2 | 65.7 | 38.1 | 41.0 | 42.9 |
| 14 | 81.3 | 81.8 | 81.8 | 64.5 | 66.2 | 67.7 | 39.2 | 42.0 | 44.2 |
| 15 | 82.8 | 82.8 | 82.8 | 65.8 | 68.6 | 69.3 | 40.3 | 43.1 | 44.9 |

**TAB. 3.** *Données qualitatives*

## 4. Conclusion

DIVCLUS-T est une méthode monothétique qui a l'avantage par rapport aux méthodes polythétiques telles que WARD et les centres mobiles, de donner une interprétation très simple et immédiate des classes et un arbre hiérarchique facile a lire et à comprendre par l'utilisateurs. Il est en outre normal que cette contrainte sur l'interprétation des classes, imposée dans le processus de classification, implique une perte de qualité au niveau du critère d'inertie. Il n'est donc pas surprenant que DIVCLUS-T soit généralement moins performant que WARD ou les centres mobiles. Le fait que DIVCLUS-T puisse être meilleur que WARD et que dans le cas contraire, la différence entre DIVCLUS-T et WARD est comparable à celle entre WARD et les centres mobiles, est un premier résultat encouragent, une étude plus approfondie restant nécessaire. Cette méthode semble cependant une bonne alternative à WARD pour les utilisateurs qui s'intéressent aux partitions en peu de classes et à leur interprétation.

## 5. Bibliographie


BREIMAN, L., FRIEDMAN, J.H., OLSHEN, R.A. and STONE, C.J. (1984) : *Classification and regression Trees*, C.A :Wadsworth.

CHAVENT, M. (1998) : A monothetic clustering method. *Pattern Recognition Letters, 19, 989-996*.

CHAVENT, M., GUINOT, C., LECHEVALLIER Y. and TENENHAUS, M. (1999) : Méthodes divisives de classification et segmentation non supervisée : recherche d'une typologie de la peau humaine saine. *Revue Statistique Appliquée, XLVII (4), 87-99*.

CHAVENT, M., BRIANT, O. and LECHEVALLIER, Y. (2006) : *DIVCLUS-T : a new descendant hierarchical clustering method*. Internal report U-05-15, Laboratoire de Mathématiques Appliquées de Bordeaux.

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

MORGAN, J.N. and SONQUIST, J.A. (1963) : Problems in the analysis of survey data, and proposal. *J. Aler. Statist. Assoc., 58, 415-434*.

PIRCON, J.-Y. (2004) : *La classification et les processus de Poisson pour de nouvelles méthodes de partitionnement*. Phd Thesis, Facultés Universitaires Notre-Dame de la Paix, Belgium.

SAPORTA, G. (1990), *Probabilités Analyse des données et Statistique*, Editions TECHNIP.

# NP-difficulté de l'approximation Robinsonienne en norme du supremum

**Victor Chepoi, Morgan Seston**

*Laboratoire d'Informatique Fondamentale de Marseille,*
*Faculté de Médecine, Université de la Méditerranée*
*27, Boulevard Jean Moulin,*
*13385 MARSEILLE METZ Cedex 5*
*morgan.seston@laposte.net*

RÉSUMÉ. *Dans le domaine de la classification, il est fondamental d'avoir des outils pour pouvoir représenter des éléments d'un ensemble en fonction d'une matrice de distance (ou plus généralement d'une dissimilarité) définie sur cet ensemble. Ce type de problèmes peut être formulé ainsi : Etant donnée une dissimilarité $d$ définie sur $X$, trouver une dissimilarité $d'$ définie sur $X$ d'un type particulier qui minimise la distance entre $d$ et $d'$ en norme $l_p$ avec $0 \leq p \leq \infty$. Nous nous intéressons ici à l'approximation par des dissimilarités de Robinson en norme $l_\infty$. Nous montrerons que ce problème est NP-difficile.*

MOTS-CLÉS : *Robinson, NP-complétude, approximation, NAE-3-SAT.*

## 1. Introduction

En classification, il est connu qu'à certains types de dissimilarités sont associés des représentations graphiques. Parmi les plus connues, on peut citer les ultramétriques que l'on peut représenter sous forme d'arbres hiérarchiques [Jar 71]. Dans notre cas, nous nous intéresserons aux dissimilarités de Robinson introduites par Robinson [Rob 51], à l'origine pour des problèmes de chronologie en archéologie. Il a été par la suite montré que ces dissimilarités sont une généralisation des ultramétriques et peuvent être représentées par des pseudo-hiérarchies (ou pyramides) faiblement indicées. Mirkin et Rodin [Mir 84] donnent un algorithme en $\mathcal{O}(n^4)$ de reconnaissance des dissimilarités de Robinson, basé sur les algorithmes de reconnaissances des hypergraphes d'intervalles. Plus tard, Chepoi et Fichet [Che 96] donneront un algorithme en $\mathcal{O}(n^3)$ pour le même problème en utilisant une stratégie diviser pour régner. Concernant l'approximation par des dissimilarités de Robinson, il existe peu de résultats. Barthelemy et Brucker [Bar 01] ont montré que le problème d'approximation d'une dissimilarité par une dissimilarité fortement de Robinson en norme $l_p$ ($p < \infty$) est NP-difficile. Les dissimilarités fortement de Robinson sont des cas particuliers des dissimilarités de Robinson que l'on représente par des pseudo-hierarchies indicées. Pour revenir au problème général, nous montrons que le problème d'approximation par une dissimilarité de Robinson est NP-difficile en norme $l_\infty$, pour les autres normes le problème est, à notre connaissance, toujours ouvert.

## 2. Préliminaires

Un ordre total $\prec$ est dit *compatible* pour une dissimilarité $d$, si pour tout $x, y, z$ tels que $x \prec y \prec z$, on a $d(x, z) \geq \max\{d(x, y), d(y, z)\}$ (si un ordre est compatible, son dual l'est aussi). On peut remarquer que si l'on ordonne, selon un ordre compatible, les éléments de la matrice associée à la dissimilarité, alors les valeurs de celle-ci sont croissantes en lignes et en colonnes quand on s'éloigne de la diagonale principale. Une dissimilarité est dite de Robinson si il existe un ordre compatible. De même, que l'on définit la compatibilité entre un ordre et

une dissimilarité, on peut définir l'$\epsilon$-compatibilité. Une dissimilarité $d$ et un ordre $\prec$ sont $\epsilon$-compatibles, si pour tout $x \prec y \prec z$, on a $d(x,z)+2\epsilon \geq \max\{d(x,y), d(y,z)\}$. Et une dissimilarité est $\epsilon$-Robinson si il existe un ordre $\epsilon$-compatible. On peut prouver qu'il est équivalent de définir un ordre $\epsilon$-compatible ainsi : un ordre est $\epsilon$-compatible si il existe une dissimilarité $d'$ compatible pour cet ordre telle que $||d-d'||_\infty \leq 2\epsilon$. Notre problème se définit alors comme suit :

**APPROXIMATION_PAR_ROBINSON** : *Etant donnée une dissimilarité $d$, trouver une dissimilarité de Robinson $d_R$ qui minimise l'erreur $||d-d_R||_\infty$.*

**RECONNAISSANCE_$\epsilon$_ROBINSON** : *Etant donné une dissimilarité $d$, est-ce que $d$ est $\epsilon$-Robinson ?*

Une notion connue de classification qui nous sera utile pour la suite est la notion de sur-dominée. La sur-dominée $d'$ d'une dissimilarité $d$ dans un sous-ensemble $\mathcal{D}'$ de $\mathcal{D}$ est la plus petite dissimilarité de $\mathcal{D}'$, si elle existe, plus grande que $d$ ($d' = \min_{d'' \in \mathcal{D}'}\{d \leq d''\}$). Pour une dissimilarité $d$ et un ordre $\prec$ fixé, il existe une sur-dominée Robinsonienne que nous noterons $d^*_\prec$ ($d^*_\prec = \min_{d' \in \mathcal{R}_\prec}\{d \leq d''\}$ où $\mathcal{R}_\prec$ est l'ensemble des dissimilarités compatibles avec l'ordre $\prec$). De plus, $d^*_\prec$ est facilement calculable, en effet : $d^*_\prec(x,y) = \max\{d(u,v) : x \prec u \prec v \prec y\}$. A partir de la sur-dominée, on peut calculer une dissimilarité optimale $\hat{d}_\prec$ pour un ordre fixé $\prec$ ($\forall\, x, y$ $\hat{d}_\prec(x,y) = d^*_\prec(x,y) - \frac{||d-d^*_\prec||_\infty}{2}$). A partir de ces résultats connus, il est facile de montrer que l'erreur optimale $\hat{\epsilon}$ pour tout ordre appartient à l'ensemble $\Delta = \{\frac{d(x,y)-d(x',y')}{2} : x, y, x', y' \in X\}$. On peut aisément en déduire que indépendamment de l'ordre, l'erreur optimale appartient aussi à $\Delta$.

## 3. Problème similaires

De nombreux problèmes similaires d'approximation d'une dissimilarité par une dissimilarité particulière ont déjà été étudiés et montrés comme étant polynomiaux ou NP-difficiles. Pour ces derniers, il existe des algorithmes d'approximation avec un facteur constant. Pour un problème de minimisation $\Pi$, *un algorithme d'approximation avec un facteur $c \geq 1$* est un algorithme qui pour toute instance $I$ de $\Pi$ s'exécute en temps polynômial dans la taille de $I$ et garantit que le coût de la solution obtenue est inférieure à $c$ fois le coût de la solution optimale [Vaz 01].

D'abord en ce qui concerne les ultramétriques, Farach et al. [Far 93] donnent un algorithme polynômial pour l'approximation d'une dissimilarité $d$ par une ultramétrique $d'$ en norme $l_\infty$. Dans [Aga 96], Agarwala et al. ont montré que l'approximation d'une dissimilarité $d$ par une semi-distance d'arbre est NP-difficile. De plus, ils donnent un algorithme avec un facteur 3 pour ce problème. L'idée de cet algorithme est de fixer un point $p$ comme racine de l'arbre et de placer les autres en respectant leur distance à $p$. Par la suite, Chepoi et Fichet [Che 00] ont présenté une approche unifiée en utilisant les sous-dominantes pour les approximations par une ultramétrique et par une distance d'arbre. Håstad et al. [Has 03] utilisent le même type d'algorithme que Agarwala et al. [Aga 96] pour obtenir un algorithme avec un facteur 2 pour l'approximation d'une dissimilarité par une semi-distance linéaire. En effet pour chaque point $p$, ils construisent l'arrangement tel que la coordonnée de $p$ est égale à 0, et les coordonnées des autres points sont égales à leur distance à $p$. Parmi ces arrangements, ils choisissent celui qui minimise sa distance à $d$. Cet algorithme a un facteur 3. Pour obtenir un facteur 2, ils décalent chaque point d'un $\epsilon$. Pour savoir si un point doit être rapproché ou éloigné de $p$, ils construisent une formule 2-SAT. Håstad et al. [Has 03] ont aussi montré qu'il est NP-difficile d'approcher une dissimilarité par une distance linéaire avec un facteur inférieur à 7/5, et à $2-\delta$ pour $0 < \delta < 1$ tant qu'il est NP-difficile de colorier un graphe 3-coloriable avec $\lceil 4/\delta \rceil$ couleurs.

## 4. NP-Complétude

Nous inspirant de la démonstration de NP-complétude de Håstad et al. [Has 03] pour le problème de l'approximation par une distance unidimensionnelle, nous utilisons une réduction au problème NOT-ALL-EQUAL-3-SAT.

Le *problème* NOT-ALL-EQUAL-3-SAT (NAE-3-SAT) est un problème de satisfaisabilité. Un problème de satisfaisabilité est défini sur un ensemble de variables booléennes $X$ et de clauses $C$. Une clause est une disjonction de

littéraux (un littéral est soit une variable $x$ soit sa négation $\bar{x}$). Dans le cas de 3-SAT, les clauses sont formées de trois littéraux. On dit qu'une instance $(X, C)$ est satisfaisable s'il existe un assignement $A : X \to (vrai, faux)$ tel que chaque clause est vraie. Dans le cas de NAE-3-SAT, on ajoute comme contrainte que pour chaque clause on doit avoir au moins un littéral faux. Le problème NAE-3-SAT est de trouver un assignement satisfaisant l'instance $(X, C)$, et le problème de décision associé est : étant donné une instance $(X, C)$, existe-t-il un assignement satisfaisant ? Dans [Sch 78], Schaefer a montré que ce problème est NP-complet. Nous avons montré que si il existe un algorithme d'approximation avec un facteur $< 3/2$ alors on peut résoudre le problème *NAE-3-SAT* en temps polynomial. Donc tant $P \neq NP$, il n'existe pas d'algorithme d'approximation avec un facteur $< 3/2$ pour notre problème.

A partir une instance $(X, C)$, nous allons construire une dissimilarité $d$ définie sur $P$. L'ensemble $P$ est défini ainsi :

- à toute variable $x$ de $X$, on associe les points $p_x$ et $p_{\bar{x}}$ dans $P$
- a toute clause $c$ de $C$, on associe les point $c_1$, $c_2$, $c_3$ dans $P$.

Enfin on ajoute à l'ensemble $P$, les points $t$ et $f$ pour vrai et faux. Du fait de la dualité des ordres, on peut s'intéresser uniquement aux ordres tels que $t \prec f$. On a une partition des ordres sur $P$ en deux ensembles :

- $\mathcal{O}_1$ : les ordres tels qu'il existe $x$ avec $p_x$ et $p_{\bar{x}}$ du même coté de $t$ ou $f$ (soit $v \in \{t, f\}$ $p_x \prec p_{\bar{x}} \prec v$, $p_{\bar{x}} \prec p_x \prec v$, $v \prec p_x \prec p_{\bar{x}}$ ou $v \prec p_{\bar{x}} \prec p_x$)
- $\mathcal{O}_2$ : les ordres tels que pour tout $x \in X$, on a $p_x$ et $p_{\bar{x}}$ de part et d'autre de $t$ et $f$ ($p_x \prec t \prec f \prec p_{\bar{x}}$ ou $p_{\bar{x}} \prec t \prec f \prec p_x$)

On construit la dissimilarité $d$ à valeurs dans $\{0, 3, 6, 9\}$, donc l'erreur optimale $\hat{\epsilon}$ associée à $d$ appartient à l'ensemble $\{0, 3/2, 3, 9/2\}$. On fixe les valeurs des distances entre les points telles que :

- tout ordre associé à un assignement qui satisfait l'instance $(X, C)$ est 3-compatible
- tout ordre associé à un assignement qui ne satisfait pas l'instance $(X, C)$ est 9/2-compatible
- tout ordre de $\mathcal{O}_1$ est 9/2-compatible.

Donc pour tout ordre $\prec_1$ de $\mathcal{O}_1$, nous fixons les valeurs des distances entre les points associés aux littéraux, $t$ et $f$ telles que $\hat{\epsilon}_{\prec_1} = 9/2$. Si $d(p_x, p_{\bar{x}}) = 9, d(p_x, t) = 0$ et $d(p_{\bar{x}}, p_x) = 0$ alors $\hat{\epsilon}_{\prec_1} \geq 9/2$ (cf. figure 1). Or on a vu que l'erreur optimale $\hat{\epsilon}_{\prec}$ associée à $d$ pour un ordre $\prec$ appartient à l'ensemble $\{0, 3/2, 3, 9/2\}$. On en déduit que l'erreur optimale associée à tout ordre de $\mathcal{O}_1$ est 9/2-compatible.

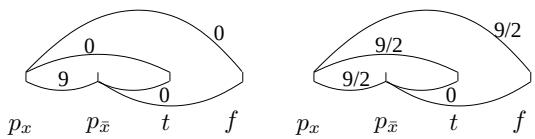



**FIG. 1.** *Un ordre de type $\mathcal{O}_1$ et son approximation.*

Maintenant, on veut que tout ordre $\prec_2$ de $\mathcal{O}_2$, on ait $\hat{\epsilon}_{\prec_2} \geq 3$ pour cela fixons $d(t, f) = 6$ (cf. figure 2).

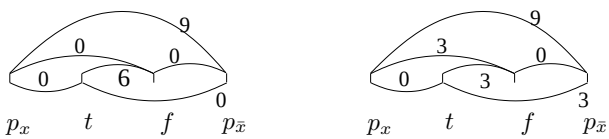



**FIG. 2.** *Un ordre de type $\mathcal{O}_2$ et son approximation.*

Enfin parmi les ordres de $\mathcal{O}_2$, il nous faut distinguer les ordres associés à un assignement satisfaisant, des autres. Les assignements ne satisfaisants pas $(X, C)$, sont tels qu'il existe trois points, associés à trois littéraux appartenant à la même clause, du même côté de $t$ et $f$. Les distances aux points associés aux clauses vont nous permettre de faire la distinction entre ces deux types d'ordres :

– $\mathcal{O}'_1$ : il existe $\{u, v, w\} = c \in C$, telle que $p_u, p_v, p_w$ sont du même coté de $t$ et $f$
– $\mathcal{O}'_2$ :pour tout $\{u, v, w\} = c \in C$, il existe un point appartenant à $\{p_u, p_v, p_w\}$ qui n'est pas du même coté de $t$ et $f$ que les 2 autres.

On finit de construire la dissimilarité $d$ telle que pour tout ordre $\prec'_1$ de $\mathcal{O}'_1$, on ait $\hat{\epsilon}_{\prec'_1} = 9/2$ et pour tout ordre $\prec'_1$ de $\mathcal{O}'_1$, on ait $\hat{\epsilon}_{\prec'_1} = 3$.

Un algorithme d'approximation avec un facteur strictement inférieur à 3/2 appliqué à une dissimilarité, associée à une instance de NAE-3-SAT satisfaisable, renverra un ordre optimal car $\hat{\epsilon} \in \{0, 3/2, 3, 9/2\}$. Pour cette démonstration, on notera que la difficulté se trouve dans le choix des valeurs de la dissimilarité. En effet, une fois les valeurs de la dissimilarité bien choisies, la démonstration est basée sur une étude de cas fastidieuse mais aisée.

## 5. Bibliographie


[Aga 96] R. Agarwala, V. Bafna, M. Farach, B. Narayanan, M. Paterson, et M. Thorup, *On the approximability of numerical taxonomy (fitting distances by tree metrics)*, SIAM Journal on Computing **17** (1999), 1073–1085.

[Bar 01] J.-P. Barthelemy, F. Brucker, *NP-hard approximation problems in overlapping clustering*, Journal of Classification **18** (2001), 159–183.

[Che 96] V. Chepoi et B. Fichet, *Recognition of Robinsonian dissimilarities*, Journal of Classification **14** (1997), 311–325.

[Che 00] V. Chepoi et B. Fichet, *$l_\infty$-approximation via subdominants*, Journal of Mathematical Psychology **44** (2000), no. 4, 600–616.

[Far 93] M. Farach, S. Kannan, et T. Warnow, *A robust model for finding optimal evolutionary trees*, Algorithmica **13** (1995), 155–179.

[Jar 71] N. Jardine et R. Sibson, *Mathematical Taxonomy*, John Wiley & Sons, (1971).

[Has 03] J. Håstad, L. Ivansson, et J. Lagergren, *Fitting points on the real line and its application to RH mapping*, Journal of Algorithms **49** (2003), no. 1, 42–62.

[Mir 84] B. Mirkin et S. Rodin, *Graphs and Genes*, Springer-Verlag, Berlin, 1984.

[Rob 51] W. S. Robinson, *A method for chronologically ordering archaeological deposits*, American Antiquity **16** (1951), 293–301.

[Sax 79] J. B. Saxe, *Embeddability of weighted graphs in k-space is strongly NP-hard*, Proceedings of the 17th Allerton Conference on Communications, Control, and Computing, 1979, pp. 480–489.

[Sch 78] T. J. Schaefer, *The complexity of satisfiability problems*, STOC '78 : Proceedings of the tenth annual ACM symposium on Theory of computing (New York, NY, USA), ACM Press, 1978, pp. 216–226.

[Vaz 01] V. Vazirani, *Approximation Algorithms*, Springer, Berlin, 2001.

# Classification avec recouvrement des classes : une extension des $k$-moyennes

**Guillaume Cleuziou**

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

RÉSUMÉ. *On constate ces dernières années un intérêt croissant pour les méthodes de classification construisant des schémas autorisant le recouvrement des classes. Ces schémas sont, en effet, particulièrement adaptés à certains types de données documentaires ou biologiques par exemple. Malheureusement les quelques techniques existantes s'avèrent peu ou mal adaptées à ces applications. Nous proposons dans cet article d'adapter l'algorithme des $k$-moyennes à cette tâche, en reconsidérant à la fois le critère de qualité et l'algorithme de recherche d'un recouvrement optimal.*

MOTS-CLÉS : *Classification non-supervisée, recouvrement des classes, $k$-moyennes.*

## 1. Introduction

La classification est un domaine de recherches en constante évolution du fait de l'émergence perpétuelle de nouvelles problématiques issues du monde réel. Cette évolution porte autant sur les méthodes employées (centres mobiles, hiérarchies, mélanges de lois, agents artificiels, etc.) que sur l'idéal recherché. Ceux qui cherchaient hier une "organisation en classes compactes et séparées" s'autorisent aujourd'hui à varier les formes, densités, tailles, nombres et agencements des classes, selon le contexte applicatif visé.

Dans cette étude nous nous intéressons à l'agencement des classes entres elles et plus particulièrement aux recouvrements (appelés aussi intersections ou empiétements) entre les classes. Dans certains domaines d'applications, ce type de schéma de classification est naturel. Par exemple en biologie un gène peut influencer plusieurs aspects du métabolisme, en recherche d'information un document (texte, image, vidéo, etc.) peut aborder plusieurs thématiques ou appartenir à plusieurs genres différents, enfin en traitement du langage un mot peut avoir plusieurs interprétations. Dans chacun de ces domaines, restreindre chaque élément à n'appartenir qu'à une seule classe entraîne une perte d'information potentiellement utile à l'utilisateur mais également cruciale pour le processus de classification.

Les approches de classification aboutissant à des classes recouvrantes n'abondent pas dans la littérature ; deux approches de référence nécessitent cependant d'être mentionnées ici : la classification pyramidale [DID 84] et la classification floue [BEZ 81]. Pourtant ces deux approches sont peu utilisées dans les domaines d'application cités précédemment car peu adaptées. Les pyramides autorisent des recouvrement trop limités et entre des classes nécessairement assez similaires tandis que la classification floue oblige à effectuer, a posteriori, un choix d'affectation peu évident, à partir des fonctions d'appartenance. L'affectation plus ou moins arbitraire des individus à des classes pré-construites est un raisonnement classificatoire que l'on rencontre dans d'autres méthodes plus récentes [LEL 93, PAN 03, CLE 04].

Nous considérerons dans cette étude que les recouvrements entre classes ne doivent pas être simplement tolérés mais, au contraire, que le processus de construction des classes doit les intégrer et en tirer profit. De même que les pyramides étendent les hiérarchies en autorisant les recouvrements, nous proposons d'adapter la recherche d'une bonne partition à celle d'un bon recouvrement. Dans cette perspective, nous présentons une extension des $k$-moyennes [MAC 67], fondée d'une part sur l'adaptation de la fonction objective à optimiser et d'autre part sur l'algorithme de recherche d'un recouvrement optimal.

## 2. Algorithme de classification avec recouvrement des classes

### 2.1. *L'algorithme des $k$-moyennes*

Étant donné un ensemble d'individus $X = \{x_1, x_2, \ldots, x_n\}$ définis dans $\mathbb{R}^p$ muni d'une métrique euclidienne $d$, l'algorithme des $k$-moyennes est fondé sur la recherche d'une partition $\mathcal{I} = \{I_1, I_2, \ldots, I_k\}$ de $X$ minimisant le critère de variance intra-classes :

$$V(\mathcal{I}) = \sum_{j=1}^{k} \sum_{\{x_i \in I_j\}} p_i d^2(x_i, c_j) \quad [1]$$

où $c_j$ désigne le centre de la classe $I_j$ et $p_i$ la masse relative à l'individu $x_i$ (traditionnellement chaque individu est pondéré de façon uniforme, avec $\sum p_i = 1$). Les deux étapes le l'algorithme des $k$-moyennes qui consistent à (1) affecter chaque individu au centre de classe le plus proche et (2) mettre à jour le centre de chaque classe en calculant son centre de gravité, permettent d'assurer la convergence du critère $V(.)$ vers une partition stable[1]. On remarquera que l'optimisation du critère de variance intra-classes ne tolère (et encore moins ne favorise) aucun recouvrement de classes. En effet, chaque affectation supplémentaire d'un individu $x_i$ à une classe $I_j$ impliquerait une augmentation du critère $V(.)$ de la quantité $p_i d^2(x_i, c_j)$. Nous proposons de modifier la fonction objective utilisée, de façon à autoriser l'affectation de chaque individu à une ou plusieurs classes.

### 2.2. *Une autre interprétation de la fonction objective*

Résumer une collection d'individus à travers un ensemble de classes permet une analyse globale des données mais suppose en même temps de concéder une partie de l'information contenue dans ces données. La fonction objective $V(.)$, étudiée précédemment, peut alors être interprétée comme un critère mesurant l'information perdue ou encore l'erreur commise en substituant chaque individu à un centre (ou représentant) de classe[2]. Dans la suite nous parlerons d'*image* d'un individu pour désigner ce substitut.

**Définition 2.1** *Soient une collection de classes $\mathcal{I} = \{I_1, I_2, \ldots, I_k\}$ formant une* **partition** *de l'ensemble d'individus $X = \{x_1, x_2, \ldots, x_n\}$ et $c_1, c_2, \ldots, c_k$ les centres respectifs des classes de $\mathcal{I}$, l'image de $x_i$ (notée $\overline{x}_i$) dans la classification est donnée par le centre $c_j$ de la classe $I_j$ à laquelle $x_i$ est affecté.*

Dans la cas où $\mathcal{I}$ n'est plus une partition mais un recouvrement de $X$, la définition 2.1 doit être étendue. Considérant que l'affectation d'un individu $x_i$ à plusieurs classes se justifie par le fait que $x_i$ partage des propriétés avec chacune de ces classes, l'image de $x_i$ doit résulter d'un compromis entre tous les centres de classes concernées.

**Définition 2.2** *Soient une collection de classes $\mathcal{I} = \{I_1, I_2, \ldots, I_k\}$ formant un* **recouvrement** *de l'ensemble d'individus $X = \{x_1, x_2, \ldots, x_n\}$ et $c_1, c_2, \ldots, c_k$ les centres respectifs des classes de $\mathcal{I}$, l'image de $x_i$ (notée $\overline{x}_i$) dans la classification est donnée par le centre de gravité[3] de l'ensemble $\{c_j | x_i \in I_j\}$.*

Par la définition 2.2, la fonction objective $V(.)$ peut être réécrite de manière à favoriser les recouvrements de classes lorsque ceux-ci permettent de capturer d'avantage d'information sur les individus concernés. On note $\overline{V}(\mathcal{I})$ ce critère évaluant la qualité d'un recouvrement $\mathcal{I}$ d'un ensemble d'individus $X$ :

$$\overline{V}(\mathcal{I}) = \sum_{x_i \in X} p_i d^2(x_i, \overline{x}_i) \quad [2]$$

### 2.3. *Algorithme de recherche d'un recouvrement optimal*

Une manière naïve de rechercher le recouvrement qui minimise le critère $\overline{V}(.)$ serait de générer tous les recouvrements possibles. Cette solution est exclue dans le cas de partitions du fait qu'il existe de l'ordre de $k^n$ partitions

1. La partition stable obtenue correspond à une situation où l'algorithme ne peut plus évoluer. Nous parlerons dans la suite d'"optimum local" pour évoquer cette situation.
2. Dans le critère $V(.)$ l'erreur commise pour un individu $x_i$ affecté à la classe $I_j$ vaut $d^2(x_i, c_j)$
3. Dans cette définition, les classes contenant $x_i$ sont pondérées de façon uniforme.

en $k$ classes pour $n$ individus. Le nombre de recouvrements possibles étant beaucoup plus grand encore, il convient de proposer un algorithme permettant d'explorer partiellement l'espace des possibilités au risque d'aboutir à un optimum seulement "local".

L'algorithme que nous présentons (Figure 1) s'inspire largement de l'algorithme des $k$-moyennes et procède par itérations de deux étapes :

1. l'affectation des individus aux centres les plus proches,
2. le calcul des nouveaux centres des classes.

Dans cet algorithme, l'heuristique d'affectation (AFFECTER($x_i$,$C^t$)) consiste à affecter d'abord $x_i$ au centre le plus proche, puis à considérer les autres centres - du plus proche au plus éloigné - en effectuant l'affectation tant que $d(x_i, \overline{x}_i)$ décroît.

**Initialisation :** ($t$=0) choisir aléatoirement $k$ centres $C^t = \{c_1^t, c_2^t, \ldots, c_k^t\}$ dans $X$,
Pour chaque $x_i \in X$ : AFFECTER($x_i$,$C^t$),
en déduire un recouvrement initial $\mathcal{I}^t = \{I_1^t, I_2^t, \ldots, I_k^t\}$.

FAIRE ($t$=$t$+1)

**Calcul des nouveaux centres $C^t$ :** pour $j$ allant de 1 à $k$ calculer le nouveau centre $c_j^t$ de $I_j$,

**Affectation :** Pour chaque $x_i \in X$ : AFFECTER($x_i$,$C^t$),
en déduire un nouveau recouvrement $\mathcal{I}^t$,

TANT QUE $\mathcal{I}^t \neq \mathcal{I}^{t-1}$

**FIG. 1.** *Algorithme (simplifié) de recherche d'un recouvrement optimal.*

L'algorithme simplifié, tel que présenté en Figure 1, nécessite quelques précisions afin d'en assurer la convergence (cf. Section 2.4). Tout d'abord concernant l'affectation des individus : après recalcul des centres, il est possible que les centres les plus proches d'un individu aient changés sans toutefois qu'une nouvelle affectation conduise à une meilleur image ; dans ce cas nous choisirons de conserver l'ancienne affectation.

La seconde précision porte sur le calcul des centres de classes. Dans cet algorithme, le centre $c_j$ d'une classe $I_j$ correspond au centre de gravité du nuage de points $N_j = \{(x_i, p_i) | x_i \in I_j\}$ où les masses $p_i$ associées à chaque individu sont données par :

$$p_i = \begin{cases} 0 & \text{si } d(\overline{x}_{i|j}, c_{j,i}) = 0, \\ \eta . d^2(\overline{x}_{i|j}, x_i) / d^2(\overline{x}_{i|j}, c_{j,i}) & \text{sinon.} \end{cases} \quad [3]$$

Dans cette expression, $\overline{x}_{i|j}$ désigne l'image partielle de $x_i$, *i.e* le centre de gravité de l'ensemble $\{c_l | x_i \in I_l \; et \; j \neq l\}$. Le terme $c_{j,i}$ désigne le centre $c_j$ idéal pour permettre à $x_i$ de "coller à son image" ($d(x_i, \overline{x_i}) = 0$) et $\eta$ est un coefficient normalisateur ($\sum p_i = 1$). Dans la suite on notera également $\overline{x}_i^b$ et $\overline{x}_i^a$ les images de $x_i$ respectivement avant (*b*efore) et après (*a*fter) la mise à jour du centre $c_j$.

Notons que dans le cas d'une partition (chaque individu appartient à une seule classe), on se ramène au calcul classique d'un centre de gravité avec des points tous de même masse $\eta$ puisque $c_{j,i} = x_i$.

### 2.4. *Convergence de l'algorithme*

Nous donnons dans cette section les principaux éléments visant à démontrer la convergence de l'algorithme à travers la décroissance du critère $\overline{V}(.)$. Chaque itération de l'algorithme est composée de deux étapes : l'affectation des individus aux centres et le recalcul des centres. Concernant l'étape d'affectation, nous avons précisé que pour chaque individu $x_i$, une nouvelle affectation n'est pas retenue si elle fait augmenter la quantité $d(x_i, \overline{x_i})$ ; et donc indirectement le critère $\overline{V}(.)$. On peut également montrer que ce critère décroît lors de l'étape de recalcul des

centres et plus précisément pour chaque centre recalculé. Étant donné une classe $I_j^t$, on note d'après le théorème de Huygens :

$$\sum_{x_i \in I_j^t} p_i.d^2(c_{j,i}, c_j^t) = \sum_{x_i \in I_j^t} p_i.d^2(c_{j,i}, c_j^{t+1}) + \sum_{x_i \in I_j^t} p_i.d^2(c_j^t, c_j^{t+1}) \quad [4]$$

On peut de plus montrer géométriquement (Thalès) l'égalité suivante :

$$\frac{d^2(\overline{x}_{i|j}, x_i)}{d^2(\overline{x}_{i|j}, c_{j,i})} = \frac{d^2(x_i, \overline{x}_i^b)}{d^2(c_j^t, c_{i,j})} = \frac{d^2(x_i, \overline{x}_i^a)}{d^2(c_j^{t+1}, c_{i,j})} \quad [5]$$

En remplaçant dans [4] les $p_i$ par leur définition donnée en [3] puis en utilisant l'égalité [5] on obtient :

$$\sum_{x_i \in I_j^t} d^2(x_i, \overline{x}_i^b) = \sum_{x_i \in I_j^t} d^2(x_i, \overline{x}_i^a) + T \qquad \text{(avec } T \geq 0) \quad [6]$$

ce qui montre que l'étape de calcul des nouveaux centres fait décroître (strictement[4]) $\overline{V}(.)$. La convergence de l'algorithme est donc assurée du fait de la décroissance (stricte) du critère $\overline{V}(.)$ sur un ensemble fini de recouvrements.

## 3. Discussion et conclusion

Dans la méthode exposée, l'heuristique d'affectation choisie n'est certainement pas de nature à minimiser la fonction objective. Il s'agit simplement d'en assurer la décroissance en essayant[5] de respecter la propriété suivante : *chaque individu doit être affecté aux centres les plus proches*. Étant donnés $k$ centres de classes, il serait en effet possible d'extraire un schéma d'affectation optimal au sens du critère $\overline{V}(.)$ mais qui ne vérifierait pas la contrainte pourtant indispensable pour assurer la cohérence des classes.

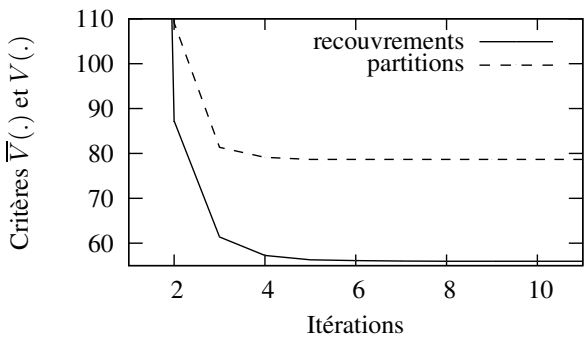



**FIG. 2.** *Convergences des critères $\overline{V}(.)$ et $V(.)$ dans les mêmes conditions initiales.*

| Étiquettes \ Classes | 1 | 2 | 3 |
|---|---|---|---|
| Iris Setosa | | | 50 |
| Iris Versicolor | 21 | 50 | 5 |
| Iris Virginica | 48 | 24 | |

**FIG. 3.** *Matrice de confusion.*

Afin d'illustrer notre approche, nous présentons une première expérimentation de l'algorithme sur la base Iris (UCI repository). Ce jeu de donné, traditionnellement utilisé en classification supervisée, disposerait d'une organisation en trois classes de 50 individus chacune, dont deux classes sont réputées difficilement séparables (classes des Iris Versicolor et Virginica). Les résultats présentés correspondent à la meilleure classification obtenue (relativement à $\overline{V}(.)$) sur 20 exécutions de l'algorithme.

On observe via la matrice de confusion proposée en Figure 3 que chaque classe d'Iris extraite par notre approche correspond à une classe prédéfinie : la classe 1 s'identifie principalement aux Iris Virginica, la classe 2 aux Iris

4. Dans [6], $T$ est strictement positif sauf dans le cas où aucun centre n'a été modifié.
5. L'algorithme proposé ne permet pas de satisfaire totalement à cette contrainte. Cependant l'heuristique d'affectation limite en pratique ces violations, qui ne remettent donc pas en cause la cohérence globale des classes.

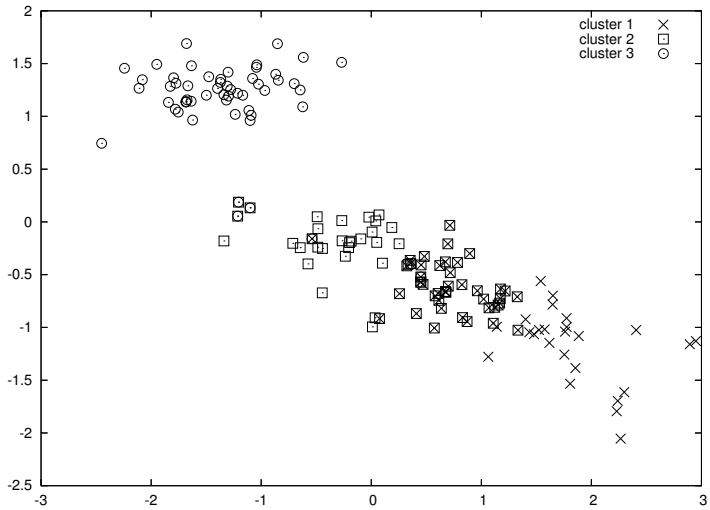



**FIG. 4.** *Visualisation des classes par projection sur les deux premiers vecteurs propres (ACP).*

Versicolor et la dernière classe presque exclusivement aux Iris Setosa. De plus, les intersections mentionnées entre les deux premières classes obtenues indiquent effectivement la confusion qui existe entre les deux classes Versicolor et Virginica. La visualisation proposée en Figure 4 atteste de la cohérence globale des classes construites mais également des recouvrements entre ces classes.

Enfin, le graphique de la Figure 2 expose sur un exemple pratique que la vitesse de convergence de l'algorithme proposé est du même ordre que pour la méthode des $k$-moyennes.

Pour conclure, nous rappelons que nous avons traité dans cet article du problème de la classification en classes recouvrantes. Constatant qu'il n'existe pas de méthodologie clairement établie pour ce problème, nous avons présenté une première approche dans laquelle la construction de classes recouvrantes est utilisée dans le processus même de classification afin d'améliorer la représentativité des classes relativement aux données initiales.

Les études à venir sur ce thème de recherche consisteront notamment à proposer une méthode d'affectation des individus conciliant à la fois les contraintes de cohérence des classes et d'amélioration du critère de qualité du recouvrement.

## 4. Bibliographie

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

[CLE 04] CLEUZIOU G., MARTIN L., VRAIN C., PoBOC : an Overlapping Clustering Algorithm. Application to Rule-Based Classification and Textual Data, R. LÓPEZ DE MÁNTARAS AND L. SAITTA, IOS PRESS, Ed., *Proceedings of the 16th European Conference on Artificial Intelligence*, Valencia, Spain, August 22-27 2004, p. 440-444.

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

[LEL 93] LELU A., Modèles neuronaux pour l'analyse de données documentaires et textuelles, Thèse de doctorat, mars 1993, Université de Paris VI.

[MAC 67] MACQUEEN J., Some methods for classification and analysis of multivariate observations, *Proceedings of the Fifth Berkeley Symposium on Mathematical statistics and probability*, vol. 1, Berkeley, 1967, University of California Press, p. 281-297.

[PAN 03] PANTEL P., Clustering by Committee, Ph.d. dissertation, 2003, Department of Computing Science, University of Alberta.

# Un algorithme efficace pour les cartes auto-organisatrices de Kohonen appliquées aux tableaux de dissimilarités

**Brieuc Conan-Guez†, Fabrice Rossi‡, Aïcha El Golli‡**

†*LITA EA3097, Université Paul Verlaine -Metz, Île du Saulcy, 57045 METZ CEDEX 1*
*brieuc.conan-guez@univ-metz.fr*
‡*Projet AxIS, INRIA, Domaine de Voluceau, Rocquencourt, B.P. 105, 78153, Le Chesnay Cedex, France*
{*fabrice.rossi,aicha.elgolli*}*@inria.fr*

*RÉSUMÉ. Dans cet article, nous proposons une nouvelle implémentation d'une adaptation des cartes auto-organisatrices de Kohonen (SOM) aux tableaux de dissimilarités : nous proposons tout d'abord une modification de l'algorithme afin d'obtenir une réduction importante de son coût théorique, puis nous introduisons certaines techniques d'implémentation qui permettent de réduire de manière importante son temps de calcul effectif. Une propriété importante de ces diverses modifications est que les résultats obtenus sont strictement identiques à ceux de l'algorithme initial.*

*MOTS-CLÉS : Tableau de dissimilarités, Carte auto-organisatrice de Kohonen, Implémentation efficace.*

## 1. Introduction

Dans beaucoup d'applications réelles, les individus étudiés ne peuvent pas être décrits efficacement par des vecteurs numériques : on pense par exemple à des données de tailles variables (séquences de protéines), ou bien à des données (semi-)structurées (graphes, arbres, documents XML). Une solution possible pour traiter de telles données est de s'appuyer sur une mesure de dissimilarité permettant de comparer les individus deux à deux.

Une variante des cartes auto-organisatrices de Kohonen (SOM, [KOH 01]) adaptée aux tableaux de dissimilarités a été proposée dans [KOH 98, KOH 02] (voir section 2). Cet algorithme souffre malheureusement d'un temps de calcul beaucoup plus élevé que celui du SOM classique. Dans ce travail, nous proposons tout d'abord une modification de cet algorithme afin d'obtenir une réduction importante de son coût théorique (section 3), puis nous introduisons certaines techniques d'implémentation qui permettent de réduire de manière importante son temps de calcul. Les résultats produits par ce nouvel algorithme sont strictement identiques à ceux de l'algorithme initial. Enfin, afin de quantifier l'amélioration des temps de calcul, la section 4 est consacrée à des expériences menées sur des données simulées.

## 2. Cartes auto-organisatrices de Kohonen adaptées aux tableaux de dissimilarités

Nous rappelons dans cette section l'adaptation du SOM aux tableaux de dissimilarités (DSOM, [KOH 98, KOH 02, GOL 04]). On considère $N$ individus $\mathcal{D} = (\mathbf{x}_i)_{1 \leq i \leq N}$ appartenant à un espace arbitraire $\mathcal{X}$, muni d'une mesure de dissimilarité $d$. On considère un SOM à $M$ modèles (ou neurones), qui sont numérotés de 1 à $M$. Le modèle $j$ est associé à un élément de $\mathcal{D}$, noté $\mathbf{m}_j$ (pour chaque modèle $j$, il existe $i$ tel que $\mathbf{m}_j = \mathbf{x}_i$). $\mathbf{m}_j$ est appelé le prototype du modèle $j$, et on note $\mathcal{M} = (\mathbf{m}_1, \ldots, \mathbf{m}_M)$. On munit l'ensemble des modèles d'une métrique $\delta$ qui correspond à la structure *a priori* imposée par le SOM. Le but du SOM est alors double : quantifier correctement l'ensemble $\mathcal{D}$ en trouvant de bons prototypes, organiser ces prototypes dans $\mathcal{D}$ afin qu'ils respectent la topologie

induite par $\delta$ sur les modèles. On note $K$ une fonction noyau décroissante (par exemple $K(x) = \exp(-x^2)$), et $h(j,k) = K(\delta(j,k))$ la fonction de voisinage.

Le SOM adapté aux tableaux de dissimilarités (DSOM) est un algorithme itératif où les prototypes et la fonction de voisinage évoluent à chaque itération $l$ : on note donc $\mathbf{m}_j^l$ et $h^l(.,.)$ ces quantités. La décroissance de $h^l$ à chaque itération assure la convergence de l'algorithme. L'algorithme du DSOM, qui est basée sur la version non stochastique du SOM (version *batch*) commence par une *phase d'initialisation* : on choisit par exemple $\mathcal{M}^0 = (\mathbf{m}_1^0, \ldots, \mathbf{m}_M^0)$ de manière aléatoire. Puis l'algorithme alterne les *phases d'affectation* et les *phases de représentation* jusqu'à la convergence. Pour l'itération $l$, on a :

1. *phase d'affectation* : chaque individu $\mathbf{x}_i$ est affecté à son modèle gagnant $c^l(i)$ en appliquant la règle d'affectation usuelle $c^l(i) = \arg\min_{j\in\{1,\ldots,M\}} d(\mathbf{x}_i, \mathbf{m}_j^{l-1})$. L'inconvénient de ce type d'affectation est qu'il ne permet pas de lever les ambiguités relatives aux prototypes (cas relativement fréquent où deux modèles distincts sont associés au même prototype). Dans notre implémentation, nous utilisons donc la modification proposée dans [KOH 98]. Finalement, on note $\mathcal{C}_j^l$ la classe associée au modèle $j$ à l'itération $l$. Les $\mathcal{C}_j^l$ forment une partition de $\mathcal{D}$.

2. *phase de repésentation* : l'algorithme calcule de nouvelles valeurs pour les prototypes (i.e. $\mathcal{M}^l$). Chaque prototype $\mathbf{m}_j^l$ est solution du problème :

$$\mathbf{m}_j^l = \arg\min_{\mathbf{m}\in\mathcal{D}} \sum_{i=1}^{N} h^l(c^l(i), j) d(\mathbf{x}_i, \mathbf{m}). \qquad [1]$$

## 3. Une implémentation efficace du DSOM

Si l'on examine le coût algorithmique du DSOM, on constate que pour une itération de l'algorithme, le coût de la phase d'affectation est en $O(NM^2)$ (voir la règle d'affectation modifiée dans [KOH 98]) et que la phase de représentation (voir équation 1) est en $O(N^2M)$[1]. Comme $N > M$ dans tous les cas, la phase de représentation domine nettement le calcul. Les modifications apportées au DSOM dans cette section vont permettre dans un premier temps de réduire la complexité théorique de la phase de représentation, puis, grâce à une implémentation efficace de l'algorithme, de réduire son temps d'exécution.

### 3.1. *Les sommes partielles*

En analysant le problème de minimisation proposé dans l'équation 1, on constate que sa structure peut être grandement simplifiée. A l'itération $l$ et pour chaque modèle $j$, on cherche à trouver pour quel $k$, $S^l(j,k)$ est minimum, où $S^l(j,k) = \sum_{i=1}^{N} h^l(c^l(i), j) d(\mathbf{x}_i, \mathbf{x}_k)$. Si l'on note $D^l(u,k) = \sum_{i\in\mathcal{C}_u^l} d(\mathbf{x}_i, \mathbf{x}_k)$, on peut exprimer $S^l(j,k)$ d'une manière plus simple :

$$S^l(j,k) = \sum_{u=1}^{M} h^l(u,j) \sum_{i\in\mathcal{C}_u^l} d(\mathbf{x}_i, \mathbf{x}_k) = \sum_{u=1}^{M} h^l(u,j) D^l(u,k). \qquad [2]$$

Il y a $MN$ différentes valeurs $D^l(u,k)$, qui peuvent être pré-calculées une fois pour toute avant la phase de représentation. Le coût de cette phase de pré-calcul est en $O(N^2)$. En effet, calculer la somme partielle $D^l(u,k)$ coûte $O(|\mathcal{C}_u^l|)$ (où $|\mathcal{C}_u^l|$ est l'effectif de $\mathcal{C}_u^l$). Puis calculer tous les $D^l(u,k)$ pour un $u$ fixé coûte $O(N|\mathcal{C}_u^l|)$. Comme $\sum_{u=1}^{M} |\mathcal{C}_u^l| = N$, le coût total est en $O(N^2)$. Selon l'équation 2, le calcul de $S^l(j,k)$ peut être effectué en $O(M)$ opérations. Comme cela doit être fait pour tous les $k$ et pour tous les $j$, le coût est en $O(NM^2)$. On a donc un coût total pour la phase de représentation de $O(N^2 + NM^2)$, à comparer avec $O(N^2M)$ pour l'algorithme initial. Comme on a $N > M$ dans toutes les situations, cette approche réduit le coût du DSOM. De plus, les résultats obtenus sont strictement identiques à ceux de l'algorithme initial.

1. à comparer avec $O(nNM)$ dans le cas du SOM *batch* classique sur des vecteurs de dimension $n$

### 3.2. *L'arrêt prématuré*

Dans la section précédente, les modifications apportées à l'algorithme du DSOM ont permis de réduire sa complexité algorithmique, dans les sections suivantes, nous allons montrer comment grâce à quelques techniques d'implémentation, il est possible d'accélérer son temps d'exécution de manière importante.

Lors de la phase de représentation, à l'itération $l$ et pour chaque modèle $j$, on cherche à calculer le minimum des $S^l(j,k)$. Ce calcul est réalisé informatiquement au sein d'une boucle en $k$, que nous appelerons la boucle externe. Cette boucle externe effectue deux opérations à chacune de ses itérations : premièrement elle calcule grâce à une seconde boucle (boucle interne) la valeur de $S^l(j,k)$ pour un $k$ donné (voir la somme en $u$ apparaissant à la droite de l'équation 2), puis elle compare le résultat trouvé avec le meilleur résultat calculé lors des itérations prédédentes. Afin d'améliorer l'efficacité de ce calcul, une idée simple consiste à déplacer cette étape de comparaison dans la boucle interne : le calcul de la boucle interne est arrêté prématurément dès lors que la sous-somme $\Sigma_k = \sum_{u=1}^{m'} h^l(u,j)D^l(u,k)$ calculée à l'itération $m'$ (avec $m' < M$) est supérieure au minimum.

Afin de favoriser la stratégie d'arrêt prématuré, la boucle interne ainsi que la boucle externe doivent être ordonnées. Pour la boucle interne, l'ordre optimal serait d'accroître $\Sigma_k$ le plus rapidement possible en sommant d'abord les grandes valeurs de $h^l(u,j)D^l(u,k)$. Pour la boucle externe, l'ordre optimal serait de commencer avec les valeurs faibles de $S^l(j,k)$ (i.e. avec un bon candidat pour le prototype du modèle $j$) : une faible valeur de $S^l(j,k)$ arrêtera la boucle interne plus tôt qu'une grande valeur. En pratique cependant, calculer ces ordres optimaux est extrêmement coûteux. Nous utiliserons donc des ordres efficaces mais non optimaux qui sont induits par la topologie du DSOM. La définition de $h^l$ implique que $h^l(u,j)$ est petit quand $\delta(u,j)$ est grand. Il est donc raisonnable d'ordonner la boucle interne en $u$ dans l'ordre décroissant des $h^l(u,j)$, i.e. dans l'ordre croissant des $\delta(u,j)$. Pour la boucle externe, nous utilisons les propriétés d'organisation du DSOM : les individus sont affectés à la classe du prototype le plus proche. Donc, la qualité *a priori* d'un individu $\mathbf{x}_k$ comme prototype pour le modèle $j$ est plus ou moins l'inverse de la distance $\delta$ entre le modèle $j$ et le modèle représentant la classe de $\mathbf{x}_k$. On ordonne donc la boucle externe dans l'ordre croissant de $\delta$. Ces divers optimisations ne modifie pas les résultats produits.

### 3.3. *La mémorisation*

Une autre source d'optimisation vient du fait que les classes $\mathcal{C}_u^l$ produites par le DSOM ont tendance à se stabiliser lors des dernières itérations de l'algorithme. Il n'est pas rare que d'une itération à l'autre, le contenu d'une (ou plusieurs) classe reste strictement identique. Dans de tels cas, les $N$ valeurs $D^l(u,k)$ correspondantes restent inchangées, et il est donc inutile de les recalculer. La mise en œuvre informatique de cette optimisation peut facilement être réalisée en associant à chaque classe $\mathcal{C}_u$ une variable booléenne. Lors de la phase d'affectation, cette variable est positionnée si l'on constate un changement de classe. Les $D^l(u,k)$ correspondants seront alors recalculés lors de la phase de représentation.

Bien que cette stratégie de mémorisation se révèle être très efficace, son mode d'action est un peu grossier. En effet, le calcul complet des $D^l(u,k)$ pour deux valeurs de $u$ (i.e., pour deux classes) peut être déclenché par le changement de classes d'un unique individu. Il semble donc intéressant de chercher une solution plus fine. Considérons en effet le cas où la classe ne subit qu'une seule modification : l'individu $\mathbf{x}_i$. On a alors :

$$D^l(c^{l-1}(i),k) = D^{l-1}(c^{l-1}(i),k) - d(\mathbf{x}_i,\mathbf{x}_k) \quad [3]$$

$$D^l(c^l(i),k) = D^{l-1}(c^l(i),k) + d(\mathbf{x}_i,\mathbf{x}_k) \quad [4]$$

Appliquer ces formules de mise-à-jour induit $2N$ additions et $N$ affectations. Si plusieurs individus se déplacent de leur ancienne classe à leur nouvelle classe, les opérations de mise-à-jour doivent être effectuées pour chaque déplacement. Dans le cas extrême où tous les individus changent de classes, le nombre total d'additions serait de $2N^2$ (et $N^2$ affectations). Ce coût de remise-à-jour est donc supérieur à un recalcul total ($N^2$ additions et $N^2$ affectations). Ceci implique que pour un nombre de modifications de classes inférieur à approximativement $\frac{N}{2}$, l'approche par mise-à-jour est plus efficace que le recalcul total. Il est impotant de noter que dans le cas où le recalcul total est nécessaire, la stratégie de mémorisation présentée au paragraphe ci-dessus peut être utilisée.

## 4. Expériences

Les différentes optimisations ont été évaluées sur un jeu de données simulées. Il consiste en $N$ points de $\mathbb{R}^2$ choisis aléatoirement et uniformément dans le carré unité. $\mathbb{R}^2$ est muni de la distance euclidienne (les dissimilarités sont calculées avant le lancement de l'algorithme). La topologie du DSOM est une grille hexagonale munie de sa distance de graphe $\delta$. On choisit $L = 100$ itérations et la fonction de voisinage est gaussienne.

| $N$ (nb d'individus)<br>$M$ (nb de modèles) | 500 | 1 000 | 1 500 | 2 000 | 3 000 |
|---|---|---|---|---|---|
| $49 = 7 \times 7$ | 11.4 / 0.8 | 53.5 / 2.3 | 135.4 / 4.8 | 261.6 / 8.5 | 865.3 / 22.5 |
| $100 = 10 \times 10$ | 24.7 / 2.4 | 115.0 / 5.6 | 283.4 / 9.8 | 557.0 / 15.3 | 1757.0 / 32.8 |
| $225 = 15 \times 15$ | | 313.7 / 30.4 | 806.6 / 46.4 | 1594.8 / 63.3 | 4455.5 / 105.3 |
| $400 = 20 \times 20$ | | | 1336.9 / 136.1 | 2525.2 / 179.1 | 7151.8 / 264.6 |

**TAB. 1.** *Temps d'exécution en secondes pour l'algorithme standard / pour l'algorithme avec sommes partielles*

Le tableau 1 donne les performances[2] de l'algorithme standard et de l'algorithme basé sur les sommes partielles. Pour l'algorithme standard, le coût est quadratique en $N$ et relativement linéaire en $M$. L'amélioration apportée par les sommes partielles est impressionante. Le rapport entre les deux algorithmes est approximativement proportionnel à $\frac{NM}{N+M^2}$, qui est le rapport théorique (la phase d'affectation n'est pas prise en compte).

| $N$ (nb d'individus)<br>$M$ (nb de modèles) | 500 | 1 000 | 1 500 | 2 000 | 3 000 |
|---|---|---|---|---|---|
| $49 = 7 \times 7$ | 1.14 / 1.6 | 1.05 / 1.92 | 1.00 / 2 | 0.97 / 2.02 | 0.98 / 2.39 |
| $100 = 10 \times 10$ | 1.41 / 1.85 | 1.33 / 1.81 | 1.23 / 1.66 | 1.15 / 1.65 | 1.08 / 1.83 |
| $225 = 15 \times 15$ | | 2.27 / 2.87 | 2.13 / 2.67 | 2.00 / 2.57 | 1.78 / 2.43 |
| $400 = 20 \times 20$ | | | 2.74 / 3.04 | 2.75 / 3.2 | 2.48 / 2.89 |

**TAB. 2.** *Taux d'accélération pour l'algorithme avec arrêt prématuré et ordre sans / avec mémorisation (référence : algorithme avec sommes partielles)*

Le tableau 2 donne les performances des algorithmes avec arrêt prématuré et ordre avec ou sans mémorisation. Ces expériences montrent clairement que chacune des optimisations proposées apporte un gain important en terme de temps d'exécution.

## 5. Conclusions

Nous avons proposé dans ce travail un nouvel algorithme pour le SOM appliqué aux tableaux de dissimilarités. Cet algorithme permet une réduction importante du coût théorique. De plus, certaines techniques d'implémentation ont permis d'accélérer d'un facteur 3 le temps d'exécution sous des conditions favorables.

## 6. Bibliographie


[GOL 04] GOLLI A. E., CONAN-GUEZ B., ROSSI F., Self-Organizing Map and symbolic data, *Journal of Symbolic Data Analysis*, vol. 2(1), 2004.

[KOH 98] KOHONEN T., SOMERVUO P. J., Self-Organizing Maps of symbol strings, *Neurocomputing*, vol. 21, 1998, page 19.

[KOH 01] KOHONEN T., *Self-Organizing Maps*, vol. 30, Springer Series in Information Sciences, 1995,1997,2001.

[KOH 02] KOHONEN T., SOMERVUO P. J., How to make large Self-Organizing Maps for nonvectorial data, *Neural Networks*, vol. 15(8), 2002, page 945.


2. L'algorithme a été implémenté en Java (JRE 5.0 de Sun) sur un PC équipé d'un processeur Pentium IV (3 GHz) sous le système d'exploitation Linux

# Classification visuelle et interactive de données en utilisant des points d'intérêt

**David Da Costa**[†⋆], **Gilles Venturini**[†]

[†]*Laboratoire d'Informatique, Ecole Polytechnique de l'Université de Tours*
*64, Avenue Jean Portalis, 37200 Tours, France.*
{*david.dacosta, venturini*}*@univ-tours.fr*

[⋆]*Agicom, Institut d'Etudes*
*3, degrés Saint Laumer, 41000 Blois, France.*
*ddacosta@agicom.fr*

RÉSUMÉ. *L'objectif de notre travail est de pouvoir représenter visuellement de grands ensembles de données et de laisser l'expert du domaine procéder interactivement à la définition d'une classification de ces données. Notre approche se base sur l'existence d'une fonction de similarité afin de traiter des données de tous types (numériques, symboliques, images, textes, etc) et permet à l'expert du domaine, grâce à l'utilisation d'une visualisation à base de points d'intérêt (POIs), de définir des classes dans les données. Nous comparons notre approche interactive avec la classification ascendante hiérarchique (CAH) sur un ensemble de bases classiques.*

MOTS-CLÉS : *Classification visuelle, interaction, points d'intérêt (POI).*

## 1. Introduction

La classification est une des tâches importantes de la fouille de données et a pour but d'affecter à chaque donnée d'un ensemble un label de classe. La grande majorité des méthodes dans ce domaine utilise des approches non interactives dans lesquelles un algorithme produit des résultats. Cependant, on peut noter que l'expert final est le seul à pouvoir valider les résultats, et il arrive même que certaines méthodes (telles que la CAH) nécessite l'intervention de l'utilisateur pour définir de manière fiable un "bon" nombre de classes. La méthode que nous proposons dans cet article se place dans la lignée des méthodes visuelles et interactives qui vont directement solliciter l'expert du domaine et obtenir ainsi un résultat directement validé par l'utilisateur. A titre d'exemple, nous pouvons citer les arbres de décision [BRE 84] [ANK 99] [SHA 96] [ALS 98] qui sont construits en divisant à plusieurs reprises l'ensemble des données en sous-ensembles disjoints et où une classe est affectée à chaque feuille de l'arbre. Des approches de classification visuelle ont été développées dans ce cadre comme StarClass [3rd03] ou encore PaintingClass [ACM03]. StarClass est une méthode visuelle et interactive de classification : elle visualise des données multidimensionnelles en utilisant les "star coordinates" [KAN 00] et l'utilisateur peut agir sur une des dimensions pour créer un arbre de décision. Elle est cependant limitée à des petits ensembles de données

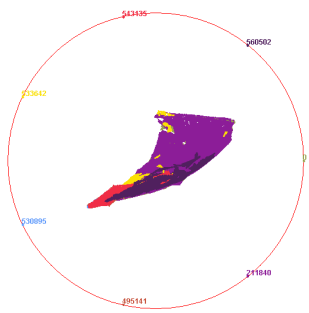
De manière similaire, le but de notre travail est de construire interactivement une classification visuelle de données numériques ou symboliques. Pour ce faire nous définissons une nouvelle méthode de visualisation basée sur les points d'intérêt à partir de laquelle l'utilisateur pourra définir des groupes de points.

## 2. Utilisation des points d'intérêt pour la classification

Le principe de notre méthode consiste, à l'instar de systèmes de recherche documentaire tels que Sqwid [MCC 97] ou Radviz [HOF 99], à placer des données appelées point d'intérêt (POIs) sur un cercle et à représenter les autres données en fonction de leur similarité à ces POIs. Chaque donnée se place sur le barycentre des POIs pondérés par la similarité entre cette donnée et chaque POI [DAC 06]. Par exemple, nous avons visualisé la base de données supervisée *Forest CoverType* [BLA 98] (voir figure 1) qui contient un total de 581012 données, 54 attributs et se compose de 7 classes. Lorsque l'on travaille ainsi avec des données supervisées, notre méthode place le premier individu de chaque classe comme point d'intérêt initial et positionne les données restantes. L'utilisateur peut ensuite réaliser de nombreuses opérations interactives. Il peut changer les POIs (en ajouter, en enlever, définir des POIs qui ne soient pas des données mais des hypothèses à tester, etc). Il peut également effectuer des zooms de différentes natures, détecter des points isolés et obtenir des informations sur ces points (voir [DAC 06] pour plus de détails). Cette visualisation est très efficace également en ce qui concerne les temps d'affichage : elle peut afficher $n$ données (et donc $n^2$ similarité) mais en calculant seulement un nombre linéaire de similarité (entre les données et les $k$ POIs). Cela nous permet de traiter des ensembles ayant jusqu'à 1 million de données en moins d'une minute sur un PC standard.

Dans cet article nous considérons des données non supervisées. Pour le choix initial des points d'intérêt, notre méthode propose $k$ données choisies aléatoirement et place les données restantes comme précédemment. L'utilisateur peut ensuite modifier ces POIs de manière interactive, en choisissant par exemple les centres des nuages observés. Il peut enlever des POIs ou en rajouter de manière à obtenir un résultat visuellement satisfaisant (par exemple des groupes compacts et bien séparés). Nous rappelons ici que les données sont de tous types (pas seulement numériques) et qu'il n'est par conséquent pas possible d'appliquer des méthodes telles que les KMeans pour définir ces points (dans la conclusion nous proposons à ce titre des perspectives). Une fois les données visualisées correctement, l'utilisateur peut sélectionner des sous-ensembles de données (avec la souris) et définir ainsi un label de classe pour toutes les données. Il construit ainsi une classification des données, comme représenté sur la figure 2.

**FIG. 1.** *Visualisation de la base de données Forest CoverType (le temps d'affichage est de 91 secondes sur un PC standard).*

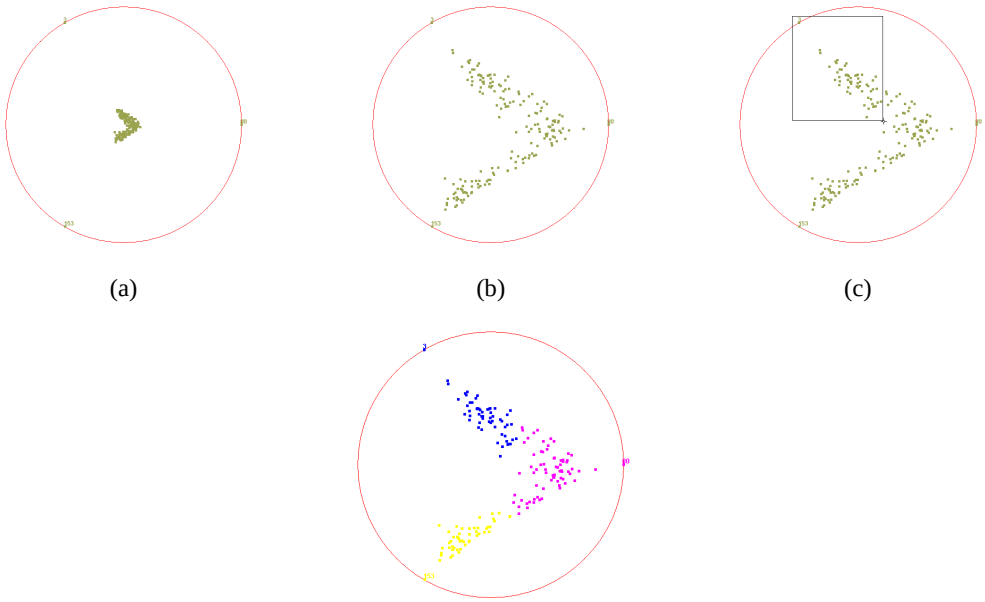


## 3. Expérimentations

Afin de déterminer l'efficacité relative de notre méthode, nous allons évaluer la classification obtenue à la fois en terme de nombre de classes trouvées $C_T$ et de pureté des classes $P_R$.

Pour une classe donnée, la pureté représente le pourcentage de données bien classées et s'obtient à partir d'une matrice de confusion. Elle se calcule donc à partir de la classe réelle de chaque donnée. En effet, pour chacune des classes trouvées nous cherchons la cardinalité du groupe issu de la classe réelle qui est la plus représentée parmi les données de la classe trouvée considérée. Ainsi la somme de toutes les cardinalités pour toutes les classes trouvées divisée par le nombre total de données représente la valeur de la pureté $P_R$.

**FIG. 2.** *Exemple d'une classification visuelle et interactive de données sur la base "WINE". L'utilisateur obtient une visualisation satisfaisante (en (a) et (b)) à l'aide d'opérations interactives, puis sélectionne des données (en (c)) et obtient une classification (en (d)).*



(a) (b) (c)

(d)

Nous comparons les résultats de notre méthode à ceux obtenus avec la CAH [LAN 67][SNE 73]. Nous avons réalisé nos tests sur 7 bases numériques et symboliques. Chaque test consiste à visualiser les données, puis à regrouper les points comme indiqué dans la section précédente. Une fois la classification calculée, nous l'évaluons. Les résultats sont présentés dans le tableau 1.

**TAB. 1.** *Résultats obtenus par la CAH et et notre méthode (POI) sur des bases de données numériques et symboliques.*

| Base de données | # Nombre de données | # Dimensions | # Classes réelles | Type | POI | | CAH | |
|---|---|---|---|---|---|---|---|---|
| | | | | | $C_T$ | $P_R$ | $C_T$ | $P_R$ |
| IRIS | 150 | 4 | 3 | Numérique | 3 | 0,89 | 3 | 0,88 |
| PIMA | 768 | 8 | 2 | Numérique | 2 | 0,69 | 3 | 0,65 |
| SOYBEAN | 47 | 35 | 4 | Numérique | 4 | 0,98 | 6 | 1,00 |
| THYROID | 215 | 5 | 3 | Numérique | 3 | 0,89 | 5 | 0,84 |
| VEHICLE | 846 | 18 | 4 | Numérique | 3 | 0,88 | 3 | 0,35 |
| WINE | 178 | 12 | 3 | Numérique | 3 | 0,81 | 6 | 0,84 |
| HAYES ROTH | 132 | 5 | 3 | Mixte | 3 | 0,41 | 4 | 0,42 |

A partir des résultats obtenus sur le nombre de classes trouvées et sur la pureté, on constate que la CAH obtient des résultats souvent moins bons que notre méthode interactive. Notre méthode se basant sur une visualisation interactive facilite l'étiquetage des nuages de points représentés au centre de notre cercle. Pour certaines données

comme *Pima*, il y a beaucoup plus de difficulté à étiqueter les données (ces données sont connues pour être très fortement bruitées, une caractéristique que l'on retrouve également avec la CAH qui obtient de mauvaises performances). D'autres bases sont également plus difficiles que d'autres à classifier. C'est le cas par exemple de la base *Vehicle*. Ces bases correspondent à des cas où les classes sont mal séparées les unes des autres, où les données sont superposées, une propriété que notre méthode détecte très facilement. Pour cette dernière base, le fait de couper le dendogramme au niveau du saut maximum du critère de Ward engendre de mauvais résultats pour la CAH.

## 4. Conclusion

Nous avons proposé dans cet article une nouvelle méthode de classification non supervisée qui est interactive et qui s'appuie sur une visualisation des données. L'intéret de cette méthode est de faire intervenir directement l'expert du domaine qui peut ainsi valider les résultats obtenus sans avoir à interpréter les sorties d'une méthode automatique. Elle peut s'appliquer à de grands ensembles de données. Sur les bases testées, nous avons pu observer que le temps passé à classifier manuellement les données est très court et tout à fait acceptable par un utilisateur (moins d'une minute). Nous pensons pouvoir étendre notre méthode en proposant des améliorations : par exemple, nous souhaitons proposer des POIs initiaux de meilleure qualité en choisissant automatiquement des données situées plutôt au centre des nuages, données détectables par le fait que leur similarité aux autres données est plus élevée en moyenne. Nous travaillons également à généraliser cette méthode en 3D et à utiliser du matériel de réalité virtuelle (écran stéréoscopique pour la visualisation et capteurs 3D pour les interactions).

## 5. Bibliographie


[3rd03] The 3rd SIAM International Conference on Data Mining, *StarClass : Interactive Visual Classification Using Star Coordinates*, 2003.

[ACM03] ACM KDD 2003 Conference, *PaintingClass : Interactive Construction, Visualization and Exploration of Decision Trees*, 2003.

[ALS 98] ALSABTI K., RANKA S., SINGH V., CLOUDS : A Decision Tree Classifier for Large Datasets, *Knowledge Discovery and Data Mining*, 1998, p. 2-8.

[ANK 99] ANKERST M., ELSEN C., ESTER M., KRIEGEL H.-P., Visual Classification : An Interactive Approach to Decision Tree Construction., *KDD*, 1999, p. 392-396.

[BLA 98] BLAKE C., MERZ C., UCI Repository of machine learning databases, 1998.

[BRE 84] BREIMAN L., FRIEDMAN J. H., OLSHEN R. A., STONE C. J., *Classification and Regression Trees*, Belmont, Wadsworth, 1984.

[DAC 06] DA COSTA D., VENTURINI G., Visualisation interactive de données avec des points d'intérêt, *Actes de la 18ème conférence francophone sur l'Interaction Homme-Machine, IHM'06*, AFIHM, ACM Press, 18–21 2006, p. XX–XX.

[HOF 99] HOFFMAN P., GRINSTEIN G., PINKNEY D., Dimensional anchors : a graphic primitive for multidimensional multivariate information visualizations, *NPIVM '99 : Proceedings of the 1999 workshop on new paradigms in information visualization and manipulation in conjunction with the eighth ACM internation conference on Information and knowledge management*, New York, NY, USA, 1999, ACM Press, p. 9–16.

[KAN 00] KANDOGAN E., Star Coordinates : A Multidimensional Visualization Technique with Uniform Treatment of Dimensions, 2000.

[LAN 67] LANCE G., WILLIAMS W., A general theory of classificatory sorting strategies : I. hierarchical systems, *Computer journal*, vol. 9, n$^o$ 4, 1967, p. 373-380.

[MCC 97] MCCRICKARD S., KEHOE C., Visualizing Search Results using SQWID, *Proceedings of the Sixth International World Wide Web Conference*, April 1997.

[SHA 96] SHAFER J. C., AGRAWAL R., MEHTA M., SPRINT : A Scalable Parallel Classifier for Data Mining, VIJAYARAMAN T. M., BUCHMANN A. P., MOHAN C., SARDA N. L., Eds., *VLDB'96, Proceedings of 22th International Conference on Very Large Data Bases, September 3-6, 1996, Mumbai (Bombay), India*, Morgan Kaufmann, 1996, p. 544-555.

[SNE 73] SNEATH P. H., SOKAL R. R., *Numerical Taxonomy*, W.H. Freeman, San Francisco, 1973.

# Nouvelle méthode de classification adaptée aux données de grandes dimensions : application aux données de biopuces

**Doulaye Dembélé**

*IGBMC, CNRS-IMSERM-ULP*
*1 rue Laurent Fries, BP 10142*
*Parc d'Innovation*
*67404 Illkirch Cedex, France*
*doulaye@titus.u-strasbg.fr*

RÉSUMÉ. *Nous proposons dans cet article une nouvelle méthode de classification adaptée aux données de grandes dimensions. Pour ces données la distance de Chebyshev semble intéressante, de plus elle nécessite moins de temps de calcul comparée à la distance Euclidienne, plus utilisée en raison de ses bonnes propriétés géométriques. La méthode proposée combine les méthodes de regroupement hiérarchique et par partition pour obtenir le nombre de classes dans les données. Des données issues d'expériences de biopuces sont utilisées pour illustrer les performances de la méthode proposée.*

MOTS-CLÉS : *classification, K-Means, distance de Chebyshev, biopuces*

## 1. Introduction

Nous proposons dans ce papier une nouvelle méthode de classification adaptée aux données de grandes dimensions. Pour ces données le problème d'espace vide est connu [DOH 00]. D'autres faiblesses sont liées à l'utilisation de la distance Euclidienne, parmi celles-ci signalons le fait que les distances entre toutes les paires de points des données ont tendance à être identiques quand la dimension augmente [DEM 94]. Dans ces conditions il sera impossible de discriminer les classes, s'il y en a, dans les données. Il est aussi montré dans [HER 02] qu'en augmentant l'ordre de la métrique de Minkowski, il est possible d'augmenter la dimension de la matrice des distances des données prises deux à deux. Le maximum de dimension pour la matrice des distances sera obtenu pour un ordre infini, c'est-à-dire en utilisant la distance de Chebyshev. Notons que l'ordre de la matrice des distances défini le degré de contraste permettant d'obtenir des classes dans les données. La distance de Chebyshev semble alors intéressante, de plus elle nécessite moins de temps de calcul comparée à la distance Euclidienne la plus utilisée en raison de ses bonnes propriétés géométriques. Ceci n'est pas négligeable pour les données de grande dimension comme celles générées par les biologistes dans le cadre de l'étude de l'expression des gènes à l'aide de la technologie de biopuces.

Nous nous intéressons ici aux méthodes de classification non paramétriques et non supervisées ou clustering. Ces méthodes peuvent être regroupées dans deux grandes familles : les méthodes hiérarchiques et les méthodes par partition. Les méthodes hiérarchiques ne nécessitent pas la connaissance *a priori* du nombre de classes dans les données. Leur résultat est représenté sous forme d'un arbre ou dendrogramme dans lequel les branches contiennent les échantillons similaires du point de vue du critère utilisé pour les construire [EVE 93, JAI 88, JAI 00]. Cette méthode est intéressante mais elle ne permet pas de re-examiner un échantillon déja placé dans une branche. Les méthodes par partition consistent à trouver le meilleur regroupement des $N$ échantillons des données en $K$ classes de manière à optimiser un critère de qualité défini *a priori*. Pour résoudre ce problème combinatoire, on utilise dans la pratique une heuristique pour avoir une solution en un temps raisonnable. Dans cette heuristique, les échantillons sont initialement reparties en $K$ classes puis itérativement, on recherche la meilleure combinaison

locale qui améliore la qualité du critère prédéfini en changeant la classe de certains échantillons. Cette étape prend fin quand il n'y a plus d'amélioration possible du critère [EVE 93, JAI 88, JAI 00]. Le principal inconvénient des méthodes par partition est la nécessité de connaître *a priori* le nombre de classes à former. Dans cet article, nous proposons une nouvelle stratégie qui combine les méthodes hiérarchiques et par partition. Dans un premier temps la matrice des distances des données est calculée et utilisée pour former un nombre maximum de classes (étape par partition sans connaissance *a priori* du nombre des classes), puis un regroupement est effectué pour réduire le nombre des classes (étape hiérarchique ascendante avec critère de validation). L'idée de combiner les méthodes de classification hiérarchique et par partition n'est pas nouvelle, voir par exemple [WON 82]. Toutefois la procédure présentée dans le paragraphe suivant est originale.

Nous présentons la méthode de classification proposée puis des résultats obtenus avec des données de biopuces sont ensuite présentés.

## 2. Nouvelle méthode de classification

La méthode de clustering non paramétrique par partition la plus utilisée est la méthode K-Means. Soient $K$ le nombre des classes à trouver, $\mathbf{c}_k$ le centre de la classe $k$ ($k = 1, 2, \ldots, K$) et $\mathbf{x}_i$ l'échantillon $i$ des données. La méthode K-Means permet d'obtenir la répartition des données après la minimisation de la fonction suivante :

$$J(\mathbf{c}_k) = \sum_{k=1}^{K} \sum_{i=1}^{N} u_{ik} d(\mathbf{x}_i, \mathbf{c}_k) \qquad [1]$$

où $d(\mathbf{x}_i, \mathbf{c}_k)$ désigne la distance entre l'échantillon $\mathbf{x}_i$ et le centre $\mathbf{c}_k$ de la classe $k$ alors que $u_{ik}$ vaut 1 si l'échantillon $\mathbf{x}_i$ appartient à la classe $k$ et 0 sinon.

À partir d'une partition initiale, la fonction (1) est itérativement améliorée en changeant la classe des échantillons, jusqu'à l'obtention d'une solution stable. Dans la relation (1), il y a deux paramètres à choisir avant le début des calculs, la distance $d(.,.)$ et le nombre $K$ des classes.

La distance Euclidienne est la plus utilisée à cause de ses bonnes propriétés géométriques. Elle définit toutefois implicitement une forme sphérique pour les classes à trouver. Pour obtenir des classes de forme ellipsoïdale, la distance de Mahalanobis est utilisée. Pour les données de grande dimension la distance de Chebyshev qui est équivalente à la métrique de Minkowski à l'ordre infini offre plus de contraste dans la matrice des distances [HER 02]. C'est pour cette raison que notre choix s'est porté sur la distance de Chebyshev.

Étant donné qu'il est souvent difficile de connaître *a priori* le nombre des classes, nous nous proposons de les déterminer directement à partir des données. L'idée dans la fonction (1) est de former des classes telles que les échantillons membres d'une classe soient plus proches que ceux d'une autre classe. Cette proximité peut être définie par un seuil sur les distances [LUK 01]. Soit $d_{seuil}$ cette distance seuil, toutes les distances des échantillons d'une classe doivent être alors plus petites que $d_{seuil}$. Le problème revient alors à déterminer ce seuil à partir des distances des données. La recherche du seuil est examinée plus loin. À partir d'un seuil approprié sur les distances, nous répartissons les données pour obtenir une valeur maximale pour $K$, puis un regroupement de certaines classes est enfin effectué. La procédure de classification proposée se résume comme suit :

1. Calculer toutes les distances entre les échantillons des données,
2. Rechercher un seuil pour les distances des échantillons d'une classe,
3. Repartir les données en utilisant le seuil trouvé,
4. Déterminer le nombre maximum $K_{max}$ de classes sans prendre en compte les classes singletons,
5. Calculer les $K_{max}$ centres des classes et les utiliser pour avoir une partition initiale des données,
6. Utiliser une méthode de regroupement et un critère de validation pour obtenir $K$ classes.

La première étape consiste à calculer les $\frac{N(N-1)}{2}$ distances des $N$ échantillons. La médiane de ces distances est utilisée pour obtenir le seuil recherché (étape 2). Ce seuil est utilisé dans la troisième étape conjointement avec

les distances pour affecter un index à chaque échantillon. Cela est fait en comparant le premier échantillon aux autres, puis le second échantillon non indexé est comparé aux autres, et ceci est poursuivi jusqu'à l'avant dernier échantillon. Le même index est associé aux échantillons de distances inférieures ou égales au seuil. L'examen des différents index permet enfin d'obtenir $K_{max}$ (étape 4). La cinquième étape consiste à calculer les centres des classes retenues puis à répartir les échantillons entre celles-ci. Dans la dernière étape, une méthode hiérarchique ascendante peut être utilisée. Il est également possible d'utiliser l'algorithme K-Means dans lequel le nombre de classes est décroissant.

### 2.1. *Conditionnement des données*

Avant d'utiliser un algorithme de classification, les données sont souvent standardisées. Cela consiste en général à transformer les données pour avoir une moyenne nulle et un écart type égal à un pour chaque échantillon. La standardisation est particulièrement utile pour les données de biopuces qui peuvent avoir des amplitudes très différentes alors que l'on est intéressé par la variation des profils. La transformation ci-dessus rend toutefois les données sphériques. Une autre transformation peut consister à ramener toutes les valeurs des données entre $0$ et $1$. Cela est obtenu en otant la valeur minimale de chaque valeur et en divisant le résultat par l'étendue, c'est-à-dire, la différence entre les valeurs maximale et minimale.

### 2.2. *Détermination du seuil des distances*

Le nombre maximum $K_{max}$ dépend de la valeur de $d_{seuil}$. Si cette valeur est élevée, $K_{max}$ sera faible et inapproprié, inversement si la valeur de $d_{seuil}$ est faible, $K_{max}$ sera élevé et générera beaucoup de classes singletons. Après des tests sur des données synthétiques, la médiane des distances fournit un bon compromis si les données sont transformées pour avoir des valeurs comprises entre $0$ et $1$. Nous utilisons cette solution heuristique pour $d_{seuil}$ en attendant les résultats d'autres études sur le sujet.

### 2.3. *Critère d'arrêt*

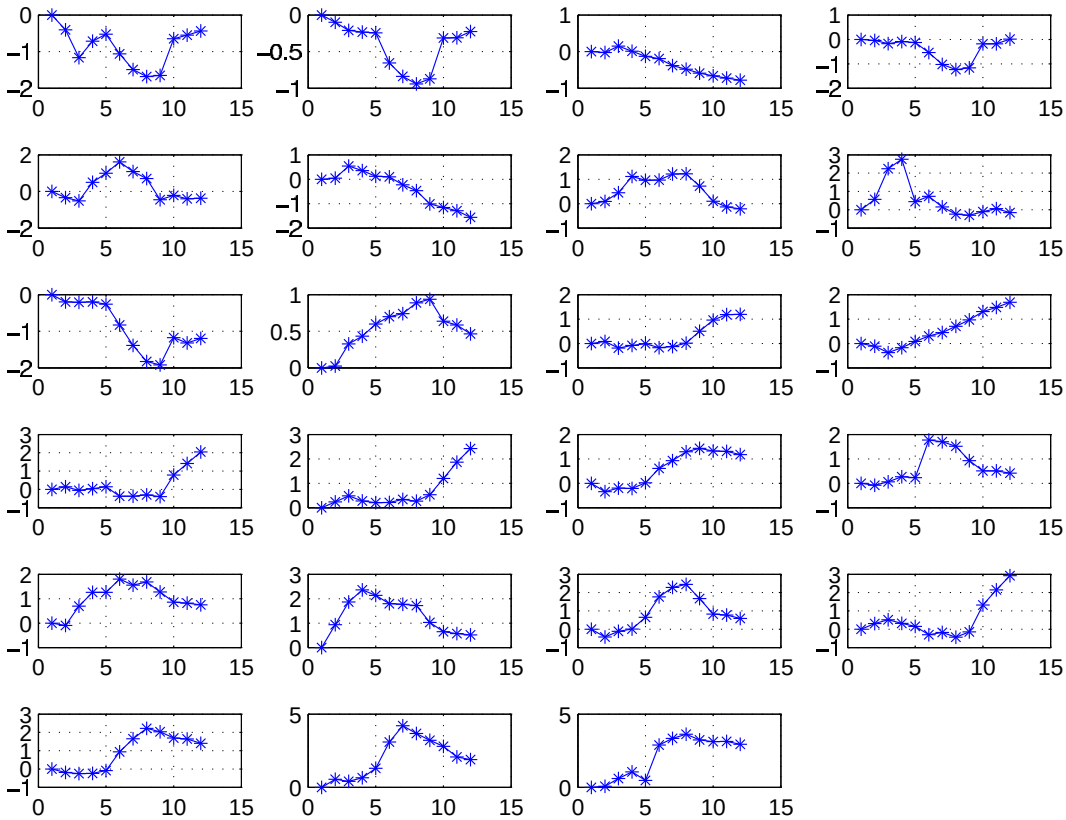
Dans l'étape de réduction du nombre des classes, les critères comparés dans [MIL 85] peuvent être utilisés. En utilisant la transformation $[0,1]$ des données, des échantillons de profils identiques peuvent être affectés à des classes différentes en fonction de leur valeur absolue moyenne. Pour regrouper les classes, nous utilisons l'information de leur forme. Ceci permet d'identifier les classes de profils voisins. Nous définissons et utilisons la co-variation des profils comme le coefficient de corrélation des écarts non centrés observés pour les valeurs successives de chaque profil (pour plus de détails, voir la version longue de ce papier).

## 3. Résultats

Pour illustrer les performances de la méthode proposée, nous avons utilisé les données de biopuces. Les biopuces sont de petits supports (lames de verre) sur lesquels des milliers de séquences d'ADN correspondant chacune à un gène sont attachées à des adresses connues (spots). L'ARN des échantillons à analyser est marqué avec une molécule fluorescente puis hybridé (par appariement entre séquences d'ADN complémentaires) sur les biopuces. Les biopuces sont ensuite scannées. Le niveau d'expression des gènes est représenté par une intensité de fluorescence. La quantification de l'image (mesure de l'intensité de fluorescence pour chacun des spots) fournit des données numériques qui servent à l'analyse.

Nous nous sommes servi des donnés d'une étude de réponses de fibroblastes humains à des concentrations de sérum au cours du temps [IYE 99]. Nous avons utilisé les données correspondant à une sélection de $517$ gènes. Ces données peuvent être recupérées à l'adresse suivante : *http ://www.sciencemag.org/feature/data/984559.shl*.

La valeur de distance seuil obtenue pour ces données est $d_{seuil} = 0.223$. En utilisant cette valeur, nous avons obtenu un nombre maximum de classes $K_{max} = 23$. Les profils de ces classes sont représentés sur la figure 1. La



*23 classes obtenues dans la phase initiale. Chaque profil est identifié par un numero et le nombre des échantillons qui la forme, e.g. la première classe $C1$ contient 22 échantillons ($N = 22$).*

matrice des co-variations des 23 profils initiaux des données de sérum a ensuite été calculée. Cette matrice (non présentée ici) montre des coefficients de co-variation élevés, $\geq 0.75$, pour les classes 13, 14 et 20 qui ont des profils similaires. Le coefficient de co-variation est la plus petite, $-0.87$, pour les classes 6 et 12 montrant une opposition de profils pour ces classes. La figure 1 contient tous les profils obtenus dans [IYE 99] avec des redondances. Avec la méthode hiérarchique ascendante les classes ont été regroupées.

## 4. Conclusion

Une nouvelle méthode de classification de données est présentée. La distance de Chebyschev est utilisée. Cette distance semble plus appropriée pour les données de grandes dimensions. La stratégie proposée consiste dans un premier temps à rechercher un nombre maximum de classes dans les données. Ceci est fait après examen de la matrice des distances des données. Puis une réducton du nombre de classes est effectuée. Pour cela une méthode hiérarchique ascendante peut être utilisée. La méthode K-Means qui offre la possibilité de re-affecter un échantillon à une autre classe peut être aussi utilisée. Pour la standardisation des données, la méthode qui permet de ramener toutes les valeurs entre 0 et 1 a été utilisée. Un travail future consiste à étudier le choix du seuil $d_{seuil}$ des distances d'une classe. Une interface conviviale pourra également faciliter l'exploitation des résultats de classification.

**Remerciements**

Merci à Bernard Jost qui a pris le temps de lire cet article. Je remercie également les rapporteurs pour leurs remarques. Ce travail a bénéfié du soutien du Centre National de la Recherche Scientifique (CNRS), de l'Institut National de la Recherche Médicale (INSERM), de l'Hôpital Universitaire de Strasbourg et du Centre National de Recherche en Génomique (CNRG).

## 5. Bibliographie

[DEM 94] DEMARTINES P., Analyse de données par réseaux de neurones auto-organisés, PhD thesis, TIRF, INPG, Grenoble, France, novembre 1994.

[DOH 00] DOHONO D. L., High-Dimensional Data Analysis : The Curses and Blessings of Dimensionality, *Am. Math. Soc. Conf. "Math Challenges of the 21st Century", Los Angeles, www-stat.stanford.edu/˜donoho*, 2000.

[EVE 93] EVERITT B. S., *Cluster Analysis*, Arnold, London, 3rd édition, 1993.

[HER 02] HERAULT J., GUÉRIN-DUGUÉ A., VILLEMAIN P., Searching for the Embedded Manifolds in High-Dimensional Data, Problems and Unsolved Questions, *SANN'2002 Proceedings - European Symposium on Artificial Neural Networks 24-26 April, Bruges, Belgium*, 2002, p. 173-184.

[IYE 99] IYER V. R., EISEN M. B., ROSS D. T., SCHULER G., MOORE T., LEE J. C. F., TRENT J. M., STAUDT L. M., JR J. H., BOGOSKI M. S., LASHKARI D., SHALON D., BOTSTEIN D., BROWN P. O., The Transcriptional Program in the Response of Human Fibroblast to Serum, *Science*, vol. 283, 1999, p. 83-87.

[JAI 88] JAIN J. K., DUBES R. C., *Algorithms for Clustering Data*, Prentice Hall, Englewood Cliff, New Jersey, 1988.

[JAI 00] JAIN A. K., DUIN R. P. W., MAO J., Statistical Pattern Recognition : A Review, *IEEE trans. PAMI*, vol. 22, n$^o$ 1, 2000.

[LUK 01] LUKASHIN A. V., FUCHS R., Analysis of Temporal Gene Expression Profiles : Clustering by Simulated Annealing and Determining the Optimal Number of Clusters, *Bioinformatics*, vol. 17, n$^o$ 5, 2001, p. 405-414.

[MIL 85] MILLIGAN G. W., COOPER M. C., An Examination of Procedures for Determining the Number of Clusters in a Data Set, *Psychometrika*, vol. 50, n$^o$ 2, 1985, p. 159-179.

[WON 82] WONG M. A., A Hybrid Clustering Method for Identifying High-Density Clusters, *Journal of American Statistical Association*, vol. 77, n$^o$ 380, 1982, p. 841-847.

# Classification hiérarchique de variables discrètes basée sur l'information mutuelle en pré-traitement d'un algorithme de sélection de variables pertinentes

**Hélène Desmier** [ab] **& Ivan Kojadinovic** [a] **& Pascale Kuntz** [a]

[a] *LINA CNRS FRE 2729, Site Polytech Nantes, rue Christian Pauc, 44306 Nantes, France*
[b] *PerformanSe SAS, Atlanpole La Fleuriaye, 44470 Carquefou, France*
{*prenom.nom*}*@polytech.univ-nantes.fr*

RÉSUMÉ. *Notre travail se situe dans le contexte de la sélection de variables pertinentes pour les problèmes de discrimination caractérisés par un grand nombre de variables potentiellement discriminantes toutes discrètes ou nominales. Dans ce cadre, nous proposons une procédure de sélection de variables pertinentes fondée sur une troncature $k$-additive de l'information mutuelle et utilisant une classification ascendante hiérarchique des variables potentiellement discriminantes afin de réduire le nombre de sous-ensembles dont la pertinence est estimée.*

MOTS-CLÉS : *Sélection de variables, classification de variables, information mutuelle, troncature $k$-additive.*

## 1. Introduction

Le problème de la sélection de variables en discrimination se rencontre généralement lorsque le nombre de variables pouvant être utilisées pour expliquer la classe d'un individu est très élevé. Le rôle de la procédure de sélection de variables consiste alors à sélectionner un sous-ensemble de variables *potentiellement discriminantes* permettant d'expliquer la classe de façon optimale. La nécessité de ce traitement préalable est essentiellement due au fait que, généralement, l'utilisation d'un nombre de variables discriminantes trop élevé dans un modèle de discrimination détériore grandement sa capacité de *généralisation* et la compréhension de la relation modélisée.

D'un point de vue structurel, une procédure de sélection de variables peut être vue comme composée de deux éléments fondamentaux [LIU 98] : une *mesure de pertinence*, utilisée pour mesurer l'*influence* d'un sous-ensemble de variables potentiellement discriminantes sur la variable qualitative à expliquer, et un *algorithme de recherche*, dont le rôle est de *parcourir* l'ensemble des sous-ensembles de variables à la recherche d'un sous-ensemble *optimal* ou *sous-optimal* au sens de la mesure de pertinence.

Dans le cadre de ce travail, nous nous intéressons au cas où les variables potentiellement discriminantes sont toutes discrètes ou nominales et nous proposons une procédure *filtre* [LIU 98] utilisant une *troncature $k$-additive de l'information mutuelle* [KOJ 05] comme mesure de pertinence. L'utilisation de l'information mutuelle en tant que mesure de pertinence a déjà été considérée à de nombreuses reprises dans la littérature (voir e.g. [HUT 05]). L'approximation que nous utilisons permet d'approcher la pertinence d'un ensemble de variables à partir des pertinences de ses sous-ensembles de faible cardinal. Afin d'éviter d'avoir à parcourir la totalité des sous-ensembles non vides de l'ensemble des variables potentiellement discriminantes ou d'avoir à recourir à des heuristiques souvent trop sous-optimales du type *sélection pas à pas*, nous proposons d'effectuer, en pré-traitement de la sélection de variables, une classification ascendante hiérarchique de l'ensemble des variables potentiellement discriminantes afin d'en identifier la *structure*.

## 2. Sélection de variables fondée sur une troncature $k$-additive de l'information mutuelle

Nous considérons, dans la suite, qu'un problème de sélection de variables se présente sous la forme d'un ensemble de variables aléatoires potentiellement discriminantes $\aleph := \{X_1, \ldots, X_m\}$ et d'un vecteur aléatoire $\overrightarrow{\mathbb{Y}}$ à expliquer. Comme indiqué précédemment, nous nous limitons au cas où toutes ces variables sont discrètes et prennent un nombre fini de valeurs ou sont nominales. Les sous-ensembles de $\aleph$ sont notés par des majuscules doubles, e.g. $\mathbb{X}$, et, étant donné un sous-ensemble $\mathbb{X} \subseteq \aleph$ de $r$ variables aléatoires, $\overrightarrow{\mathbb{X}}$ désigne un vecteur aléatoire de dimension $r$ dont les composantes sont des éléments distincts de $\mathbb{X}$.

Dans ce contexte probabiliste, il semble naturel de mesurer la pertinence des sous-ensembles de $\aleph$ à l'aide d'une *mesure de dépendance* [JOE 89]. En d'autres termes, nous considérerons qu'un sous-ensemble $\mathbb{X}$ non vide de $\aleph$ est d'autant plus pertinent que le degré de *dépendance fonctionnelle* entre les vecteurs aléatoires $\overrightarrow{\mathbb{X}}$ et $\overrightarrow{\mathbb{Y}}$ est élevé. La mesure de dépendance à utiliser peut par exemple être choisie parmi les mesures d'*écart à l'indépendance*. Nous avons opté pour l'*information mutuelle* (voir e.g. [JOE 89, COV 91]) en raison de ses propriétés intéressantes et de son lien avec l'entropie de Shannon.

En pratique, nous ne disposons pas de la distribution de probabilité de $(X_1, \ldots, X_m, \overrightarrow{\mathbb{Y}})$ mais uniquement de $n$ réalisations indépendantes de ce vecteur aléatoire à partir desquelles la mesure de pertinence peut être estimée.

### 2.1. *Mesure de pertinence*

Nous définissons la pertinence d'un sous-ensemble non vide $\mathbb{X}$ de $\aleph$ par

$$\omega(\mathbb{X}) := I(\overrightarrow{\mathbb{X}}; \overrightarrow{\mathbb{Y}}) = \sum_{x,y} p_{(\overrightarrow{\mathbb{X}}, \overrightarrow{\mathbb{Y}})}(x,y) \log \frac{p_{(\overrightarrow{\mathbb{X}}, \overrightarrow{\mathbb{Y}})}(x,y)}{p_{\overrightarrow{\mathbb{X}}}(x) p_{\overrightarrow{\mathbb{Y}}}(y)},$$

où $p_{(\overrightarrow{\mathbb{X}}, \overrightarrow{\mathbb{Y}})}$, $p_{\overrightarrow{\mathbb{X}}}$ et $p_{\overrightarrow{\mathbb{Y}}}$ désignent respectivement la distribution jointe et les distributions marginales des vecteurs aléatoires $\overrightarrow{\mathbb{X}}$ et $\overrightarrow{\mathbb{Y}}$ et où $I(\overrightarrow{\mathbb{X}}; \overrightarrow{\mathbb{Y}})$ est l'*information mutuelle* entre $\overrightarrow{\mathbb{X}}$ et $\overrightarrow{\mathbb{Y}}$ [JOE 89, COV 91]. Il peut être vérifié que cette mesure de pertinence est monotone par rapport à l'inclusion, ce qui n'est pas sans poser des problèmes pratiques [KOJ 05]. La mesure $\omega$ est classiquement estimée par maximum de vraisemblance (proportions).

### 2.2. *Approximations $k$-additive de la mesure de pertinence*

En utilisant des notions de combinatoire telles que la *transformée de Möbius* (voir e.g. [GRA 97]), il a été montré [KOJ 05] que la pertinence d'un sous-ensemble $\mathbb{X} := \{X_{i_1}, \ldots, X_{i_r}\}$ de $\aleph$ peut être réécrite comme

$$\omega(\mathbb{X}) = \sum_{X_j \in \mathbb{X}} I(X_j; \overrightarrow{\mathbb{Y}}) - \sum_{\{X_j, X_k\} \subseteq \mathbb{X}} I(X_j; X_k; \overrightarrow{\mathbb{Y}}) + \sum_{\{X_j, X_k, X_l\} \subseteq \mathbb{X}} I(X_j; X_k; X_l; \overrightarrow{\mathbb{Y}}) - \cdots + (-1)^{r+1} I(X_{i_1}; \ldots; X_{i_r}; \overrightarrow{\mathbb{Y}}), \quad (1)$$

où $I(X_{i_1}; \ldots; X_{i_r})$ désigne l'information mutuelle entre $r > 2$ vecteurs aléatoires [WIE 96]. Cette généralisation de l'information mutuelle classique peut être interprétée comme une mesure de l'*interaction simultanée* entre $r > 2$ vecteurs aléatoires. Si elle est nulle, les $r$ vecteurs aléatoires n'interagissent pas simultanément. Il est important de noter que l'information mutuelle entre plus de deux vecteurs aléatoires n'est pas nécessairement positive [COV 91].

La pertinence de $\mathbb{X}$ est ainsi calculée d'abord en sommant les pertinences des singletons contenus dans $\mathbb{X}$, puis en soustrayant les informations mutuelles entre paires de variables de $\mathbb{X}$ et $\overrightarrow{\mathbb{Y}}$, ensuite en ajoutant les informations mutuelles entre variables des sous-ensembles de 3 éléments de $\mathbb{X}$ et $\overrightarrow{\mathbb{Y}}$, etc. Les informations mutuelles qui sont rajoutées ou enlevées peuvent être vues comme des *termes correcteurs* ou des termes d'*ordre supérieur* et s'apparentent aux termes d'interaction utilisés dans le contexte de l'*analyse de la variance* ou des *modèles log-linéaires* [AGR 02, Chap. 8].

Afin d'obtenir une approximation de l'information mutuelle moins coûteuse en terme de temps de calcul, il a été proposé [KOJ 05] de procéder à une *troncature $k$-additive* de $\omega$ pour un $k \in \{1, \ldots, m\}$, c'est-à-dire de négliger

les *termes correcteurs* d'ordre supérieur à $k$ dans l'Eq. (1). Prendre la troncature 1-additive de $\omega$ est équivalent à considérer que la pertinence d'un sous-ensemble est égale à la somme des pertinences des singletons qu'il contient, i.e, que $\omega$ est additive. Bien que cette hypothèse soit à la base de la plupart des approches filtres (voir e.g.[HUT 05, §8]), dans la plupart des situations réelles, une telle simplification est trop extrême car, généralement, l'ensemble des variables potentiellement discriminantes contient des variables redondantes.

La troncature 2-additive, notée $\omega^{(2)}$, apparaît plus appropriée car elle prend partiellement en compte les interactions entre variables potentiellement discriminantes sans être trop complexe en terme de nombre de coefficients. En effet, $\omega^{(2)}$ est complètement définie à partir de ses valeurs sur les singletons et les paires de variables potentiellement discriminantes, i.e, pour tout $\mathbb{X} \subseteq \aleph$ non vide, il peut être montré que [GRA 97, KOJ 05] que

$$\omega^{(2)}(\mathbb{X}) = \sum_{\{X_i, X_j\} \subseteq \mathbb{X}} \omega(\{X_i, X_j\}) - (|\mathbb{X}| - 2) \sum_{X_i \in \mathbb{X}} \omega(\{X_i\}). \quad (2)$$

Utiliser $\omega^{(2)}$ est très avantageux du point de vue du temps de calcul : une fois les pertinences des singletons et des paires de $\aleph$ estimées, la pertinence approchée de n'importe quel sous-ensemble de $\aleph$ peut être immédiatement calculée à l'aide de l'équation précédente. Du point de vue de la qualité de l'approximation, nous pouvons voir, en considérant l'Eq. (1), que plus la dépendance entre variables de $\aleph$ est faible, meilleure sera l'approximation de $\omega$ par sa troncature 2-additive. Pour plus de détails, voir [KOJ 05].

## 3. Classification hiérarchique ascendante de variables pour identifier la *structure* de $\aleph$

Le deuxième élément fondamental d'une procédure de sélection de variables est un algorithme de recherche. Afin d'éviter d'avoir à parcourir la totalité des sous-ensembles non vides de $\aleph$ ou d'avoir à recourir à des heuristiques souvent trop sous-optimales du type *sélection pas à pas*, nous proposons d'effectuer une classification ascendante hiérarchique de $\aleph$ afin d'en identifier la *structure*.

### 3.1. *Classification ascendante hiérarchique de variables fondée sur l'information mutuelle*

Un algorithme de classification ascendante hiérarchique est classiquement défini par deux éléments : une mesure de similarité (ou dissimilarité) et un critère d'agrégation entre classes. Les partitions compatibles avec la hiérarchie de classes obtenue sont généralement évaluées (en vue par exemple du choix d'une partition) en fonction de leur *homogénéité* et de leur *séparation*. Une première façon simple de mesurer l'homogénéité et la séparation d'une partition consiste à calculer le *diamètre* moyen et l'*écart* moyen respectivement de ses classes (voir e.g. [HAN 97]).

En tant que mesure de similarité, nous avons opté encore une fois pour l'information mutuelle, cette fois-ci normalisée :

$$I^*(X;Y) := \frac{I(X;Y)}{\min[H(p_X, H(p_Y)]}.$$

où $H$ désigne l'entropie de Shannon [SHA 48]. La quantité $I^*(X;Y)$ est clairement comprise entre 0 et 1. De plus, $I^*(X;Y) = 1$ si et seulement si $X$ et $Y$ sont fonctionnellement dépendantes [JOE 89, Th. 2.3].

### 3.2. *Algorithme de recherche*

Idéalement, l'objectif serait de retenir, parmi les partitions les plus homogènes compatibles avec la hiérarchie obtenue, la moins fine. D'un point de vue pratique, il faut tempérer l'objectif précédent en trouvant un compromis entre une forte homogénéité et un faible nombre de classes. Nous nous contentons ici dans un premier temps d'identifier un "coude" sur le graphique donnant le diamètre moyen des partitions compatibles en fonction de leur taille [HAR 96]. L'heuristique que nous proposons alors est de n'estimer que la pertinence des sous-ensembles composés d'au plus une variable de chaque classe, les variables d'une même classe pouvant être considérées comme "suffisamment dépendantes", ce qui nous conduit également à choisir le lien moyen comme critère d'agrégation. L'homogénéité des classes de la partition retenue joue donc un rôle important. La pertinence des sous-ensembles étant mesurée par le biais de la troncature 2-additive de l'information mutuelle (voir Eq. (2))

pénalisant les sous-ensembles contenant des variables liées, un certain degré de dépendance inter-classes est envisageable en pratique.

Une fois une partition compatible sélectionnée, il est demandé à l'utilisateur de donner le nombre maximum $p$ de variables discriminantes à retenir. L'algorithme (implémenté sur la plateforme R [R D 05]) est donné ci-après :

| | |
|---|---|
| *Entrées :* | - $\mathcal{P} = \{C_1, C_2, ..., C_k\}$, une partition de $\aleph$ en $k$ classes,<br>- $p$, le cardinal maximum des sous-ensembles, fixée par l'utilisateur. |
| *Sortie :* | Les "meilleurs" sous-ensembles de cardinal 1 à $\min(p, k)$ compatibles avec $\mathcal{P}$. |
| *Algorithme :* | **Pour** $i = 1$ **à** $\min(p, k)$<br>**Pour tous** les sous-ensembles de taille $i$ compatibles avec $\mathcal{P}$<br>Estimer la pertinence de cet ensemble<br>**Fin pour**<br>Afficher le meilleur sous-ensemble de taille $i$<br>**Fin pour** |

## 4. Un exemple très simple

Afin d'illustrer l'approche proposée, nous avons généré un problème très simple de faible taille (dont nous connaissons ainsi la *structure*). Nous considérons un ensemble de 22 variables discrètes potentiellement discriminantes. Les variables $X_1, \ldots, X_5$ et $X_{21}, X_{22}$, mutuellement indépendantes, à valeurs dans $\{1, 2, 3, 4\}$, sont distribuées selon une loi uniforme. Les variables $X_6, \ldots, X_{10}$ sont définies par $X_i := 4 - X_{i-5}$, les variables $X_{11}, \ldots, X_{15}$ par $X_i := X_{i-10}^2$. Enfin, les variables $X_{16}, \ldots, X_{20}$ sont définies par $X_i := \min(X_1, X_2)$. La variable aléatoire $Y$ à expliquer est définie par $Y := \max(X1, X2, X3) + \min(X_4, X_5)$. Nous avons ensuite généré $n = 800$ réalisations du vecteur aléatoire $(X_1, \ldots, X_{22}, Y)$. La partition retenue est la partition en 8 classes :

$$\{\{X_1, X_6, X_{11}\}, \{X_2, X_7, X_{12}\}, \{X_3, X_8, X_{13}\}, \{X_4, X_9, X_{14}\}, \{X_5, X_{10}, X_{15}\}, \{X_{16}, \ldots, X_{20}\}, \{X_{21}\}, \{X_{22}\}\}.$$

Les sous-ensembles de 1,2,...,5 variables renvoyés sont respectivement $\{X_5\}$, $\{X_4, X_5\}$, $\{X_3, X_4, X_5\}$, $\{X_1, X_3, X_4, X_5\}$ et $\{X_1, X_2, X_3, X_4, X_5\}$.

## 5. Bibliographie


[AGR 02] AGRESTI A., *Categorical Data Analysis*, Wiley, 2002, Second edition.

[COV 91] COVER T., THOMAS J., *Elements of Information Theory*, John Wiley and Sons, 1991.

[GRA 97] GRABISCH M., $k$-order additive discrete fuzzy measures and their representation, *Fuzzy Sets and Systems*, vol. 92, n$^o$ 2, 1997, p. 167–189.

[HAN 97] HANSEN P., JAUMARD B., Cluster analysis and mathematical programming, *Mathematical programming*, vol. 79, 1997, p. 191–215.

[HAR 96] HARDY A., On the number of clusters, *Computational Statistics and Data Analysis*, vol. 23, 1996, p. 83–96.

[HUT 05] HUTTER M., ZAFFALON M., Distribution of mutual information from complete and incomplete data, *Computational Statistics and Data Analysis*, vol. 48, 2005, p. 633–657.

[JOE 89] JOE H., Relative entropy measures of multivariate dependence, *J. Am. Statist. Assoc.*, vol. 84, 1989, p. 157–164.

[KOJ 05] KOJADINOVIC I., Relevance measures for subset variable selection in regression problems based on k-additive mutual information, *Computational Statistics and Data Analysis*, vol. 49, n$^o$ 4, 2005, p. 1205–1227.

[LIU 98] LIU H., MOTODA H., *Feature selection for knowledge discovery and data mining*, Kluwer Academic Publishers, 1998.

[R D 05] R DEVELOPMENT CORE TEAM, R : A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, 2005, ISBN 3-900051-00-3.

[SHA 48] SHANNON C. E., A mathematical theory of communication, *Bell Systems Technical Journal*, vol. 27, 1948, p. 379–623.

[WIE 96] WIENHOLT W., SENDHOFF B., How to determine the redundancy of noisy chaotic time series, *International Journal of Bifurcation and Chaos*, vol. 6, n$^o$ 1, 1996, p. 101–117.

# Algorithme de construction de la hiérarchie faible associée à une mesure de dissimilarité

**Jean Diatta**

*IREMIA, Université de la Réunion*
*15 avenue René Cassin - BP 7151*
*97715 Saint-Denis messag cedex 9, France*
*Jean.Diatta@univ-reunion.fr*

RÉSUMÉ. *Nous proposons un algorithme de construction de la hiérarchie faible associée à une mesure de dissimilarité.*

MOTS-CLÉS : *Algorithme, Classe faible, Dissimilarité, Hiérarchie faible.*

## 1. Introduction

Les hiérarchies faibles sont des structures de classification qui généralisent les hiérarchies [BAN 89]. Elles permettent l'empiétement des classes et peuvent être entièrement construites en temps polynomial à partir de mesures de dissimilarité. En effet, chaque hiérarchie faible est la hiérarchie faible associée à une mesure de dissimilarité donnée [DIA 94].

Dans cette note, nous proposons un algorithme de construction de la hiérarchie faible associée à une mesure de dissimilarité. Cet algorithme est fondé sur le fait que chacune des classes composant cette hiérarchie faible est engendrée par une paire d'entités et satisfait une certaine propriété d'inclusion. Cette propriété d'inclusion permet de construire certaines de ces classes en n'effectuant qu'un nombre limité de tests. Elle permet également de ne construire qu'une seule fois les classes engendrées par plusieurs paires d'entités.

## 2. Hiérarchies faibles

Les hiérarchies faibles ont été introduites à la fin des années 80 en affaiblissant la condition caractéristique des hiérarchies dites fortes [BAN 89]. Soit $E$ un ensemble fini non vide. Une *hiérarchie faible* sur $E$ est une collection $\mathcal{W}$ de sous-ensembles de $E$, satisfaisant la propriété :

> l'intersection de trois membres $X, Y, Z$ de $\mathcal{W}$ est toujours l'intersection de deux membres parmi ces trois, i.e., $X \cap Y \cap Z \in \{X \cap Y, X \cap Z, Y \cap Z\}$.

EXEMPLE 1 *Soit $E = \{a, b, c, d, e\}$. Alors $\mathcal{W} = \{a, \{a, b\}, \{a, c\}, \{a, d\}, \{a, e\}, E\}$ est une hiérarchie faible sur $E$.*

### 2.1. *Hiérarchies faibles et mesures de dissimilarité*

Une mesure de dissimilarité sur $E$ est une fonction $d : E \times E \to \mathbb{R}$ vérifiant $d(x, x) = 0$, $d(x, y) \geq 0$ et $d(x, y) = d(y, x)$. Les mesures de dissimilarité jouent un rôle important en classification où elles sont souvent

utilisées pour construire des classes ayant un degré de dissimilarité intra classe faible et/ou un degré de dissimilarité inter classes élevé. Les classes faibles introduites dans [BAN 89] dans le cadre des similarités sont parmi ces classes.

Un sous-ensemble $X$ de $E$ est une *classe faible* associée à une mesure de dissimilarité $d$ (ou *classe $d$-faible*), si son *indice d'isolation $d$-faible*

$$\mathbf{i}_d^w(X) := \min_{\substack{x,y\in X \\ z\notin X}} \{\max\{d(x,z), d(y,z)\} - d(x,y)\}$$

est strictement positif. En d'autres termes, pour tous $x, y$ dans la classe et tout $z$ extérieur à la classe, au moins l'une des dissimilarités $d(x,z)$ et $d(y,z)$ est strictement supérieure à la dissimilarité $d(x,y)$.

La proposition 1 ci-dessous montre que les classes faibles associées à une mesure de dissimilarité forment une hiérarchie faible [BAN 89].

PROPOSITION 1 *Soit $d$ une mesure de dissimilarité sur $E$. Alors les classes faibles associées à $d$ forment une hiérarchie faible appelée la* hiérarchie faible associée à $d$.

### 2.2. *Hiérarchies faibles et* 2*-boules*

Dans le paragraphe 2.1, nous avons vu que les hiérarchies faibles sont liées aux dissimilarités via les classes faibles. Il s'avèrent que ces classes faibles sont des intersections spéciales de boules, appelées 2-boules.

Soit $d$ une mesure de dissimilarité sur $E$. Soit $x, y$ deux éléments (non nécessairement distincts) de $E$ et soit $r$ un nombre réél positif ou nul. La $d$-boule (ou simplement boule) de centre $x$ et de rayon $r$ est l'ensemble $B^d(x,r)$ des éléments de $E$ dont le degré de dissimilarité d'avec $x$ est au plus $r$, *i.e.*, formellement, $B^d(x,r) = \{z \in E : d(x,z) \leq r\}$ ; la $(d,2)$-boule (ou simplement 2-boule) engendrée par $x, y$ est l'ensemble $B_{xy}^d$ défini par $B_{xy}^d = B(x, d(x,y)) \cap B(y, d(x,y))$.

La proposition ci-dessous caractérise une classe faible comme étant un sous-ensemble contenant toutes les 2-boules engendrées par ses paires d'éléments (non nécessairement distincts) [DIA 94].

PROPOSITION 2 *Soit $d$ une mesure de dissimilarité sur $E$. Un sous-ensemble $X$ de $E$ est une classe $d$-faible si et seulement s'il satisfait la propriété d'inclusion : $\forall x, y \in X : B_{xy}^d \subseteq X$.*

De la proposition 2 peut être aisément déduit que toute classe faible est une 2-boule [DIA 94].

PROPOSITION 3 *Soit $d$ une mesure de dissimilarité sur $E$. Si un sous-ensemble $X$ de $E$ est une classe $d$-faible, alors $X = B_{xy}^d$, où $x, y$ sont tels que $d(x,y) = \max_{u,v\in X} d(u,v)$.*

Une 2-boule $B_{xy}^d$ sera dite *faiblement isolées* si elle est une $d$-classe faible. Il résulte des propositions 1 and 3 que la hiérarchie faible associée à une mesure de dissimilarité $d$ est l'ensemble des 2-boules faiblement isolées de $d$, augmenté de l'ensemble vide. Cela nous fournit un moyen pour construire la hiérarchie faible associée à une mesure de dissimilarité. Par ailleurs, la propriété directe suivante sera utile pour la construction des 2-boules faiblement isolées [DIA 98].

PROPOSITION 4 *Soit $d$ une mesure de dissimilarité sur $E$. Soit $B_{xy}^d$ une 2-boule faiblement isolée contenant $u, v$. Si $d(u,v) \geq d(x,y)$ et $B_{uv}^d$ est faiblement isolée, alors $B_{uv}^d = B_{xy}^d$.*

## 3. Algorithme de construction de la hiérarchie faible associée à une mesure de dissimilarité

### 3.1. *Construction globale des* 2*-boules faiblement isolées*

Etant donnée une mesure de dissimilarité $d$ sur $E$, l'algorithme 1 ci-dessous construit une 2-boule $B_{ij}^d$ lorsque celle-ci est faiblement isolée, en présentant des entités successivement choisies dans un ensemble $X$. La 2-boule $B_{ij}^d$ est initialisée à son générateur (ligne 1), et son complémentaire $\overline{B}_{ij}^d$ est initialisé à l'ensemble vide (ligne 2). Alors des entités $k$ sont choisies dans $X$ et testées pour savoir si elles appartiennent à $B_{ij}^d$ ou à $\overline{B}_{ij}^d$ (ligne 4). Après affectation de $k$ soit à $B_{ij}^d$ ou à $\overline{B}_{ij}^d$ (ligne 5 ou 12), on teste si l'ensemble courant $B_{ij}^d$ est faiblement isolé relativement à l'ensemble courant $\overline{B}_{ij}^d$ ; si ce test est négatif, la construction de $B_{ij}^d$ s'arrête (lignes 6-10 ou 13-17). Si $B_{ij}^d \cup \overline{B}_{ij}^d = X$, alors $B_{ij}^d$ est faiblement isolée.

ALGORITHME 1 (GWI2B($i, j, X$))
**Input :** Un sous-ensemble fini non vide $X$ d'entités et une paire d'entités $\{i, j\}$.
**Output :** La 2-boule $B_{ij}^d$ lorsqu'elle est faiblement isolée.

1: Set $B_{ij}^d = \{i, j\}$
2: Set $\overline{B}_{ij}^d = \varnothing$
3: **for** $k \in X$ **do**
4: **if** $\max\{d(i,k), d(j,k)\} \leq d(i,j)$ **then**
5: $B_{ij}^d \leftarrow k$
6: **for** $u \in B_{ij}^d$ and $v \in \overline{B}_{ij}^d$ **do**
7: **if** $\max\{d(k,v), d(u,v)\} \leq d(k,u)$ **then**
8: mark $B_{ij}^d$ as already considered and go to STOP
9: **end if**
10: **end for**
11: **else**
12: $\overline{B}_{ij}^d \leftarrow k$
13: **for** $u, v \in B_{ij}^d$ **do**
14: **if** $\max\{d(u,k), d(v,k)\} \leq d(u,v)$ **then**
15: mark $B_{ij}^d$ as already considered and go to STOP
16: **end if**
17: **end for**
18: **end if**
19: **end for**
20: mark $B_{ij}^d$ as weakly isolated
21: mark $B_{ij}^d$ as already considered
22: STOP

### 3.2. *Construction locale de* 2*-boules faiblement isolées*

Dans l'algorithme 2 ci-dessous, la 2-boule $B_{ij}^d$ est supposée faiblement isolée. Pour $u, v \in B_{ij}^d$, soit (1) $d(i,j) \leq d(u,v)$, soit (2) $d(i,j) > d(u,v)$. Dans le cas (1), si $B_{uv}^d$ est faiblement isolée, alors, par la proposition 4, on aura $B_{uv}^d = B_{ij}^d$ et il ne sera donc pas nécessaire de garder $B_{uv}^d$ puisque $B_{ij}^d$ est supposée être déjà considérée (lignes 3-4). Dans le cas (2), l'algorithme 1 y est utilisé pour construire $B_{uv}^d$ (si faiblement isolée), en présentant des entités choisies successivement dans $B_{ij}^d$ (ligne 6). En effet, comme $B_{ij}^d$ est supposée être faiblement isolée, $B_{uv}^d \subseteq B_{ij}^d$ d'après la proposition 2. D'où les entités extérieures à $B_{ij}^d$, donc extérieures à $B_{uv}^d$, ne peuvent empêcher $B_{uv}^d$ d'être faiblement isolée. Si $B_{uv}^d$ est faiblement isolée, alors l'algorithme 2 est récursivement appelé après insertion de la 2-boule $B_{uv}^d$ dans $\mathcal{F}$ (lignes 7-10).

ALGORITHME 2 (LWI2B($i, j, \mathcal{F}$))
**Input :** Une paire d'entités $\{i, j\}$ et un ensemble (potentiellement vide) $\mathcal{F}$ de 2-boules.
**Output :** Un ensemble (potentiellement vide) $\mathcal{F}$ de 2-boules $B^d_{uv}$, où $u, v \in B^d_{ij}$ et $B^d_{uv}$ est faiblement isolée.

```
1: for u, v ∈ B^d_ij do
2:    if B^d_uv is not already considered then
3:       if d(i, j) ≤ d(u, v) then
4:          mark B^d_uv as already considered
5:       else
6:          GWI2B(u, v, B^d_ij)
7:          if B^d_uv is weakly isolated then
8:             F←B^d_uv
9:             LWI2B(u, v, F)
10:         end if
11:      end if
12:   end if
13: end for
```

### 3.3. *Construction de la hiérarchie faible associée à une mesure de dissimilarité*

L'algorithme 3 ci-dessous est l'algorithme principal de construction de la hiérarchie faible associée à $d$. La construction d'une 2-boule $B^d_{ij}$ dépend du choix de $i$ et $j$. Si $i$ et $j$ ne sont pas choisis dans une 2-boule $B^d_{uv}$ déjà marquée comme étant faiblement isolée, alors l'algorithme 1 est utilisé en présentant des entités choisies successivement dans $E$ (ligne 4). Chaque fois qu'une 2-boule faiblement isolée $B^d_{ij}$ est construite et $B^d_{ij}$ insérée dans $\mathcal{F}$, l'algorithme 2 est exécuté par rapport à cette 2-boule (lignes 5-8).

ALGORITHME 3 (WH($E, d$))
**Input :** Un ensemble fini non vide d'entités $E$ et une mesure de dissimilarité $d$ sur $E$.
**Output :** La hiérarchie faible $\mathcal{F}$ associée à $d$.

```
1: Set F := ∅
2: for i, j ∈ E do
3:    if B^d_ij is not already considered then
4:       GWI2B(i, j, E)
5:       if B^d_ij is weakly isolated then
6:          F←B^d_ij
7:          LWI2B(i, j, F)
8:       end if
9:    end if
10: end for
```

L'analyse détaillée de cet algorithme montre qu'il améliore significativement celui initialement proposé par [BAN 89], décrit ci-dessous. Une raison simple à cela est que dans l'algorithme de Bandelt et Dress, l'isolation faible de chaque classe est testée par rapport à chaque élément du complémantaire de cette classe dans $E$ tout enier. Or, dans l'algorithme que nous proposons, pour ceratines classes (au moins celles engendrées par un singleton), cette isolation faible n'est testée que par rapport aux éléments des complémentaires respectifs de ces classes dans des parties propres respectives de $E$ (voir algorithme 2). Par ailleurs, contrairement à l'algorithme de Badelt et Dress, celui que nous proposons ne construit jamais deux fois la même classe, ce qui permet de faire l'économie de tests (coûteux) d'égalité de classes. Voici une description de l'algorithme de Bandelt et Dress.

On suppose que $E = \{e_1, \ldots, e_n\}$. Alors on construit successivement la hiérarchie faible $\mathcal{H}_k^d$ associée à la restriction de $d$ à $\{e_1, \ldots, e_k\}$ ($k \leq n$). On initialise l'algorithme en posant $\mathcal{H}_0^d = \varnothing$. Si $\mathcal{H}_k^d$ est déterminée pour $k < n$, alors pour toute classe $C$ appartenant à $\mathcal{H}_k^d$, on vérifie si les deux propriétés suivantes sont satisfaites :

(a) $\max\{d(e_i, e_{k+1}), d(e_j, e_{k+1})\} > d(e_i, e_j)$ pour tous $i, j \leq k$ avec $e_i, e_j \in C$,

(b) $\max\{d(e_i, e_j), d(e_j, e_{k+1})\} > d(e_i, e_{k+1})$ pour tous $i, j \leq k$ avec $e_i \in C$ et $e_j \notin C$.

Ainsi, $\mathcal{H}_{k+1}^d$ contient $C$ si et seulement si (a) est vrai, et il contient $C \cup \{k+1\}$ si et seulement si (b) est vrai. Finalement, $\mathcal{H}_n^d$ est la hiérarchie faible associée à $d$.

L'interprétaion des classes faibles n'est pas toujours très facile, étant donné qu'elles sont construites à partir uniquement d'une mesure de dissimilarité. Toutefois, si l'on se place dans le cadre d'un contexte à descriptions fermées par borne inférieure, c'est-à-dire, un contexte dans lequel l'espace de description des entités est un inf-demitreillis, alors le descripteur des entités induit une correspondance de Galois entre l'ensemble des sous-ensembles d'entités et l'espace de description de ces entités [BIR 67]. On peut alors contruire ladite hiérarchie faible de Galois associée à une mesure dissimilarité [DIA 05]. Ainsi, les classes construites seront accompagnées de leurs descriptions respectives et seront donc plus facilement interprétables.

## 4. Bibliographie


[BAN 89] BANDELT H.-J., DRESS A. W. M., Weak hierarchies associated with similarity measures : an additive clustering technique, *Bull. Math. Biology*, vol. 51, 1989, p. 113–166.

[BIR 67] BIRKHOFF G., *Lattice theory*, 3rd edition, Coll. Publ., XXV, American Mathematical Society, Providence, RI, 1967.

[DIA 94] DIATTA J., FICHET B., From Apresjan hierarchies and Bandelt-Dress weak hierarchies to quasi-hierarchies, DIDAY E., LECHEVALIER Y., SCHADER M., BERTRAND P., BURTSCHY B., Eds., *New Approaches in Classification and Data Analysis*, Springer-Verlag, 1994, p. 111–118.

[DIA 98] DIATTA J., Approximating dissimilarities by quasi-ultrametrics, *Discrete Mathematics*, vol. 192, 1998, p. 81–86.

[DIA 05] DIATTA J., Galois Weak Hierarchies : Theoretical and Computational issues, MIRKIN B., MAGOULAS G., Eds., *5th United Kingdom Workshop on Computational Intelligence*, 2005, p. 1–8.

# Une commémoration positive de la valeur de la méthode des moindres quarrés

**Antoine de Falguerolles**

*Université de Toulouse III (Paul Sabatier)*
*Laboratoire de Statistique et Probabilités, 118, route de Narbonne,*
*31062 Toulouse Cedex 9*
*falguero@cict.fr*

RÉSUMÉ. *La méthode des moindres carrés et ses avatars classiques tels les moindres carrés généralisés, les moindres carrés pondérés itérés, et les moindres carrés (pondérés itérés) alternés ont colonisé avec un certain succès la statistique. L'impérialisme des moindres carrés est parfois dénoncé au nom d'arguments assez convaincants. Mais il nous semble cependant opportun de commémorer, deux cents ans après, le rôle novateur qu'a joué dans cette affaire Adrien Marie Legendre (1752-1833) avec ses publications de 1805 et 1806.*

MOTS-CLÉS : *Moindres carrés (Moindres quarrés), Moindres carrés généralisés, Analyse de covariance.*

## 1. Introduction

« La règle par laquelle on prend le milieu entre les résultats de différentes observations, n'est qu'une conséquence très-simple de notre méthode générale que nous appellerons *Méthode des moindres quarrés*. »

En 1805, Adrien Marie Legendre (1752-1833) publiait sous forme d'annexe à un livre une méthode permettant de trouver une solution moyenne d'un système incompatible d'équations linéaires (Legendre, 1805). Cette annexe était publiée à nouveau en 1806 et *circa* 1830 (Legendre, 1806, ca 1830). Ces deux dernières sont disponibles sur le réseau international (*internet*)[1]. C'est l'exemple d'application de la méthode tel qu'il figure dans les présentation de 1806 qui fait l'objet de cette publication. En effet, dans un langage contemporain, Legendre donne un intéressant exemple de régression linéaire généralisée (erreurs à modèle de type moyenne mobile d'ordre 1). La démarche, élégante sinon originale, n'est pas sans rappeler celle utilisée un siècle et demi après dans des situations où les erreurs suivent un modèle autorégressif d'ordre 1 (Durbin et Watson, 1950, 1951). De par sa simplicité et sa flexibilité, la méthode des moindres carrés était promise à un bel avenir. Mais à quel horizon ? Il est toujours difficile de mesurer le laps de temps mis pour qu'une méthode sorte réellement du cercle étroit de ses inventeurs. D'ailleurs, n'avait-elle pas déjà un passé puisque Carl Friedrich Gauss (1777-1855) publiait en 1809 qu'il en avait eu l'idée en 1795 ! Mais qu'en est-il du grand public ? Le temps d'une génération pourrait en être une estimation raisonnable. Un exemple publié en 1834 indique que l'ajustement d'une droite par la méthode des moindres carrés était parfaitement maîtrisé par un ingénieur du corps des ponts et chaussées, Georges Muntz (1807, après 1869). L'exemple dépasse encore une fois la simple application d'une méthode puisque Muntz s'attaque en fait à la résolution d'un problème d'analyse de la covariance en des temps où l'idée de considérer une variable indicatrice comme une variable statistique classique était encore inconnue (Muntz, 1834). Ce sont ces deux exemples qui vont être re-visités dans cet article commémoratif.

1. La publication de 1806 est disponible grâce à l'obligence du Service interétablissement de coopération documentaire de Toulouse (SICDT) et de l'Observatoire Midi-Pyrénées, la seconde grâce à celle de la bibliothèque numérique « gallica ». Elles sont aussi disponibles à l'adresse suivante : http://www.lsp.ups-tlse.fr/Fp/Falguerolles/FACSIMILE/index.html

## 2. Les équations à faire concourir dans les moindres carrés

L'exemple de Legendre concerne la mesure du méridien de Paris et, notamment, la mesure de l'aplatissement de la terre. Cet aspect de la modélisation n'est pas évoqué ici mais on pourra se reporter très utilement à l'article commémoratif de Georges Balmino (2005).

Les données observées sont constituées des latitudes (exprimées en degrés) de $5$ lieux d'observations (Dunkerque, le Panthéon à Paris, Evaux, Carcassonne et Montjouy) et des mesures (exprimées en modules de deux toises) des $4$ arcs compris entre ces lieux. L'histoire de ces données est magistralement contée dans les ouvrages de Denis Guedj (1986) et de Ken Alder (2002, traduction française en 2005).

Notant $i$ ($i = 0, \ldots, 4$) les « lieux de l'observation », $L_i$ leurs latitudes, et $S_i$ les mesures des $4$ arcs $i$ ($i = 1, \ldots, 4$) entre les lieux de latitudes $L_{i-1}$ et $L_i$, les deux modèles proposées par Legendre ont pour expression formelle :
1. Variable réponse : $S_i$ ; prédicteur linéaire de la forme : $\beta_1(L_i - L_{i-1}) + \beta_2 K \sin(L_i - L_{i-1}) \cos(L_i + L_{i-1})$ où $K$ est un coefficient connu, et où $\beta_1$ et $\beta_2$ désignent les coefficients génériques inconnus du modèle.
2. Variable réponse : $L_i - L_{i-1}$ ; prédicteur linéaire de la forme : $\frac{S_i}{K'} + \beta_1 \frac{S_i}{K'} + \beta_2 K'' \sin(L_i - L_{i-1}) \cos(L_i + L_{i-1})$ où $K'$ et $K''$ sont des coefficients connus, et où $\beta_1$ et $\beta_2$ désignent encore les coefficients génériques inconnus du modèle.

Legendre privilégie le second modèle. Son choix est motivé par de meilleures possibilités d'interprétation physique des coefficients, notamment l'estimation du coefficient $\alpha$ d'aplatissement de la terre[2]. En effet, il s'avère que, si l'aplatissement $\alpha$ est considéré comme inconnu, le premier modèle dépend en fait de trois coefficients inconnus $\beta_1$, $\beta_2$ et $\alpha$ liés par la relation $\beta_2 = \beta_1 \alpha$ ; le modèle est donc bilinéaire ou bi-additif en $\beta_1$ et $\alpha$. Il s'avère aussi que, dans le second modèle, les paramètres $\beta_1$ et $\beta_2$ sont indépendants (avec $\beta_2 = \alpha$). D'où son choix. Mais il est alors douteux, d'un point de vue de stricte orthodoxie statistique, de retrouver dans le prédicteur linéaire une fonction de la variable réponse.

En tout état de cause, Legendre obtient ce que nous appelons de nos jours un modèle de régression linéaire. Rappelons-en les ingrédients en des termes contemporains : un vecteur réponse, $\underline{y}$, et un modèle linéaire pour son prédicteur, $\underline{\mu} = \mathbf{X}\underline{\beta}$. La méthode des moindres carrés consiste alors à estimer les coefficients $\underline{\beta}$ du prédicteur en minimisant $|| \underline{y} - \underline{\mu}(\underline{\beta}) ||^2$ et le génie de Legendre est d'avoir montré que le problème se ramenait à celui de la recherche de la solution d'un système d'équations linéaires, $\mathbf{X}'\mathbf{X}\underline{\beta} = \mathbf{X}'\underline{y}$, admettant, sous des conditions assez évidentes, une solution unique [3]. On peut penser qu'un des intérêts de la méthode était qu'elle se ramenait techniquement au calcul des coefficients d'un système (de Cramer) de petite taille et à sa résolution par éliminations successives des variables.

Toutefois, au terme de son analyse, Legendre est fortement déçu par la valeur estimée de l'aplatissement de la terre qu'il obtient dans cet exemple. En effet, la valeur de l'aplatissement pouvait être approchée par d'autres moyens de mesures et était donc connue par ailleurs. Mais ces résultats décevants n'altèrent en rien la confiance que Legendre place dans la méthode des moindres carrés dont il loue la « simplicité » et la « fécondité ».

## 3. L'analyse numérique de Legendre

Les observations sont naturellement ordonnées le long du méridien de Paris et Legendre veut intégrer à son ajustement les effets d'une propagation possible des erreurs de mesure le long de ce méridien. De façon intuitive, l'ajustement ne doit pas minimiser la somme des carrés des erreurs ($u_i = y_i - \mu_i$) mais la somme des carrés

2. En fait Legendre définit l'aplatissement comme le rapport de la différence ($a - b$) des demi-axes de l'ellipse au demi-petit axe ($b$).
3. Précisons que ces équations ne s'appelaient pas les équations normales, que l'on ne parlait pas encore de système de Cramer, que le mot de régression ne sera introduit que plus tard et dans un tout autre contexte …

d'erreurs latentes ($e_i$) gouvernant les premières. C'est cette démarche qui est examinée ci-après en la replongeant dans un cadre contemporain et en lui conférant une certaine généralité.

Legendre propose donc un modèle linéaire plausible de dépendance entre les $u_i$ ($i = 1, \ldots, 4$) et les erreurs latentes $e_i$ ($i = 0, 1, \ldots, 4$) : $u_i = e_i - e_{i-1}$. Legendre utilise alors le fait que la somme des erreurs latentes ($\sum e_i$) doit être nulle pour construire un problème augmenté dont les $e_i$ sont les erreurs. Le procédé de décorrélation ne serait pas nouveau et Stigler (Stigler, 1986, p. 60) l'attribue à Pierre Simon Laplace (1749–1827). Mais l'argument invoqué par Legendre en l'absence notamment de terme constant (intercept) est-il recevable ?

Suivant la démarche présentée par Legendre mais en la généralisant quelque peu, supposons que le vecteur des erreurs ($\underline{u}$ de dimension $n$) soit une transformation linéaire d'un vecteur d'erreurs latentes ($\underline{e}$ de dimension $n+k$) : $\underline{u} = \mathbf{L}\underline{e}$, la matrice $\mathbf{L}$ de dimension $(n, n+k)$ étant supposée de rang $n$. Soit alors une matrice $\boldsymbol{\ell}$ de dimension $(k, n+k)$ et de rang $k$ telle que $\boldsymbol{\ell}\mathbf{L}' = \mathbf{0}$.

En posant $\boldsymbol{\ell}\underline{e} = \underline{0}$, l'argument de Legendre, et compte tenu des hypothèses, on a :

$$\begin{bmatrix} \underline{u} \\ \underline{0} \end{bmatrix} = \begin{bmatrix} \mathbf{L} \\ \boldsymbol{\ell} \end{bmatrix} \underline{e} \qquad \text{soit encore} \qquad \underline{e} = \begin{bmatrix} \mathbf{L} \\ \boldsymbol{\ell} \end{bmatrix}^{-1} \begin{bmatrix} \underline{u} \\ \underline{0} \end{bmatrix};$$

il est alors facile de vérifier que $||\underline{e}||^2 = \underline{e}'\underline{e} = \underline{u}(\mathbf{L}\mathbf{L}')^{-1}\underline{u} = ||\underline{u}'||^2_{(\mathbf{L}\mathbf{L}')^{-1}}$.

Legendre se ramène ainsi à la résolution d'un problème ordinaire de moindres carrés entre un vecteur réponse augmenté $\underline{y_1} = \begin{bmatrix} \mathbf{L} \\ \boldsymbol{\ell} \end{bmatrix}^{-1} \begin{bmatrix} \underline{y} \\ \underline{0} \end{bmatrix}$ et une matrice expérimentale augmentée $\mathbf{X_1} = \begin{bmatrix} \mathbf{L} \\ \boldsymbol{\ell} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X} \\ \mathbf{0} \end{bmatrix}$. Autrement dit, par un simple pré-traitement des données, lié à un choix heureux de modèle pour les erreurs latentes (la matrice $\mathbf{L}$) et un choix *ad hoc* de la matrice $\boldsymbol{\ell}$, Legendre résout un problème particulier de moindres carrés généralisés.

Les valeurs numériques des matrices implicitement utilisées par Legendre dans son application figurent ci-dessous :

$$\mathbf{L} = \begin{bmatrix} -1 & 1 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & -1 & 1 \end{bmatrix} \qquad \mathbf{L}\mathbf{L}' = \begin{bmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{bmatrix}$$

$$\begin{bmatrix} \mathbf{L} \\ \boldsymbol{\ell} \end{bmatrix} = \begin{bmatrix} -1 & 1 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & -1 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

$$\begin{bmatrix} \mathbf{L} \\ \boldsymbol{\ell} \end{bmatrix}^{-1} = \frac{1}{5} \begin{bmatrix} -4 & -3 & -2 & -1 & 1 \\ 1 & -3 & -2 & -1 & 1 \\ 1 & 2 & -2 & -1 & 1 \\ 1 & 2 & 3 & -1 & 1 \\ 1 & 2 & 3 & 4 & 1 \end{bmatrix}$$

De nos jours, la méthode utilisée par Legendre, augmentation et inversion, peut faire sourire car il existe des méthodes numériques d'emploi nettement plus simple pour effectuer des moindres carrés généralisés. Connaissant $\mathbf{L}\mathbf{L}'$, on peut procéder sans augmentation des données en effectuant des moindres carrés ordinaires avec des données transformées $\mathbf{H}\underline{y}$ et $\mathbf{H}\mathbf{X}$. Un vaste choix s'offre pour $\mathbf{H}$ : $\mathbf{H} = (\mathbf{L}\mathbf{L}')^{-\frac{1}{2}}$, l'inverse $\mathbf{C}^{-1}$ de la décomposition de Cholesky[4] $\mathbf{C}$ de $\mathbf{L}\mathbf{L}'$, la décomposition de Cholesky $\mathbf{D}$ de l'inverse $(\mathbf{L}\mathbf{L}')^{-1}$ de $\mathbf{L}\mathbf{L}'$ ...

4. André-Louis Cholesky (officier français, 1875-1918) imagina sa méthode à l'occasion de travaux de géodésie effectués au service géographique de l'État-major de l'armée.

Dans l'exemple de Legendre, tout se passe bien, trop bien même. La situation est-elle aussi simple pour d'autres situations de dépendance des erreurs ? Il est assez naturel de replacer l'approche de Legendre dans le cadre des modèles *ARMA(p,q)* introduits par Georges Box et Gwilym Jenkins (1970) mais en se limitant aux modèles les plus simples : le schéma de type moyenne mobile d'ordre 1 ($u_i = e_i + \theta e_{i-1}$, $|\theta| \leq 1$), autrement dit un modèle *ARMA(0,1) = MA(1)*, et le schéma de type autorégressif d'ordre 1 ($u_i = \theta u_{i-1} + e_i$, $|\theta| < 1$), autrement dit un modèle *ARMA(1,0) = AR(1)*. Pour une discussion de ces aspects on pourra lire l'article récemment soumis par Antoine de Falguerolles et Didier Pinchon. Mais on notera en passant que l'approche considérée par Legendre, augmentation et inversion, se généralise aussi sans difficulté à tous les cas *MA(1)*. Il suffit de choisir $\boldsymbol{\ell} = \big(-\theta, \theta^2, -\theta^3, \ldots\big)$. L'inversion de la matrice augmentée $\mathbf{L}_1(\theta)$, $\mathbf{L}_1(\theta)' = [\mathbf{L}(\theta)' \; \boldsymbol{\ell}'(\theta)]$, devient alors un peu plus technique.

## 4. Georges Muntz et le chemin de l'analyse de covariance

Presque trent ans plus tard, en 1834, Georges Muntz publie les résultats d'une étude faite sur l'évolution du prix du transport total du mètre cube de divers matériaux de construction (et notamment de calcaire concassé) en fonction de la distance à parcourir. Il pose pour la moyenne du prix demandé ($y$, la variable réponse) un modèle de prédicteur linéaire simple classique ($b + ax$). Il fait alors « concourir toutes les équations à la détermination de $a$ et $b$ : en les assujettissant à la condition que la somme de leurs quarrés soit un *minimum* ». Il rappelle alors les formules des équations normales (sans les désigner sous ce nom [5]) et leur solution.

De fait, les données sont plus complexes. Muntz dispose dans l'exemple du calcaire concassé de $15$ observations relatives à des transports sur des chemins normaux ($y_{1j}$) et de $3$ observations relatives à des transports sur des chemins en terre glaise ($y_{2j}$). Comment estimer une ordonnée commune (partie fixe du prix) et deux pentes distinctes (une par groupe) ? On peut penser que si Muntz avait su coder sa variable qualitative groupe il aurait su aussi, comme Legendre l'avait présenté en 1805, faire une régression linéaire multiple.

Toutefois, Muntz résout le problème en profitant du fort déséquilibre des deux groupes ($n_1 > n_2$). Sa procédure comporte deux étapes : estimation de $b$ (ordonnée commune aux deux groupes ) et de $a_1$ (pente spécifique d'un groupe) sur le groupe le plus nombreux (numéroté 1) par une régression linéaire simple classique ; estimation de $a_2$ sur le groupe le moins nombreux en moyennant les réponses transformées $\dfrac{y_{2j} - b}{x_{2j}}$.

Les deux étapes relèvent bien sûr de l'application des moindres carrés et le lecteur pourra vérifier que Muntz définit ainsi des estimateurs linéaires. Certes, ces estimateurs ne possèdent pas l'optimalité garantie par le théorème de Gauss-Markov dans une estimation globale de $b$, $a_1$ et $a_2$. Mais les estimations obtenues sont ici très proches des estimations optimales.

## 5. Conclusion

Les deux exemples ainsi re-visités relèvent de la statistique multidimensionnelle descriptive (ou encore empirique). En effet, ni Legendre ni Muntz n'introduisent de modèle probabiliste pour rendre compte des dissonances entre les relations observées. On notera que l'appendice de 1806 contient 22 occurrences du terme erreur mais en dehors de tout concept explicitement probabiliste. Pourtant, dans son mémoire publié vers 1830[6], Legendre reproduit presque mot à mot l'exposé de 1805 (et de 1806) mais introduit la méthode des moindres quarrés en rappelant que « M. le comte Laplace ayant trouvé par des considérations fondées sur le calcul des probabilités, que la méthode des moindres quarrés doit être employée préférablement à toute autre, pour trouver la valeur moyenne la plus exacte d'un ou de plusieurs éléments inconnus … ». Mais la manière dont les données sont traitées reste exemplaire pour des statisticiens. Un regret personnel en passant. Il est un peu surprenant de noter l'absence totale

5. Emmanuel Carvallo les désigne sous le nom d'« équations résultantes » dans son ouvrage de 1912.
6. Il s'agit d'un texte qui aurait été lu le 24 septembre 1811.

de graphiques statistiques dans ces publications. Et pourtant, Legendre est l'auteur d'un traité de géométrie de base souvent réédité et orné de fines illustrations !

Enfin, cette commémoration des moindres carrés ne saurait être complète sans que soit évoqué l'excellent ouvrage publié par Åke Björck il y a quelques années et consacré aux méthodes numériques pour les moindres carrés (1996). Le statisticien ignore souvent l'obscure algorithmique qui se met en œuvre à l'occasion de régressions notamment et les rend numériquement acceptables. Le livre de Björck permet d'en mesurer les enjeux. L'ouvrage était attendu et il avait fait lors de sa parution l'objet d'une publicité assez amusante qui souligne l'intérêt des commémorations. L'éditeur, la puissante *Society for Industrial and Applied Mathematics* pour ne pas la nommer, n'hésitait pas dans ce prospectus à rappeler "*1795 – Gauss discovers the method of least-squares . . .1995 – Björck writes a monograph that covers the full spectrum of relevant problems and methods in least squares.*" Si la seconde partie du message publicitaire n'est pas exagérée, force est de constater que la première reste un sujet de controverse. C'est Legendre qui a publié, le premier, un exposé précis des moindres carrés et déposé ainsi le nom de ce que Gauss persistait à appeler en 1809 *"meine Methode"*.

## 6. Bibliographie


Ken ALDER. *The measure of all things*. Little, Brown, Boston, 2002.

Georges BALMINO. Legendre et la mesure du méridien: 200 ans après. In *Guide des données astronomiques 2006 pour l'observation du ciel, Annuaire du Bureau des Longitudes*, pages 341–358, Paris, 2005. EDP Sciences.

Åke BJÖRK. *Numerical methods for least squares problems*. SIAM, Philadelphia, 1996.

Georges BOX and Gwilym JENKINS. *Time series analysis : Forecasting and control*. Holden-Day, San Francisco, 1970.

Emmanuel CARVALLO. *Le calcul des probabilités et ses applications*. Gauthier-Villars, Paris, 1912.

Antoine de FALGUEROLLES et Didier PINCHON. Une commémoration du bicentenaire de la publication en 1805 (et 1806) de la méthode des moindres quarrés par Adrien Marie Legendre. *soumis à publication*, 2006.

James DURBIN and Geoffrey S. WATSON. Testing for serial correlation in least squares regression. *Biometrika*, 37:409–428, 1950.

James DURBIN and Geoffrey S. WATSON. Testing for serial correlation in least squares regression. *Biometrika*, 38:159–178, 1951.

Denis GUEDJ. *La méridienne (1792-1799)*. Seghers, Paris, 1986.

Adrien Marie LEGENDRE. *Nouvelles méthodes pour la détermination des orbites des comètes avec un supplément contenant divers perfectionnemens de ces méthodes et leur application aux deux comètes de 1805*. Courcier, Paris, 1806.

Adrien Marie LEGENDRE. *Nouvelles méthodes pour la détermination des orbites des comètes*. Firmin Didot, Paris, ca 1805. Cité par Stigler.

Adrien Marie LEGENDRE. *Mémoire sur la méthode des moindres quarrés et sur l'attraction des ellipsoïdes homogènes*. s.n., s.l., ca 1830.

Georges MUNTZ. Note sur l'évaluation du prix du transport des matériaux de construction dans l'arrondissement du Nord. *Annales des Ponts et Chaussées*, 167:86–100, premier semestre 1834.

Stephen M. STIGLER. *The history of statistics, The measurement of uncertainty before 1900*. Harvard University Press, Cambridge, 1986.

# Sélection de modèle PLS par rééchantillonnage bootstrap

**A.FARAJ – H.NOCAIRI – M.CONSTANT**

Institut Français du Pétrole
1&4 Av. de Bois-Préau
92500 Rueil Malmaison
E-mails : abdelaziz.faraj@ifp.fr – hicham.nocairi@ifp.fr – michel.constant@ifp.fr

RÉSUMÉ. Le problème de sélection de modèle en régression PLS est primordial pour la modélisation de phénomènes physiques quand le nombre des variables est trop important. Les méthodes de sélection consistent à retenir, parmi les modèles ayant un bon pouvoir de prédiction, ceux qui font intervenir le minimum de variables explicatives. Celle que nous présentons est basée sur l'utilisation du bootstrap. Elle permet de calculer la distribution empirique des coefficients du modèle et de n'en conserver que les plus significatifs grâce à des tests statistiques. Elle permet de mesurer, par ailleurs, le pouvoir prédictif des modèles de régression construits. L'approche est illustrée sur un jeu de données.

MOTS-CLÉS. Régression PLS, bootstrap, validation croisée, sélection de variables, sélection de modèle.

## 1 Introduction

Dans l'industrie pétrolière, la plupart des processus se présentent sous la forme d'un système à entrées-sorties. Il est souvent nécessaire de faire recours à des modèles pour expliciter les relations pouvant exister entre les variables d'entrée et les réponses qui leur sont associées. La régression PLS apparaît, dans la plupart des cas, comme la méthode la plus appropriée pour construire ces modèles. Non seulement elle est bien adaptée quand les variables explicatives présentent des fortes colinéarités ou quand leur nombre dépasse celui des individus, elle est aussi une méthode factorielle qui a l'avantage d'apporter un point de vue exploratoire sur les données.

Mais souvent il devient nécessaire, dans un souci de diminution de coût d'expérimentation et pour éviter des modèles confus et/ou d'interprétation difficile, de réduire le nombre de variables explicatives. On fait alors appel à des méthodes de sélection de variables.

Nous présentons, dans ce papier, une méthode de sélection de variables en régression PLS basée sur le ré-échantillonnage par bootstrap. Elle est basée sur un processus itératif de sélection de variables ; version légèrement modifiée de la PLS-bootstrap proposée par Lazraq et al. [LAZ 03]. Des échantillons aléatoires sont tirés avec remise dans l'ensemble des points disponibles, et servent à la construction de plusieurs modèles. Les distributions empiriques des coefficients permettent via des tests statistiques de sélectionner les variables pertinentes. On calcule ensuite les prédictions de ces modèles sur les individus n'ayant pas été tirés (individus de test). Des indices statistiques (erreur quadratique de test, biais, variance, …) sont calculés à chaque étape de sélection afin d'évaluer les capacités prédictives des modèles.

## 2 Sélection de variables par algorithme PLS-bootstrap

Soit $\mathbf{X}=(\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^J)$ la matrice $N_xJ$ des J variables explicatives ($\mathbf{x}^j$ est le vecteur de dimension N représentant la $j^{ème}$ colonne de $\mathbf{X}$) et $\mathbf{Y}$ la matrice de la (ou des) variable(s) à expliquer (on supposera dans ce qui suit, sans perdre en généralité, qu'on a une seule variable à expliquer $\mathbf{y}$).

On notera $\mathbf{Z}=\{(\mathbf{x}_i,y_i),\ i=1,\ \ldots,\ N\}$ l'ensemble des individus où le vecteur $\mathbf{x}_i=(x_i^1,\ x_i^2,\ \ldots,\ x_i^J)^T$ représente la $i^{ème}$ ligne de $\mathbf{X}$ et le scalaire $y_i$ le $i^{ème}$ élément de $\mathbf{y}$.

La régression PLS consiste à calculer, par itérations successives, un modèle linéaire $\mathbf{y}=\mathbf{Xb}$, où les éléments du vecteur $\mathbf{b}=(b_1,\ b_2,\ \ldots,\ b_J)^T$ de dimension J sont les J coefficients du modèle (les variables $\mathbf{X}$ et $\mathbf{y}$ sont centrées-réduites). L'algorithme NIPALS est l'un des plus connus pour la mise en œuvre de la PLS [TEN 95].

Le bootstrap [EFR 93] est une technique de ré-échantillonnage basée sur des tirages aléatoires avec remise dans les données. Son but est de substituer à une distribution inconnue $\mathcal{F}$, dont sont issues les données, la distribution empirique $\mathcal{F}^*$ calculée à partir d'échantillons aléatoires. Ces échantillons aléatoires sont calculé de la manière suivante. Soit $z=\{z_1,\ z_2,\ \ldots,\ z_N\}$ l'échantillon de taille N, représentant les données dont on dispose, issu d'une population de distribution inconnue $\mathcal{F}$. A partir de cet échantillon, on construit un échantillon $z^*=\{z_1^*,\ z_2^*,\ \ldots,\ z_N^*\}$ de même taille N, qu'on appellera échantillon bootstrap, par N tirages aléatoires avec remise parmi les N observations de l'échantillon de départ. Dans l'échantillon bootstrap $z^*$, une observation $z_i$ de $z$ peut apparaître une ou plusieurs fois ou ne pas apparaître du tout. L'astérisque indique que $z^*$ n'est pas identique à $z$ mais en est une duplication aléatoire.

La PLS-bootstrap est une méthode de sélection de variables pour la régression PLS qui a été développée par Lazraq et al. [LAZ 03]. Nous en proposons la version suivante, pour un nombre L de bootsrap et un seuil $\alpha$ préalablement fixés par l'utilisateur :

**Étape 1 : Répéter pour $\ell$ =1, 2, …, L**

1.1 : Construire l'échantillon aléatoire $\mathbf{Z}^{*\ell}$ de taille N tiré avec remise dans $\mathbf{Z}$ : $\mathbf{Z}^{*\ell}=\{(\mathbf{x}_i,y_i),\ i\in C^{*\ell}\}$ est l'échantillon bootstrap $\ell$ qui servira en tant qu'ensemble d'apprentissage (i.e. pour la construction du modèle $\hat{\mathbf{y}}^{*\ell}$). $C^{*\ell}$ est l'ensemble des indices des individus ayant été tirés (certains peuvent être dupliqués plusieurs fois ; on a $\left|C^{*\ell}\right|=N$).

1.2 : Construire l'échantillon des individus non tirés : $\overline{\mathbf{Z}}^{*\ell}=\{(\mathbf{x}_i,y_i),\ i\in \overline{C}^{*\ell}\}$ désigné par le terme anglais *out of bag* (oob). $\overline{C}^{*\ell}$ est l'ensemble des indices des individus n'ayant pas été tirés dans l'échantillon bootstrap $\ell$. $\overline{\mathbf{Z}}^{*\ell}$ servira comme ensemble de test pour le modèle $\hat{\mathbf{y}}^{*\ell}$.

1.3 : Construire le modèle $\hat{\mathbf{y}}^{*\ell}=\mathbf{X}^{*\ell}\,\mathbf{b}^{*\ell}$ de régression PLS où $\mathbf{b}^{*\ell}=(b_1^{*\ell},\ b_2^{*\ell},\ \ldots,b_J^{*\ell})^T$ est le vecteur colonne des coefficients du modèle et $\mathbf{X}^{*\ell}$ la matrice des données d'apprentissage dont les lignes sont les individus $\mathbf{x}_i$ où $i\in C^{*\ell}$.

1.4 : Calculer les prédictions du modèle pour les individus i non tirés (échantillon de test) : $\hat{y}_i^{*\ell}=(\mathbf{b}^{*\ell})^T\mathbf{x}_i^{*\ell}$, $\mathbf{x}_i^{*\ell}=(x_i^{1*\ell},\ x_i^{2*\ell},\ \ldots,\ x_i^{J*\ell})^T$ où $i\in \overline{C}^{*\ell}$.

1.5 : Calculer l'erreur quadratique moyenne de test : $EQMT^{*\ell}$ (cf. **(1)**) à partir des individus de l'échantillon de test.

1.6 : Calculer le coefficient de validation $Q^{*\ell 2}$ (cf. **(2)**) à partir des individus de l'échantillon de test.

**Étape 2 : Pour i=1, …, N, calculer**

2.1 : Les ensembles $\Lambda^i=\{\ell,\ i\in C^{*\ell}\}$ des indices $\ell$ des échantillons bootstrap contenant i et $\Lambda^{-i}=\{\ell,\ i\in \overline{C}^{*\ell}\}$ des indices $\ell$ des échantillons bootstrap ne contenant pas i dont les cardinaux respectifs sont notés $\left|\Lambda^i\right|$ et $\left|\Lambda^{-i}\right|$.

2.3 : L'erreur de prédiction $e_{(-i)}^{*\ell}$ pour chaque bootstrap $\ell\in\Lambda^{-i}$

2.3 : La variance de prédiction $\sigma^{*}_{(-i)}{}^{2}$ à partir des $\left|\Lambda^{-i}\right|$ modèles booststrap.

2.4 : Le biais $\mathcal{B}^{*}_{(-i)}$ de prédiction à partir des $\left|\Lambda^{-i}\right|$ modèles booststrap.

Les notations "(-i)" des indices désignent que $\sigma^{*}_{(-i)}{}^{2}$, $\mathcal{B}^{*}_{(-i)}$ et $e^{*\ell}_{(-i)}$ sont calculés par des modèles que les individus i n'ont pas servi à construire.

**Étape 3 : Répéter pour j=1, 2, …, J**

3.1 : Calculer l'intervalle de confiance, au seuil α, $\mathbf{I}^{*}_{j}(\alpha)$ pour le coefficient $b_j$ de la variable $\mathbf{x}^{j}$ à partir de l'échantillon bootstrap $\{ b^{*\ell}_{j},\ \ell = 1, L \}$

3.2 : Éliminer les variables $\mathbf{x}^{j}$ pour lesquelles $0 \in \mathbf{I}^{*}_{j}(\alpha)$.

**Répéter les étapes 1 à 3 avec les variables $\mathbf{X}^{j}$ retenues, jusqu'à ce qu'aucune variable ne soit éliminée (convergence de l'algorithme).**

On retient le modèle associé au couple (L,α) qui réalise le maximum de la médiane de la distribution $Q^{*\ell 2}$ et le minimum de sa variance.

L'algorithme présenté ci-dessus converge au bout de quelques itérations (dépassant rarement 5). Or, selon la nature des donnée, les modèles sont susceptibles de dégradation au fur et à mesure des itérations. Il est alors nécessaire d'examiner individuellement chacune de ces itérations pour sélectionner celle qui donne le meilleur modèle ; ce qui est possible grâce aux indicateurs $EQMT^{*\ell}$, $Q^{*\ell 2}$, $e^{*\ell}_{(-i)}$, $\sigma^{*}_{(-i)}{}^{2}$ et $\mathcal{B}^{*}_{(-i)}$ calculées lors des étapes 1 et 2 de l'algorithme, définies par :

$$EQMT^{*\ell} = \frac{1}{\left|\overline{C}^{*\ell}\right|} \sum_{i \in \overline{C}^{*\ell}} (\hat{y}^{*\ell}_{(-i)} - y_i)^2 \qquad \textbf{(1)}$$

$$Q^{*\ell 2} = \mathrm{Cor}^2(\hat{\mathbf{y}}^{*\ell}, \mathbf{y}) \qquad \textbf{(2)}$$

avec $\hat{\mathbf{y}}^{*\ell} = (\hat{y}^{*\ell}_{(-i)})^{T}$ et $\mathbf{y} = (y_i)^{T}$ où $i \in \overline{C}^{*\ell}$.

$$e^{*\ell}_{(-i)} = \hat{y}^{*\ell}_{(-i)} - y_i \text{ pour } \ell \in \Lambda^{-i} \qquad \textbf{(3)}$$

$$\sigma^{*}_{(-i)}{}^{2} = \frac{1}{\left|\Lambda^{-i}\right|} \sum_{\ell \in \Lambda^{-i}} (\hat{y}^{*\ell}_{(-i)} - \overline{\hat{y}}^{*}_{(-i)})^2 \qquad \textbf{(4)}$$

$$\mathcal{B}^{*}_{(-i)} = \overline{\hat{y}}^{*}_{(-i)} - y_i \qquad \textbf{(5)}$$

$\overline{\hat{y}}^{*}_{(-i)} = \frac{1}{\left|\Lambda^{-i}\right|} \sum_{\ell \in \Lambda^{-i}} \hat{y}^{*\ell}_{(-i)}$ est la moyenne des prédictions des L modèles bootstrap au point i.

$EQMT^{*\ell}$, $Q^{*\ell 2}$, $e^{*\ell}_{(-i)}$, $\sigma^{*}_{(-i)}{}^{2}$ et $\mathcal{B}^{*}_{(-i)}$ ne servent pas dans le processus de sélection des variables. Leur intérêt est de rendre compte, *a posteriori*, de la qualité des modèles construits au fur et à mesure de l'algorithme. Ils permettent, de cette façon, de distinguer la (ou les) itération(s) correspondant aux meilleurs ensembles de variables sélectionnées (i.e. celles associées aux modèles dont les qualités de prédiction sont les meilleures). Ils renseignent, à terme, sur la nature (linéaire ou non linéaire) des relations qui existent entre les variables explicatives et la réponse.

## 3 Application

La méthode est appliquée à un jeu de données réalisé dans le cadre d'un projet de débitmétrie polyphasique. Il s'agit de modéliser les fractions volumiques d'eau, d'huile et de gaz dans le mélange pétrolier en sortie de puits. Un système de mesure basé sur la propagation d'ondes électromagnétiques

dans le fluide a été développé. Les paramètres calculés à partir du signal enregistré et les conditions expérimentales (salinité, pression et température du fluide) servent en tant que variables explicatives. Elles sont au nombre de 24 pour une campagne de 122 mesures expérimentales. La PLS-bootstrap a été utilisée pour diminuer le nombre des variables afin de construire le modèle. On s'intéresse ici à la modélisation de la variable wc (*water cut*), qui correspond au pourcentage d'eau dans le mélange. L'algorithme a convergé après 4 itérations au bout desquelles 11 variables ont été sélectionnées. Les distributions des EQMT et Q2 au cours des 4 itérations sont données à la figure 1 ci-dessous.

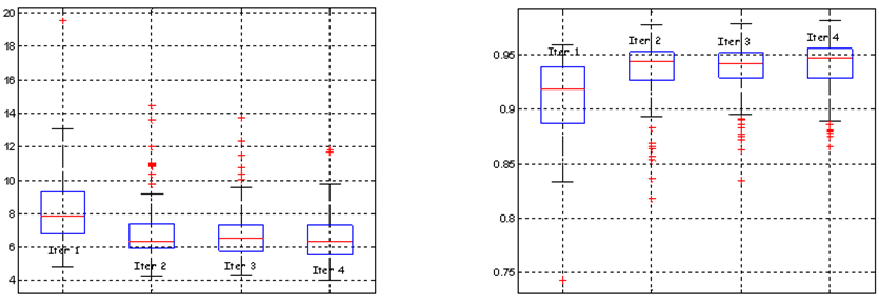



**Figure 1** : Distributions de l'erreur quadratique moyenne de test EQMT et du coefficient de validation Q2 pour les 4 itérations.

Les prédictions de la variable wc sont représentées à la figure 2 ci-dessous. On remarque qu'un certain nombre de points sont mal prédits. Il y a 2 raisons à cela : existence de mesures aberrantes (erreurs de saisie) et inversion des phases liquides dans le mélange (phénomène physique caractérisé par une réponse différente du signal selon qu'il y ait inclusion d'eau dans l'huile ou inclusion d'huile dans l'eau). La figure 3 permet d'identifier ces points par la projection, pour 50 bootstrap, des individus sur le plan factoriel défini par les 2 composantes PLS (t1,t2) pour les itérations 1 (à gauche du graphique) et 4 (à droite du graphique). Ce sont les points qui sortent du nuage.

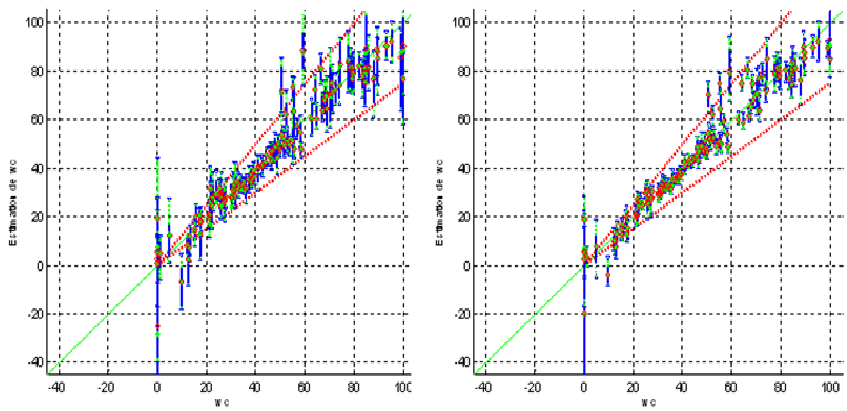


**Figure 2** : wc modélisé vs wc mesuré. Les trait pleins en bleu indiquent les intervalles de confiance de prédiction à 95 %

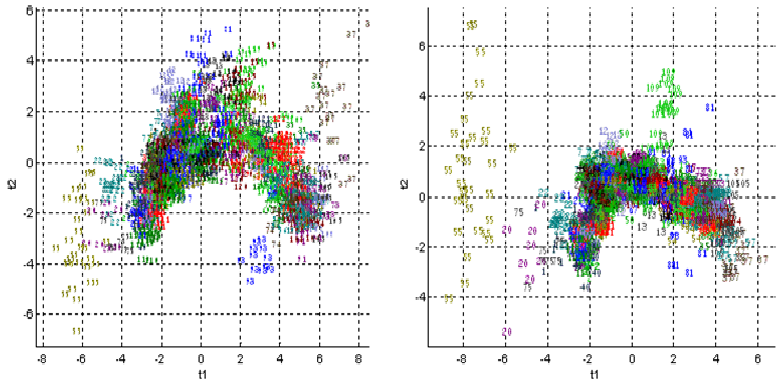



**Figure 3** : Projection des individus sur le plan (t1,t2) des composantes PLS pour les itérations 1 (graphique de gauche) et 2 (graphique de droite).

## 4 Bibliographie

[EFR 93], EFRON B., TIBSHIRANI R. *An introduction to the bootstrap*, Chapman and Hall, London 1993.

[FAR 04], FARAJ A., CONSTANT M. *Utilisation du bootstrap pour la sélection de variables et la typologie des individus en régression PLS*. 39èmes Journées de Statistique de la SFdS, Montpellier 2004, France.

[LAZ 03] LAZRAQ A., CLÉROUX R., GAUCHI J.-P. *Selecting both latent and explanatory variables in the PLS1 regression model* " Chemometrics and Intelligent Laboratory Systems, 2003, Vol **66**, 117-126.

[NOC 05] Noçairi H., Faraj A., 2005, Optimisation de la sélection des variables pertinentes pour modèle de régression PLS par bootstrap. Chimiométrie 30 nov – 1 déc. 2005, Lille

[TEN 95] TENENHAUS M., G GAUCH C. P. MÉNARDO C. Régression PLS et applications, *Revue Statistique Appliquée,* 1995, Vol **XLIII** (1), 7-63.

# Une base pour les règles d'association d'un contexte binaire valides au sens de la mesure de qualité $M_{GK}$

**Daniel R. Feno(*)(**), Jean Diatta(*), André Totohasina(**)**

*(*)Université de la Réunion*
*15 avenue René Cassin - BP 7151*
*97715 Saint-Denis messag cedex 9-France*
*fenodaniel2@yahoo.fr, jean.diatta@univ-reunion.fr*
*(**)Université d'Antsiranana- BP O*
*201 Antsiranana-Madagascar*
*totohasina@yahoo.fr*

RÉSUMÉ. *Ce papier concerne les règles d'association valides au sens de la mesure de qualité $M_{GK}$ qui est normalisée en ce sens que ses valeurs sont comprises entre $-1$ et $+1$ et reflètent les situations de référence telles que l'incompatibilité, la dépendance négative, l'indépendance, la dépendance positive, et l'implication logique entre la prémisse et le conséquent d'une règle. Par ailleurs, $\mathrm{M_{GK}}$ se trouve être la normalisée associée à la plupart des mesures de qualité proposées dans la littérature. Nous donnons une base pour l'ensemble des règles d'assoication $\mathrm{M_{GK}}$-valides, c'est-à-dire, une famille minimale de règles à partir de laquelle toutes les règles $\mathrm{M_{GK}}$-valides peuvent être dérivées par application d'axiomes d'inférence donnés. Cette base est en fait la réunion de quatre bases : la base des règles positives exactes, la base des règles négatives exactes, la base des règles positives approximatives et la base des règles négatives approximatives.*

MOTS-CLÉS : *Règle d'association, Mesure de qualité, Base*

## 1. Introduction

La fouille de données est un domaine de recherche actif dont l'importance n'a cessé de croître ces dernières années du fait de son rôle comme un outil approprié pour faire face à la croissance explosive de la taille de données stockées. Plusieurs techniques de fouille de données ont été proposées. La fouille des règles d'association figure parmi les plus populaires des méthodes de fouille des données. Les règles d'association sont utiles pour la découverte des relations au sein de très grandes bases de données. Plusieurs algorithmes de fouille de règles d'association, fondés sur les mesures *support* et *confiance*, ont été proposés dans la littérature : APRIORI [AGR 93], CLOSED [PAS 99], CLOSET [PEI 00]. Toutefois, l'ensemble des règles d'association valides au sens d'une mesure de qualité des règles comporte souvent un très grand nombre de règles dont plusieurs peuvent être rédondantes par rapport à des axiomes d'inférence donnés. Ainsi d'un point de vue informatif, il est intéressant de n'en générer qu'une base c'est-à-dire un ensemble minimal (au sens de l'ordre d'inclusion) à partir duquel, il peut être reconstitué par application de ces axiomes d'inférence.
Dans cette note, nous proposons une base pour les règles d'association valides au sens de la mesures $M_{GK}$ introduite dans [GUI 00] et dont les propriétés mathématiques ont été étudiées dans [TOT 05]. Cette base est la réunion de quatre bases : la base des règles positives exactes, la base des règles négatives exactes, la base de règles positives approximatives et la base des règles négatives approximatives. On notera que, selon [FEN 06], la mesure de qualité $\mathrm{M_{GK}}$ est normalisée en ce sens que ses valeurs sont comprises entre $-1$ et $+1$ et reflètent les situations de référence telles que l'incompatibilité, la dépendance négative, l'indépendance, la dépendance positive, et l'implication logique entre la prémisse et le conséquent d'une règle. De plus, ses propriétés permettent de considérer tout naturellement les règles d'association dites négatives, c'est-à-dire, dont la prémisse et/ou le conséquent est la

négation d'un motif. Par ailleurs, $\mathrm{M_{GK}}$ se trouve être la normalisée associée à la plupart des mesures de qualité proposées dans la littérature.

## 2. Règles d'association

Dans cet article, nous nous plaçons dans le cadre d'un contexte binaire $(\mathcal{E}, \mathcal{V})$ où $\mathcal{E}$ est un ensemble fini d'entités et $\mathcal{V}$ un ensemble fini de variables booléennes définies sur $\mathcal{E}$. Les sous ensembles de $\mathcal{V}$ seront appelés *motifs*.
Une *règle d'association* de $(\mathcal{E}, \mathcal{V})$ est un couple $(X, Y)$ de motifs, noté $X \rightarrow Y$, où $Y$ est non vide. Les motifs $X$ *et* $Y$ seront appelés respectivement la "*prémisse*" et la "*conclusion*" de la règle $X \rightarrow Y$.
Etant donnés deux motifs $X$ et $Y$ :

- $X'$ désignera l'ensemble des entités vérifiant le motif $X$, *i.e.*, $X' = \{e \in \mathcal{E} : \forall x \in X, x(e) = 1\}$.
- $\overline{X}$ désignera la négation de $X$, *i.e.*, $\overline{X}(e) = 1$ si et seulement s'il existe $x \in X$ tel que $x(e) = 0$ ($\overline{X}$)' est le complémentaire de $X'$).

La validité des règles d'association est évaluée par une ou plusieurs mesures de qualité pour ne déterminer que les règles d'association pertinentes au sens de ces mesures. Les plus connues de ces mesures de qualité sont sans doute le *support* et la *confiance* [AGR 93]. Pour un ensemble $A$, désignons par $|A|$ la cardinalité de $A$. Le support d'un motif $X$ est le nombre réel défini par $supp(X) = \frac{|X'|}{|\mathcal{E}|}$. Le support d'une règle $X \rightarrow Y$ noté $supp(X \rightarrow Y)$ est défini par : $supp(X \rightarrow Y) = supp(X \cup Y) = \frac{|(X \cup Y)'|}{|\mathcal{E}|}$. Il indique la proportion d'entités vérifiant à la fois la prémisse et la conclusion de la règle. La confiance d'une règle $X \rightarrow Y$ notée $conf(X \rightarrow Y)$ est définie par : $conf(X \rightarrow Y) = \frac{supp(X \rightarrow Y)}{supp(X)}$. Elle indique la proportion d'entités vérifiant la conclusion parmi ceux vérifiant la prémisse.
Dans ce papier, nous nous intéressons à la mesure de qualité $M_{GK}$ [GUI 00] définie par :

$$M_{GK}(X \rightarrow Y) = \begin{cases} \frac{P(Y'|X') - P(Y')}{1 - P((Y')} & \text{si } P(Y'|X') \geq P(Y') \text{ (dépendance positive)} \\ \frac{P(Y'|X') - P(Y')}{P(Y')} & si P(Y'|X') \leq P(Y') \text{ (dépendance négative),} \end{cases}$$

où $P$ est la probabilité uniforme discrète définie sur $(\mathcal{E}, \mathcal{P}(\mathcal{E}))$, *i.e.*, $\forall X' \subseteq \mathcal{E}, P(X') = \frac{card(X')}{card(\mathcal{E})}$. Soit, en fonction du concept de $confiance$ :

$$M_{GK}(X \rightarrow Y) = \begin{cases} \frac{conf(X \rightarrow Y) - supp(Y)}{1 - supp(Y)} & si\ X\ favorise\ Y,\ i.e.,\ si\ conf(X \rightarrow Y) \geq supp(Y) \\ \frac{conf(X \rightarrow Y) - supp(Y)}{supp(Y)} & si\ X \text{ défavorise } Y,\ i.e.,\ si\ conf(X \rightarrow Y) < supp(Y). \end{cases}$$

Notre intérêt pour la mesure $M_{GK}$ est motivé par le fait que celle-ci est, entre autres, la normalisée de la plupart des mesures de qualité proposées (dont $M_{GK}$ lui-même) dans la littérature [FEN 06] et qu'elle intègre un lien implicatif. De plus, d'un côté plus la valeur prise par $M_{GK}$ est proche de 1, plus la règle correspondante approche l'implication logique ; d'autre part, plus $M_{GK}$ s'approche de $-1$, plus les deux motifs prémisse et Conséquent de la règle ainsi évaluée tendent à être incompatibles, et alors il convient d'explorer les règles dites *négatives* associées, *i.e.* les règles du type $(X \rightarrow \overline{Y})$, ou $(\overline{X} \rightarrow Y)$. Les règles non négatives sont dites *positives*.

Soit $\alpha \in \mathbb{R}$. Une règle d'association positive $X \rightarrow Y$ (resp. négative $X \rightarrow \overline{Y}$) sera dite $\alpha$-*valide* au sens d'une mesure de qualité $\mu$ si $\mu(X \rightarrow Y) \geq \alpha$ (resp. $\mu(X \rightarrow \overline{Y}) \geq \alpha$).

## 3. Base des règles d'association valides au sens de la mesure de qualité $M_{GK}$

L'ensemble des règles d'association valides au sens d'une mesure de qualité des règles comporte souvent un très grand nombre de règles dont plusieurs peuvent être rédondantes par rapport à des axiomes d'inférence donnés. Ainsi d'un point de vue informatif, il est intéressant de n'en générer qu'une base c'est-à-dire un ensemble minimal (au sens de l'ordre d'inclusion) à partir duquel, il peut être reconstitué par application de ces axiomes d'inférence.

Dans cette note, nous proposons une base des règles valides au sens de $M_{GK}$ pour un seuil minimum $\alpha \in [0,1]$ donné. Cette base est la réunion de quatre bases : la base BPE des règles positives exactes (*i.e.* les règles $X \rightarrow Y$ telles que $M_{GK}(X \rightarrow Y) = 1$), la base BNE des règles négatives exactes (*i.e.* les règles $X \rightarrow \overline{Y}$ telles que $M_{GK}(X \rightarrow \overline{Y}) = 1$), la base BPA des règles positives approximatives (*i.e.* les règles $X \rightarrow Y$ telles que $\alpha \leq M_{GK}(X \rightarrow Y) < 1$ ) et la base BNA des règles négatives approximatives (*i.e.* les règles $X \rightarrow \overline{Y}$ telles que $\alpha \leq M_{GK}(X \rightarrow \overline{Y}) < 1$).

### 3.1. *Base des règles positives exactes*

Rappelons que les règles positives exactes au sens de la mesure $M_{GK}$ sont les règles $X \rightarrow Y$ telles que $M_{GK}(X \rightarrow Y) = 1$. Or l'égalité $M_{GK}(X \rightarrow Y) = 1$ est équivalente à $conf(X \rightarrow Y) = 1$. Donc l'ensemble des règles positives exactes au sens de la mesure $M_{GK}$ est identique à l'ensemble des règles positives exactes au sens de la mesure confiance. Par conséquent, la base de Guigues-Duquenne [GUI 86] pour les règles positives exactes au sens de la confiance est une base pour les règles positives exactes au sens de $M_{GK}$, par rapport aux axiomes d'inférence de Armstrong [ARM 74] suivants : (PE1) pour tout motif $X$, $X \rightarrow X$ ; (PE2) si $X \rightarrow Y$ et $Y \rightarrow Z$, alors $X \rightarrow Z$ ; (PE3) si $X \rightarrow Y$ et $Z \rightarrow T$, alors $(X \cup Z) \rightarrow (Y \cup T)$.
Considérons les applications $f$ et $g$ définies par $f : E \longmapsto E'$ pour tout $E \subseteq \mathcal{E}$ avec $E' = \{x \in \mathcal{V} : x(e) = 1 \forall e \in E\}$ et $g : X \longmapsto X'$ pour tout $X \subseteq \mathcal{V}$. Alors l'application $\varphi = f \circ g$ est un opérateur de fermeture sur $\mathcal{P}(\mathcal{V})$, *i.e.* $\varphi$ vérifie les trois conditions suivantes : (F1)$X \subseteq Y$ implique $\varphi(X) \subseteq \varphi(Y)$, (F2) $X \subseteq \varphi(X)$, (F3) $\varphi(\varphi(X)) = \varphi(X)$. Un motif $X$ sera dit $\varphi$-fermé si $\varphi(X) = X$. Un motif $X$ est dit $\varphi$-critique s'il n'est pas $\varphi$-fermé et $\varphi(Z) \subset X$ pour tout motif $Z$ $\varphi$-critique strictement contenu dans $X$ [CAS 03]. Une autre caractérisation des motifs $\varphi$-critiques peut être trouvée dans [DIA 05].
La base de Guigues -Duquenne pour les règles exactes au sens de la confiance, donc pour les règles positives exactes au sens de $M_{GK}$ est définie par :

$$\text{BPE} = \{X \rightarrow \varphi(X) \setminus X : X \; est \; \varphi - critique\}.$$

### 3.2. *Base des règles négatives exactes*

Le résultat suivant permet de caractériser les règles négatives exactes au sens de la mesure $M_{GK}$ en fonction du support des règles positives correspondantes.

**Proposition 1** *Soient $X$ et $Y$ deux motifs tels $supp(X) \neq 0$ et $supp(Y) \neq 0$. Alors $M_{GK}(X \rightarrow \overline{Y}) = 1 \Leftrightarrow M_{GK}(X \rightarrow Y) = -1 \Leftrightarrow conf(X \rightarrow Y) = 0 \Leftrightarrow supp(X \rightarrow Y) = 0$.*

Le résultat de la proposition 1 nous conduit à considérer la bordure positive de l'ensemble des motifs de support nul noté $Bd^+(0)$ [MAN 97]. La bordure positive des motifs de support nul est l'ensemble des motifs maximaux (au sens de l'ordre d'inclusion) de support non nul, *i.e.*, formellement :

$$Bd^+(0) = \{X \subseteq \mathcal{V} : supp(X) > 0 \; et \; pour \; tout \; x \notin X, \; supp(X \cup \{x\}) = 0\}.$$

Considérons maintenant les axiomes d'inférence ci-après : (NE1) si $X \rightarrow \overline{Y}$, alors pour tout motif $T$ tel que $supp(YT) > 0$ on a $X \rightarrow \overline{YT}$ ; (NE2) si $X \rightarrow \overline{Y}$, alors pour tout $Z \subset X$ tel que $supp(ZY) = 0$, on a $Z \rightarrow \overline{Y}$. Alors on montre que les axiomes (NE1) et (NE2) sont correctes, c'est-à-dire que toute règle d'association déduite par application de (NE1) et/ou (NE2) à partir d'une ou plusieurs règles d'association exactes au sens de $M_{GK}$ est négative exacte au sens de $M_{GK}$. Par ailleurs, on montre que l'ensemble

$$\text{BNE} = \{X \rightarrow \overline{\{x\}} : X \in Bd^+(0) \text{ et } x \notin X\}$$

est une base pour les règles négatives exactes au sens de $M_{GK}$ par rapport aux axiomes d'inférence (NE1) et (NE2).

### 3.3. *Base des règles positives approximatives*

Notons qu'une règle valide au sens de la mesure de qualité $M_{GK}$ est nécessairement une règle où la prémisse favorise la conclusion. En effet, $X$ défavorise $Y$ signifie que la réalisation $X$ diminue la chance de $Y$ d'être réalisé. Dans ce cas il est alors plus pertinent de considérer la règle $X \rightarrow \overline{Y}$ puisque $X$ favorise $\overline{Y}$ lorsque $X$ défavorise $Y$. Soit $\alpha \in [0, 1]$ un nombre réel. Le résultat suivant caractérise les règles positives approximatives $\alpha$-valides au sens de $M_{GK}$ en fonction de leur confiance.

**Proposition 2** *Soient $X$ et $Y$ deux motifs tels que $X$ favorise $Y$. Alors $\alpha \leq M_{GK}(X \rightarrow Y) < 1$ si et seulement si $supp(Y)(1-\alpha)+\alpha \leq conf(X \rightarrow Y) < 1$.*

Considérons maintenant l'axiome d'inférence (PA) ci-dessous :
(PA) si $X \rightarrow Y$ alors pour tous $Z, T$ tels que $\varphi(X) = \varphi(Z)$ et $\varphi(Y) = \varphi(T)$ on a $Z \rightarrow T$.
Alors on montre que (PA) est correct, c'est-à-dire que toute règle d'association déduite par application de (PA) à partir d'une règle positive approximative au sens de $M_{GK}$ est positive approximative au sens de $M_{GK}$. Par ailleurs, on montre que l'ensemble

$$\text{BPA} = \{X \rightarrow Y : \varphi(X) = X, \varphi(Y) = Y, supp(Y)(1-\alpha)+\alpha \leq conf(X \rightarrow Y) < 1\}$$

est une base pour les règles d'associaition positives approximatives au sens de $M_{GK}$, par rapport à l'axiome d'inférence (PA).

### 3.4. *Base des règles négatives approximatives*

Le résultat suivant caractérise les règles négatives approximatives $\alpha$-valides au sens de $M_{GK}$ en fonction de la confiance des règles positives correspondantes.

**Proposition 3** *Soient $X$ et $Y$ deux motifs tels que $X$ favorise $\overline{Y}$. Alors $\alpha \leq M_{GK}(X \rightarrow \overline{Y}) < 1$ si et seulement si $0 < conf(X \rightarrow Y) \leq supp(Y)(1-\alpha)$.*

Considérons enfin l'axiome d'inférence (NA) ci-dessous :
(NA) si $X \rightarrow \overline{Y}$, alors pour tous $Z, T$ tels que $\varphi(X) = \varphi(Z)$ et $\varphi(Y) = \varphi(T)$, on a $Z \rightarrow \overline{T}$.
On montre que (NA) est correct, c'est-à-dire que toute règle d'association déduite par application de (NA) à partir d'une règle négative approximative au sens de $M_{GK}$ est négative approximative au sens de $M_{GK}$. Par ailleurs, on montre que l'ensemble

$$\text{BNA} = \{X \rightarrow \overline{Y} : \varphi(X) = X, \varphi(Y) = Y, 0 < conf(X \rightarrow Y) \leq supp(Y)(1-\alpha)\}$$

est une base pour les règles d'association négatives approximatives au sens de $M_{GK}$, par rapport à l'axiome d'inférence (NA).

En résumé la base des règles d'association que nous proposons est la réunion des bases BPE, BNE, BPA et BNA.

## 4. Bibliographie


[AGR 93] AGRAWAL R., IMIELINSKI T., SWAMI A., Mining association rules between sets of items in large databases, BUNEMAN P., JAJODIA S., Eds., *Proc. of the ACM SIGMOD International Conference on Management of Data*, vol. 22, Washington, 1993, ACM press, p. 207–216.

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

[CAS 03] CASPARD N., MONJARDET B., The lattices of closure systems, closure operators, and implicational systems on a finite set : a survey, *Discrete Applied Mathematics*, vol. 127, 2003, p. 241–269.

[DIA 05] DIATTA J., Charactérisation des ensembles critiques d'une famille de Moore finie, *Rencontres de la Société Francophone de Classification*, Montréal, Canada, 2005, p. 126–129.

[FEN 06] FENO D., DIATTA J., TOTOHASINA A., Normalisée d'une mesure probabiliste de qualité des règles d'association : étude de cas, *Actes du 2nd Atelier Qualité des Données et des Connaissances*, Lille, France, 2006, p. 25–30.

[GUI 86] GUIGUES J. L., DUQUENNE V., Famille non redondante d'implications informatives résultant d'un tableau de données binaires, *Mathématiques et Sciences humaines*, vol. 95, 1986, p. 5–18.

[GUI 00] GUILLAUME S., Traitement des données volumineuses. Mesures et algorithmes d'extraction des règles d'association et règles ordinales, PhD thesis, Université de Nantes, France, 2000.

[MAN 97] MANNILA H., TOIVONEN H., Levelwise search and borders of theories in knowledge discovery, *Data Mining Knowledge Discovery*, vol. 1, 1997, p. 241–258.

[PAS 99] PASQUIER N., BASTIDE Y., TAOUIL R., LAKHAL L., Efficient mining of association rules using closed itemset lattices, *Information Systems*, vol. 24, 1999, p. 25–46.

[PEI 00] PEI J., HAN J., MAO R., CLOSET : An Efficient Algorithm for Mining Frequent Closed Itemsets, *ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, 2000, p. 21-30.

[TOT 05] TOTOHASINA A., RALAMBONDRAINY H., ION : A pertinent new measure for mining information from many types of data, *IEEE SITIS'05*, 2005, p. 202–207.

# Illustration d'une méthode d'évaluation supervisée par un problème de classification de courbes

**Sylvain Ferrandiz, Marc Boullé**

*France Télécom R&D*
*2, avenue Pierre Marzin, 22300 Lannion*
*sylvain.ferrandiz@francetelecom.com*
*marc.boulle@francetelecom.com*

RÉSUMÉ. *La récolte des données est de moins en moins contrainte par l'aspect technique de sa mise en œuvre. En conséquence, il est aujourd'hui possible de suivre dans le temps toute caractéristique mesurée. Lors de la préparation d'une table de données et de la construction d'un modèle, le statisticien doit ainsi compter avec la présence d'un nombre croissant de variables dynamiques. A côté des questions usuellement traitées en phase de préparation, comme la sélection des variables pertinentes, toute variable dynamique soulève de plus un problème de représentation. Afin d'automatiser la prise de décision, aujourd'hui basée sur la connaissance métier, nous appliquons une méthode d'évaluation supervisée pour quantifier la pertinence d'une variable dynamique. L'évaluation est automatique et régularisée, ce qui profite à la qualité des choix opérés. Le propos est illustré sur un problème de classification de courbes de consommation téléphonique.*

MOTS-CLÉS : *Classification supervisée, préparation de données, variables dynamiques.*

## 1. Préparation de données et variables dynamiques

Avec l'émergence des systèmes d'information au tournant des années 90, la récolte des données brutes a été rendue complètement indépendante de toute finalité statistique. Modéliser directement de telles données est devenu impossible. La phase de préparation des données, dont l'objectif est de construire une table de données pour modélisation à partir des données brutes, est donc devenue une partie critique et souvent coûteuse en temps du processus de fouille de données [CHA 00].

L'évolution des moyens techniques le permettant, il est aujourd'hui possible de suivre dans le temps une caractéristique, et ce sur une longue période. A côté des variables usuelles, qu'on qualifie ici de *statiques*, sont donc de plus en plus présentes des variables *dynamiques* : mesure de l'activité cardiaque en médecine, mesure de la pression en météorologie, mesure de la consommation téléphonique en télécommunication, mesure de la propagation des ondes en sismologie. Une variable dynamique est ainsi formée par une suite de mesures et se distingue d'une variable multivariée par son caractère séquentiel.

De par la décorrélation entre la récolte des données et la modélisation, la précision des mesures et l'échelle des temps de mesure sont sans aucun rapport avec les besoins d'une étude statistique, car seulement limitées par les contraintes techniques. Dès lors, la préparation de données dynamiques passe nécessairement par une phase de recherche de représentation : d'une représentation brute (les données observées) il faut passer à une représentation cohérente pour modélisation subséquente, et ce pour chacune des variables dynamiques. Au cours de cette recherche de représentation, d'autres problèmes que celui de l'échelle des temps de mesure sont à traiter, comme le bruit sur les mesures, le non alignement des temps de mesure, le facteur d'échelle entre les individus, etc.

En phase de préparation, hormis le problème de représentation, les variables dynamiques sont placées dans le même contexte que les variables statiques et sont naturellement amenées à subir les mêmes traitements. Si on cherche à expliquer une variable cible symbolique, les tâches principales de préparation sont l'évaluation de la dépendance entre variable(s) explicative(s) et variable cible, ainsi que la sélection de variables explicatives.

En pratique, la sélection d'une représentation pour chaque variable dynamique et la sélection de variables dynamiques sont basées sur la connaissance métier : en reconnaissance de la parole, il est "usuel" de travailler avec les log-périodogrammes des signaux ; en téléphonie, il est "usuel" de travailler avec un découpage en tranches horaires prédéterminé ; en médecine, il est "usuel" de considérer à la fois un électroencéphalogramme, un électro-oculogramme et un électromyogramme afin d'étudier la phase de sommeil paradoxal. La connaissance métier porte sur un phénomène particulier et s'accumule au fur et à mesure qu'on valide de nouvelles hypothèses sur ce phénomène.

On applique ici une méthode d'évaluation non paramétrique et générique jugeant la qualité d'une représentation à l'étude d'une variable dynamique. Ses qualités favorisent l'automatisation de la prise de décision et la rendent indépendante du domaine d'application. Le contexte est celui de la classification supervisée. Le critère d'évaluation $c$ est introduit dans [FER 06b] et l'algorithme d'optimisation dans [FER 06a]. On se propose dans ce papier d'illustrer l'apport de la méthode sur un problème d'évaluation supervisée de variables dynamiques. La section 2 dérive du critère $c$ une méthode d'évaluation de la pertinence d'une représentation. La section 3 montre son apport à travers une expérimentation sur un problème de classification de profils de consommation téléphonique.

## 2. Une approche informationnelle de l'évaluation

Dans le cas d'une variable statique continue, [BOU 06] aborde la question de l'évaluation de la pertinence vis-à-vis d'une variable cible symbolique comme un problème de modélisation. Les modèles considérés sont les partitions de la variable continue en intervalles. Une approche informationnelle permet de définir un critère s'interprétant comme la probabilité que le modèle explique les données. La sélection du modèle le plus probable conduit à une méthode de discrétisation d'une variable statique continue. La probabilité que ce modèle explique les données s'utilise alors comme un indicateur de pertinence de la variable descriptive relativement à la variable cible.

Dans [FER 06b], l'approche est adaptée afin de traiter le cas où l'on dispose d'une mesure de similitude entre les instances de l'échantillon. Toute partition de Voronoi induite par un sous-ensemble d'instances constitue un modèle. Le partitionnement d'une variable en intervalles est ainsi généralisé en un partitionnement de l'espace en cellules. La probabilité qu'un modèle explique les données est explicitée et la sélection du modèle le plus probable conduit à une méthode de sélection d'instances. Là encore, la probabilité associée au modèle sélectionné constitue un indicateur supervisé de pertinence. La méthode est évaluée dans [FER 06a] en tant que méthode de sélection d'instances pour la classification par le plus proche voisin.

En pratique, la représentation des données dynamiques conduit à définir une matrice de similitude. Par exemple, une fois la transformée de Fourier appliquée, il est usuel d'utiliser une distance euclidienne pondérée. On considère donc qu'une représentation $R$ n'est autre qu'une matrice de similitude. Ainsi, on utilise le critère $c(M, R)$ présenté dans [FER 06b], qui mesure de manière supervisée l'intérêt d'une partition de Voronoi $M$ relative à un sous-ensemble de l'échantillon et définie à l'aide de la matrice $R$. Dès lors, si on note

$$c^*(R) = min_M\ c(M, R),$$

la fonction $c^*$ fournit une évaluation de la qualité de la représentation $R$ et permet ainsi de comparer différentes représentations. Pour une matrice de similitude $R$ donnée, il suffit d'appliquer un algorithme d'optimisation combinatoire et d'attribuer à $R$ la valeur rencontrée optimale du critère $c(M, R)$. Une heuristique d'optimisation efficace est présentée dans [FER 06a]. Le critère $c$ est le suivant :

$$c(M, R) = \log N + \log \binom{N + K - 1}{K} + \sum_{k=1}^{K} \log \binom{N_k + J - 1}{J - 1} + \sum_{k=1}^{K} \log \frac{N_k!}{N_{k1}! \ldots N_{kJ}!}.$$

où $N$ désigne le nombre d'instances de l'échantillon, $J$ le nombre de classes cibles, $K$ le nombre de groupes de la partition, $N_k$ le nombre d'instances dans la $k^{eme}$ cellule et $N_{kj}$ le nombre d'instances dans la $k^{eme}$ cellule portant la $j^{eme}$ étiquette.

Le critère $c$ est non paramétrique et régularisé. Il quantifie le compromis entre le nombre de groupes de la partition (les deux premiers termes) et la distribution de la cible (les deux derniers termes), ce qui correspond à un compromis entre complexité du modèle et ajustement du modèle aux données de l'échantillon. La régularisation est un moyen sûr d'endiguer le phénomène de sur-apprentissage. Etant de surcroît non paramétrique, l'évaluation se passe de validation ou de validation croisée. On dispose ainsi de plus d'instances pour ajuster le modèle, ce qui augmente sa qualité.

Afin de travailler avec un indicateur normalisé, on considère la transformation suivante de $c^*$ :

$$g^*(R) = 1 - \frac{c^*(R)}{c_0(R)},$$

où $c_0(R)$ est la valeur du critère $c(M, R)$ pour le modèle $M$ constitué par un seul groupe. Comme $c(M, R)$ n'est autre que l'opposé du logarithme d'une probabilité et comme une telle quantité s'interprète comme une longueur de codage (d'après les travaux de [SHA 48]), $g^*(R)$ mesure un gain de compression. Le gain de compression $g^*(R)$ est supérieur à 0 (dès lors que la partition en un unique groupe est évaluée durant l'optimisation) et inférieur à 1. Si $g^*(R) = 0$, la représentation $R$ n'apporte aucune information sur la variable cible. Plus la valeur de $g^*(R)$ est proche de 1, plus les classes cibles sont séparées. Il est à noter que ce critère est générique et ne se limite pas à l'évaluation de représentations de données dynamiques.

## 3. Classification de profils de consommation

On illustre les apports de notre méthode par une expérimentation sur des données de consommation en téléphonie fixe. C'est un problème de classification de profils de consommation suivant 4 classes cibles A, B, C et D. La distribution des classes cibles est uniforme sur l'échantillon. On dispose de 168 variables descriptives continues, chacune mesurant la consommation téléphonique sur une tranche horaire de la semaine et ce pour 2636 instances. On applique la méthode à la variable dynamique constituée par les 168 variables descriptives. La mesure de similitude adoptée est la métrique $L_1$.

L'évaluation fournit un gain de compression de 0.051, ce qui est très faible et caractérise un fort mélange des classes cibles. En plus de quantifier la pertinence d'une variable, elle fournit un support à la discrimination réalisée : une distribution des classes cibles et un prototype accompagnent chaque groupe. La méthode partitionne ici les instances en 7 groupes et les caractéristiques de trois d'entre eux sont reportées sur la figure 1. Les distributions relatives à chacun de ces groupes sont représentées par des histogrammes groupés. Dans chaque groupe, en calculant la valeur moyenne de chacune des 168 variables, on obtient un profil de consommation moyen caractéristique de ce groupe. Ce profil est plus parlant que le simple profil de consommation du prototype car il tient compte de toutes les instances du groupe.

Le modèle étant visualisable, il est facilement interprétable. Par exemple, on voit que les individus du groupe 7 sont en grand nombre (35% des instances), qu'ils une consommation moyenne plus élevée que la moyenne globale, et que ce comportement est majoritairement caractéristique de la classe A (la répartition dans les classes cibles A, B, C, D est $(41\%, 26\%, 15\%, 17\%)$). Le groupe 1 est quant à lui plus discriminant (la répartition dans les classes cibles est $(16\%, 20\%, 57\%, 6\%)$) avec un profil de consommation atypique (pics de consommation élevés), mais est de taille réduite ($4\%$ des instances). Le groupe 4 discrimine lui aussi la classe C, moins fortement tout de même que le groupe 1, et se différentie par une consommation moyenne très faible.

Il est à noter que ce n'est pas la visualisation en elle-même qui est nouvelle. Elle peut en effet être utilisée conjointement à toute méthode fournissant un ensemble de prototypes. La nouveauté réside dans le fait que la méthode d'évaluation proposée ici optimise exactement les paramètres de cette visualisation. En effet, les ca-

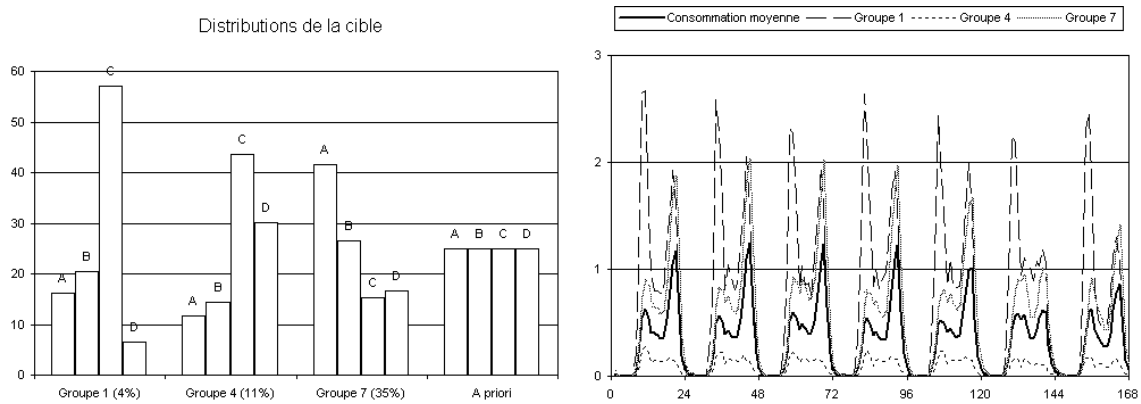



**FIG. 1.** *Caractéristiques des groupes 1, 4 et 7. A chaque profil est associée une distribution des classes cibles. Les groupes 1 et 4 contiennent majoritairement des instances de classe C. Les instances du groupe 1 correspondent à de très fortes consommations, avec des pics très marqués. Celles du groupe 4 correspondent aux faibles consommations.*

ractéristiques visualisées (taille des groupes, distribution des classes cibles dans les groupes, prototypes) sont exactement celles apparaissant dans le critère et les modèles. La visualisation n'en est que plus pertinente.

## 4. Conclusion

A cours d'un processus de fouille de données, le statisticien traite aujourd'hui aussi bien des variables statiques que des variables dynamiques. Ces dernières soulèvent avec plus de force la question du choix d'une représentation. Nous avons appliqué dans ce papier un procédé d'évaluation supervisée de la pertinence d'une représentation.

En évitant de passer par une phase de validation, on utilise plus de données pour l'apprentissage. Ceci assure une pertinence plus forte de l'hypothèse validée. En gérant le sur-apprentissage, l'hypothèse sélectionnée ne "colle" pas aux données, ce qui assure sa robustesse. C'est l'adoption d'une approche informationnelle qui conduit à une telle évaluation.

Le bon comportement attendu a été illustré sur un jeu de données réel, dans le but de classer de manière supervisée des courbes de consommation téléphonique. Les expérimentations ont montré l'apport explicatif de la structuration des données : la méthode optimise ce que l'utilisateur voit. De par sa généricité, l'évaluation n'est pas limitée au problème de représentation des données dynamiques.

## 5. Bibliographie

[BOU 06] BOULLÉ M., MODL : a bayes optimal discretization method for continuous attributes, *Machine learning*, , A paraître en 2006.

[CHA 00] CHAPMAN P., CLINTON J., KERBER R., KHABAZA T., REINARTZ T., SHEARER C., WIRTH R., CRISP-DM 1.0 : step-by-step data mining guide, 2000.

[FER 06a] FERRANDIZ S., BOULLÉ M., Sélection supervisée d'instances : une approche descriptive, *Actes de la conférence sur l'extraction et la gestion des connaissances*, vol. 2, 2006, p. 421–432.

[FER 06b] FERRANDIZ S., BOULLÉ M., Supervised evaluation of Voronoi partitions, *Journal of intelligent data analysis*, , A paraître en 2006.

[SHA 48] SHANNON C., A mathematical theory of communication, rapport, 1948, Bell systems technical journal.

# Modélisation de la synchronisation du réseau des routes aériennes en Europe associée avec une approche Data Mining

**Trung Tuyen HOANG, Henri LY, Tao PHAM DINH**

*EUROCONTROL Experimental Centre, 91222 Brétigny sur Orge, France*
*LMI-Institut National des Sciences Appliquées, Rouen-France*
*{trung-tuyen.hoang.ext, henri.ly}@eurocontrol.int, pham@insa-rouen.fr*

RÉSUMÉ. *La recherche opérationnelle est un outil puissant pour contribuer à la modélisation des problèmes de l'ATM . Cependant, associée à un nombre élevé de contraintes et des données, l'élaboration d'un modèle mathématique devient complexe et difficile. En effet ce problème est très sensible à la fluctuation des données et des éléments de contraintes (constituants des noeuds et des arcs du réseau). Ces derniers, répartis à travers différentes bases de données des " Stakeholder " ATM , doivent être vérifiés et validés pour éviter l'accumulation des incertitudes, des erreurs ainsi que la génération des résultats biaisés à la sortie du modèle. De ce fait, la phase de prétraitement des données est primordiale : les techniques et recherches actuelles de Data Mining et de classification pourront contribuer indéniablement à la fiabilité des données du modèle.*

MOTS-CLÉS : *ATM,Synchronisation, Data Mining, Modèle Linéaire*

## 1. Description du problème

En l'état actuel, pour chaque vol, il est alloué une fenêtre de temps entre -5 et +10 minutes pendant laquelle le contrôleur est libre de décider si l'avion est autorisé à décoller ou à atterrir. Cette marge est bonne au niveau de sécurité mais améliorable au niveau de l'efficacité économique vis-à-vis des retards et de la capacité de gestion (nombre d'aéronef géré). Donc actuellement, il existe des équipes qui vont procéder des simulations ayant pour objectif de réduire la marge de la fenêtre sans baisser le niveau concernant la sécurité. Par exemple on fixe une marge de -2 minutes et +3 minutes concernant la tolérance d'arrivée et comment on synchronise le trafic ?

Synchronisation du trafic aérien : c'est le processus d'établissement et de maintenance de la sécurité, de l'efficacité et de l'ordre du flux de trafic. Elle inclut la provision, le traitement de la file d'attente des avions en route ainsi qu'au sol. Elle opère également sur les vols individuels.

## 2. Modélisation Mathématique

Compte tenu que le but est de fournir un modèle mathématique pour soutenir à cette procédure (reserrer la fenêtre temporelle et synchroniser le réseau) le TTA sera pris comme l'élément principal d'input, puis on remonte sur l'itinéraire du vol pour déterminer le créneau de départ de l'aéronef. D'autre part, en fixant le TTA, l'aéroport de départ, l'aéroport d'arrivée, on devrait pouvoir déterminer l'heure de départ et la route correspondante de telle sorte que le vol arrive dans la fenêtre temporelle prédéterminée. En premier temps, on va concentrer sur la détermination de temps de départ en fixant le temps d'arrivée. La route est considérée comme connue.

### 2.1. *Les paramètres de la modélisation*

On reprise les notations de Bertsimas et Stock ([1]). Notre principe étant suivant : pour chaque TTA, on estime le TTOT. On revient au problème de minimisation du coût de service (dans notre cas, c'est la minimisation du coût total des retards) sous les constraintes de capacités et surtout la constrainte de fenêtre temporelle d'arrivée.Une fois on a obtenu des solutions optimales, on peut mettre à jour TTA et ainsi que l'heure de passage dans les secteurs.

Soient : L'ensemble des vols $\mathcal{F} = \{1, ..., F\}$. L'ensemble des aéroports $\mathcal{K} = \{1, ..., K\}$ On note également des périodes de temps $\mathcal{T} = \{1, ..., T\}$, On prend comme référence le temps $t$ pour désigner la période de temps $t$. Donc les données du problème sont suivantes :

$N_f$ = nombre de secteurs dans la trajectoire du vol $f$
$P(f, i)$= le $i^{eme}$ secteur dans le vol $f$
$P_f = (P(f, i) : 1 < i < N_f)$= ensemble des secteurs dans le vol $f$
$D_k(t)$= la capacité des vols déclarée de départ de l'aéroport $k$ à $t$
$A_k(t)$= capacité des vols déclarés d'arrivée de l'aéroport $k$ à $t$
$S_j(t)$=capacité de secteur $j$ à $t$
$d_f$= heure de départ programmée (estimée) de vol $f$
$r_f$= heure d'arrivée programmée (voulue) de vol $f$
$c_f^a$ = le coût du maintien du vol $f$ en route pour une unité de temps
$c_f^g$ =le coût de maintien du vol $f$ au sol pour une unité de temps
$l_{fj}$= le temps minimal pour le vol (l'aéronef) $f$ traverse le secteur $j$
$T_f^j$= l'ensemble de temps que le vol $f$ peut être entré dans le secteur $j$

Les variables de décision sont les suivantes :

$$w_{ft}^j = \begin{cases} 1 & \text{si le vol } f \text{ entre au secteur } j \text{ avant } t \\ 0 & \text{sinon} \end{cases}$$

et

$$u_{ft}^j = \begin{cases} 1 & \text{si le vol } f \text{ entre au secteur } j \text{ à } t \\ 0 & \text{sinon} \end{cases}$$

De cette définition, on a :

$$u_{ft}^j = w_{ft}^j - w_{f,t-1}^j, \text{ donc } w_{ft}^j = \sum_{t'<t} u_{ft'}^j$$

### 2.2. *Les contraintes*

1. Le premier secteur du vol étant l'aéroport de départ, par conséquent le temps que le vol est maintenu au sol est égal à la différence entre l'heure de départ réelle et l'heure de départ programmée.

$$g_f = \sum_{t \in T_f^k, k=P(f,1)} t u_{ft}^k - d_f = \sum_{t \in T_f^k, k=P(f,1)} t(w_{ft}^j - w_{f,t-1}^j) - d_f$$

2. Le temps que le vol f est maintenu en route peut être interprété comme l'heure d'arrivée réel diminué de l'heure d'arrivée programmée et aussi diminué du temps que le vol $f$ maintenu au sol ( avant de départ)

$$a_f = \sum_{t \in T_f^k, k=P(f,N_f)} t u_{ft}^k - r_f - g_f = \sum_{t \in T_f^k, k=P(f,N_f)} t(w_{ft}^j - w_{f,t-1}^j) - r_f - g_f$$

3. La déviation des retards d'arrivée est égale à la différence entre l'heure d'arrivée réelle et l'heure d'arrivée programmée.

Donc

$$p_f = \sum_{t \in T_f^k, k=P(f,N_f)} tu_{ft}^k - r_f = \sum_{t \in T_f^k, k=P(f,N_f)} t(w_{ft}^j - w_{f,t-1}^j) - r_f$$

4. Compte tenu que le nombre de vols de départ d'un aéroport ne peut dépasser la capacité des vols de départ de cet même aéroport, on a l'inégalité suivante

$$\sum_{f:P(f,1)=k} (w_{ft}^k - w_{f,t-1}^k) \leq D_k(t), k \in \mathcal{K}, t \in \mathcal{T}$$

5. Le nombre d'aéronefs d'arrivée sur un aéroport à un moment donné ne devrait pas dépasser la capacité de cet aéroport à ce moment là, on a l'inégalité suivante :

$$\sum_{f:P(f,N_f)=k} (w_{ft}^k - w_{f,t-1}^k) \leq A_k(t), k \in \mathcal{K}, t \in \mathcal{T}$$

6. Le nombre d'avions traversant un secteur à un moment donné ne devrait pas dépasser la capacité de ce secteur, mathématiquement on a la relation suivante :

$$\sum_{f:P(f,i)=j,P(f,i+1)=j',i<N_f} (w_{ft}^j - w_{f,t-1}^{j'}) \leq S_k(j), j \in \mathcal{K}, t \in \mathcal{T}$$

7. $a \leq p_f \leq b$, cette condition signifie que ce vol arrive dans la fenêtre temporelle prédéterminée

8. $w_{f,t}^j - w_{f,t-1}^j \geq 0, \quad f, j \in P_f, \quad t \in T_f^j$

9. $w_{ft}^j \in \{0,1\}, \quad f,j \in P_f, t \in T_f^j$

### 2.3. *Programme d'optimisation*

Notre critère est de minimiser du coût total des retards qui peut être décomposée en deux composantes : le coût de retard en route, le coût de retard au sol

$$Min \sum_f [c_f^g g_f + c_f^a a_f]$$

En rempplaçant les expressions de $g_f$ et $a_f$ on a :

$$Min \sum_f [c_f^g ( \sum_{t \in T_f^k, k=P(f,1)} t(w_{ft}^j - w_{f,t-1}^j) - d_f) + c_f^a ( \sum_{t \in T_f^k, k=P(f,N_f)} t(w_{ft}^j - w_{f,t-1}^j) - r_f - g_f)]$$

L'heure de départ programmée peut être estimée comme

$$d_f = min\{t : t \in T_f^k, k = P(f,1)\}$$

## 3. Validation de la modélisation avec les donénes fiables

### 3.1. *La nécessité de vérification des données*

Les problèmes de fiabilité des données issues des partenaires sont réels. En effet, ces données ne sont pas homogènes, de nature différente, gérées et utilisées par différents acteurs qui n'ont pas :

- La même vision des priorités ;
- Les mêmes contraintes ;
- Les mêmes préoccupations ni
- Les mêmes types de systèmes gestionnaires des bases de données (SGBD).

### 3.2. *Les méthodes de vérification de données*

*Classification*

Compte tenu du grand volume d'information disponible dans le monde ATM, il nous importe de classifier (segmenter) les types d'information au sein desquels on peut distinguer des sous ensembles homogènes pour le traitement et d'analyses différenciés afin de pouvoir les intégrer dans notre modèle de réseau.

Donc au premier abord, nous cherchons à identifier et classifier les éléments contraignants rentrant en jeu pour les noeuds et les arcs pour ce modèle de synchronisation. Une première recherche nous permet d'établir une liste des éléments suivants :

- Le nombre de vols décollé (au départ) d'un aéroport $k$ à l'instant $t$
- Le nombre de vols à l'arrivée d'un aéroport $k$ à l'instant $t$
- Le nombre total des vols gérant par le secteur donné
- Le nombre de vols spéciaux
- Le nombre de terminaux (aéroports)
- La distribution du trafic
- La fréquence de congestion
- La séparation minimale en vertical et à l'horizontal (fonction même du type de l'aéronef)
- Les conditions météorologiques

*Prioritisation*

La prioritisation consiste à identifier les éléments parmi la liste ci-dessus (pouvant jouer un rôle prépondérant croissant) pouvant influencer dynamiquement le modèle. Enfin, il convient de trouver aussi, si nécessaire, des paramètres de pondérations pour chaque élément $i$ de cette liste. Cette démarche pourra être identifiée et faite par les chercheurs.

D'autre part, la classification ainsi que la prioritisation pourront aussi être constituées après une enquête auprès des ANSPs et les autres partenaires afin de déterminer un consensus des éléments choisis parmi la liste ci-dessus. On établiera un questionnaire de type Likert.

Cependant, afin de tester sur la cohérence et la fiabilité des échelles, on pratiquera un test avec $\alpha$ de Cronbach. Si $\alpha \geq 0,7$ la fiabilité sera démontrée. Enfin, on fera aussi un test de $\chi^2$ d'indépendance (concernant les données des interviews recueillies auprès des stakeholders) pour démontrer que les réponses obtenues auprès des partenaires ne seront pas dûes à un effet de hasard et que les variables sont en relation. Le $\chi^2$ calculé sera comparé au $\chi^2$ lu avec un degré de liberté $i$ correspondant et $\alpha = 0,05$ ($p = 0,95$)

*Vérification*

La vérification et la validation des données sont primordiales. L'injection des données sans vérification et calcul de probabilité induiront des risques pour chaque élément de contraintes et par voie de conséquence à la génaration des résultats biaisés à la sortie du modèle. Pour l'instant aucun modèle (du processus data mining) n'a été choisi pour vérifier les données obtenues. Néanmoins, plusieurs pistes sont possibles à savoir l'adoption d'un ou plusieurs modèles combinés : modèle linéaire, modèle quadratique, modèle cubique ...

Néanmoins, très rapidement ce projet de modélisation de data mining doit être concrétisé avec le respect des principes importants pour fiabiliser, harmoniser et valider des diverses données obtenues.

Ce projet n'échappe pas à la règle 20/80 de PARETO. En effet, notre planification est similaire au tableau* (à quelques pourcentages près) qui permet d'identifier et de comparer le pourcentage de temps accordé par rapport aux taux de succès de chaque étape.

Ainsi la boucle constituant la recherche des solutions, la préparation des données, les accès aux données, leurs modélisations, les vérifications des données et leur injection dans le modèle synchronisation du réseau sera complète et bouclée.

| | Temps accordé en % | | Taux de succès en % | Taux par grande phase |
|---|---|---|---|---|
| 1.Analyse du problème | 10 | 20 | 15 | 80 |
| 2.Recherche et analyse de solutions possibles | 9 | | 14 | |
| 3.Implémentations des spécifications | 1 | | 51 | |
| 4.Data Mining | | 80 | | 20 |
| 4.a Préparation des données | 60 | | 15 | |
| 4.b Etude, Analyse | 15 | | 3 | |
| 4.c Modélisation des données | 5 | | 2 | |

**TAB. 1.** **Tableau extrait du document de Dorian Pyle*

## 4. Conclusions et Perspectives

Le problème de synchronisation de trafic joue un rôle central dans la gestion du trafic aérien. Il permet de bien gérer, controler le trafic ainsi que la ponctualité des vols. Une modélisation mathématique de type stochastique semble nécessaire mais la technique des fouilles des données permettra de traiter le prolème plus efficace. Les recherches sont en cours et un document spécifique leur sera entièrement consacrées ultérieurement.

## 5. Glossaire

**ANSPs** : ATM National Provider : Fournisseurs nationaux des services de gestion du trafic aérien.

**ATC** : Air traffic Control : Contrôle de trafic aérien

**ATFM** : Air Traffic Flow Management : Gestion des courants de trafic

**ATM** : Air Traffic Management, Gestion du trafic aérien

**C-ATM** : Co-operative Air Traffic Management

**FL** : Flight Level : niveau de vol

**TTA** : Time Target for Arrival, temps d'arrivé prévu

**TMA** : Terminal Area : Région terminale

**TTOT** : Target Take-Off Time, Temps de décollage prévu

**STAKEHOLDER** : Partenaire de l'ATM comprenant les ANPs, les compagnies aériennes, les usagers.

## 6. Bibliographie


[Ahu93] Ravindra.K.Ahuja, Thomas L.Magnati, James B.Orlin, *Network Flows : Theory, Alogorithms, and Applications*, Prentice Hall, 1993

[Stoc98] D.Bertsimas, S.Stock, *The Air Traffic Flow Management Problem with Enroute Capacities*, Operations Research, Vol. 46, pp. 406-422, 1998

[CATM05] *C-ATM Co-operative Air Traffic Management Medium Term Concept*, C-ATM Project, dec-2005.

[Mai91] George Maignan, *Le contrôle de la circulation aérienne*, PUF , Avril 1991.

[Yu01] Gang Yu, *Operations Research in the Airline Industry*, Kluwer Academic Publishers, Autum 2001

[Pyle99] Dorian Pyle, *Data Preparation for Data Mining*, Morgan Kaufmann Publisers, San Francisco, 1999.

[Ber03] M. Berthold and David Hand, *Intelligent Data Analysis*, Springer Berlin Edition, 2003

[Tuf05] Stéphane Tuffery, *Data mining et Statistique décisionnelle*, Edition Technip 2005

# Extraction de concepts guidée par le contexte

**Lobna Karoui, Nacéra Bennacer, Marie-Aude Aufaure**

*École Supérieure d'Électricité*
*Plateau de Moulon 3 rue Joliot Curie*
*91192 Gif-sur-Yvette cedex, France*

*RÉSUMÉ. Les ontologies constituent la brique supportant les échanges et le partage des informations en étendant l'interopérabilité syntaxique du web en une interopérabilité sémantique. Le succès du web sémantique dépend du degré d'automatisation de la construction des ontologies, de leur déploiement et de leur prolifération. Dans cet article, nous présentons une méthode incrémentale d'extraction de concepts ontologiques à partir de documents HTML en vue de construire une ontologie du domaine. Nous exploitons les caractéristiques structurelles des documents HTML afin de localiser et de définir un contexte approprié pour chaque terme en respectant sa position dans le corpus. Notre définition contextuelle permet de sélectionner les co-occurrents sémantiquement proches et de définir une mesure de pondération appropriée pour chaque couple de termes. Afin d'obtenir des classes de termes, nous avons défini les principes algorithmiques d'une méthode de clustering guidée par le contexte. Notre approche se base sur une évaluation interactive et incrémentale de la qualité des clusters par l'utilisateur. Nous avons expérimenté ces principes algorithmiques sur un corpus du domaine portant sur le tourisme. Les premiers résultats obtenus montrent que la prise en compte du contexte des termes améliore considérablement la pertinence des concepts extraits.*

*MOTS-CLÉS : Ontologie, Web sémantique, extraction, concept ontologique, clustering*

## 1 Introduction

Les ontologies constituent la brique supportant les échanges et le partage des informations en étendant l'interopérabilité syntaxique du web en une interopérabilité sémantique. Le succès du web sémantique dépend de la construction des ontologies. L'automatisation de cette construction semble être une solution prometteuse et de nombreux travaux existent dans la littérature. Dans [FAU 98], les auteurs présentent un système d'apprentissage, nommé ASIUM, à partir de textes techniques. Ils ont développé une méthode de clustering qui permet de classer les termes apparaissant avec le même verbe et avec le même rôle syntaxique ou la même préposition fournie par un analyseur syntaxique. De la même manière qu'ASIUM, le système SVETLAN présenté par Chalendar et Grau [CHA 00] utilise les verbes qui permettent de catégoriser les noms. Il est capable d'apprendre des catégories de noms à partir de textes, quel que soit leur domaine. Maedche et Staab [MEA 01] proposent un environnement d'apprentissage d'ontologies (Text-To-Onto) basé sur une architecture générale de découverte de structures conceptuelles à partir de différentes sources (XML, DTD, schéma de BD, etc.). Cet environnement possède une librairie de méthodes d'apprentissage et des outils linguistiques pour extraire des concepts, leurs relations taxonomiques et non taxonomiques. DODDLE II [SUG 04] est un environnement de développement d'ontologies permettant d'extraire des relations taxonomiques en utilisant à la fois les termes du domaine et WorldNet. Afin d'extraire des relations non taxonomiques, les auteurs utilisent les règles d'associations. SYNDIKATE [HAH 01] est un système pour l'acquisition automatique de connaissances à partir de textes allemands basé sur des procédures de compréhension de textes. Il extrait des relations non taxonomiques à partir de l'interprétation sémantique du texte.

Dans cet article, nous présentons une approche incrémentale d'extraction de concepts ontologiques à partir de documents HTML en vue de construire une ontologie du domaine. Cette approche s'inscrit dans le

cadre d'une méthodologie unifiée présentée dans [BEN 05]. Nous focalisons notre intérêt sur la prise en compte de la structure HTML dans le document afin de définir un contexte approprié pour chaque terme en sélectionnant ses co-occurrents sémantiquement proches et en respectant ses différentes positions dans le texte. La mesure de pondération affectée à chaque couple de termes qui découle ainsi du contexte de chaque terme est plus précise car elle écarte les termes n'appartenant pas au contexte du terme considéré. Notre approche se base sur une évaluation interactive et incrémentale de la qualité des clusters par l'utilisateur. Nous avons expérimenté nos principes algorithmiques en utilisant l'algorithme des cartes de kohonen sur un corpus en langue française portant sur le domaine du tourisme. Les résultats obtenus montrent bien que la prise en compte du contexte structurel des mots améliore considérablement la pertinence des concepts extraits.
Dans la section suivante, nous exposons le module d'extraction de concepts ontologiques. La section 3 détaille nos expériences et les résultats obtenus. Dans la section 4, nous concluons sur le travail présenté.

## 2 Principes Algorithmiques d'Extraction de Concepts Ontologiques

Notre approche se base sur une architecture composée principalement d'une étape de prétraitement, une étape de traitement et une étape de formalisation et d'évaluation. L'étape de prétraitement utilise différents modules pour la constitution, la représentation, le traitement et l'analyse du corpus [BEN 05]. L'analyse porte à la fois sur la structure et la nature du corpus, ainsi que sur les aspects linguistiques. L'objectif de l'analyse de la nature du corpus est de construire un corpus suffisamment riche et varié pour bien couvrir les concepts du domaine. L'analyse de la structure permet de caractériser le corpus en examinant les balises dominantes et les plus représentatives du domaine, les liens entre elles et les termes associés. Quant à l'analyse linguistique, elle permet de caractériser le corpus d'un point de vue morphologique et syntaxique. Le module de représentation permet non seulement de structurer l'intégralité du corpus sous forme relationnelle mais aussi de l'enrichir avec les informations recueillies des différentes analyses. L'étape de traitement exploite cette représentation relationnelle et en particulier les caractéristiques structurelles du corpus afin de définir un contexte approprié à chaque terme selon sa position dans le corpus.
Cette section décrit la manière dont les concepts ontologiques d'un domaine sont extraits en fonction de leur contexte. Ce contexte est un ensemble de circonstances (situations) qui entourent l'objet d'étude et reflète son environnement concret. Il fournit un support pour l'activité d'apprentissage et pour l'interprétation sémantique. Dans notre cas, l'objet étudié est le mot, l'activité d'apprentissage est le clustering, l'interprétation sémantique est l'opération d'évaluation et de labellisation des classes de termes et la définition du contexte est déduite des analyses structurelles et linguistiques du corpus. Notre définition de contexte est représentée par une hiérarchie par rapport à laquelle le contexte est instancié en respectant les différentes positions du terme dans une balise html (contexte structurel). Le contexte structurel est basé sur l'existence ou non de relations entre les balises html.
L'existence d'une relation structurelle entre les éléments HTML peut révéler une relation sémantique implicite entre les termes associés. Le fait d'instancier le contexte par rapport au lien structurel permet de cerner et révéler les concepts relatifs aux termes apparaissant dans par exemple les balises <h1> → <p> ; <caption> → <td> (titre d'un tableau → cellule d'un tableau) ; <TITLE_URL> (titre d'un lien hypertexte) → les titres d'une partie d'un document ; <TITLE_URL> → les titres du document référencés ; etc. Nous distinguons deux types de lien structurel : un lien physique qui dépend de la structure du document HTML (entre la balise <h1> et la balise <p> associée) et un lien logique qui n'est pas visuel puisque les éléments ne sont pas nécessairement consécutifs (entre <TITLE_URL> et les titres du document référencés). Pour caractériser ces liens entre les balises, nous avons défini une hiérarchie contextuelle (H.C.) pour déterminer les termes reliés dans le corpus par liaison structurelle.

Lorsque deux termes se retrouvent dans la même unité (paragraphe ou document), nous parlons de cooccurrence de ces deux mots dans cette unité contextuelle. En respectant cette structure, nous établissons des liaisons entre les termes si :

- Les termes sont encadrés par la même balise bloc (TAB. 1 : Exemple 1). Dans ce cas, on parle de cooccurrence par voisinage et le contexte est fixé à la balise en soi (<H1>).

- Les termes sont encadrés par des balises qui à leur tour sont reliées par un lien physique ou logique défini dans la hiérarchie contextuelle. Dans ce deuxième cas (TAB. 1 : Exemple 2), nous parlons de « cooccurrence par liaison » et le contexte est l'association des deux balises (<title> + <h1>).

| Exemple 1 | Exemple 2 |
|---|---|
| <H1> événement maritime </H1> | <TITLE> Catégories de logements et d'établissements d'hébergement </TITLE> <KEYWORDS> *** </KEYWORDS> <HYPERLINK> *** <TITLE_URL> *** <H1> Résidences de tourisme </H1> <P> un établissement touristique ayant certaines caractéristiques communes avec un hôtel…… </P> |

Tab. 1 – Exemples de contextes d'utilisation

La cooccurrence par liaison est une cooccurrence pour laquelle le contexte n'est pas fixé à une unité figée mais plutôt générique et instancié selon l'appartenance du terme à une balise. Dans l'exemple 2 (TAB. 1), si nous considérons le terme « logement », en respectant la liaison logique existante entre <TITLE> et <H1> (figurant dans H.C), nous trouvons les co-occurrents de « logement » dans la réunion des deux balises bien qu'elles soient éloignées. Ce second type de liaison (logique) est sémantique puisqu'un titre de document aura une relation avec les sous titres du même document. Si nous considérons le terme « résidences », nous retrouvons ses co-occurrents dans l'association des deux balises <h1> et <p> qui sont reliées par un lien physique conformément à H.C. Ces deux balises représentent le contexte instancié pour le terme « résidences» en respectant son appartenance à <h1>. Si ce même terme existe dans une autre balise, le contexte sera différent et sera une nouvelle instance. Dans les cas ou nous ne retrouvons ni un lien logique ni un lien physique entre deux balises, nous considérons la balise seule en tant qu'unité contextuelle et nous appliquons la cooccurrence par voisinage dans la même balise html. Dans l'exemple 1 (TAB. 1), nous avons comme co-occurrent de « événement » le terme « maritime » dans la balise <h1>. Cette balise représente le contexte du terme « événement ». L'application du contexte générique en relation avec la structure html et les liens sémantiques existants entre les balises permet de représenter l'adaptabilité d'un terme dans le corpus. Notre modèle contextuel, en tenant compte de la position d'un terme, prends en considération diverses situations dans lesquelles le terme a été cité. Le calcul de pondération d'un terme par rapport à son co-occurrent dépend des différents contextes (instanciés grâce à H.C) dans lesquels le mot apparaît. La pondération d'un terme est calculée en utilisant l'indice d'équivalence [MIC 88] qui permet d'évaluer la force de lien entre deux termes.

## 3 Expérimentations et Evaluation des Résultats

Afin d'évaluer notre modèle contextuel, nous avons appliqué deux définitions de contextes sur le même corpus. Le premier contexte est un contexte statique permettant d'encadrer un mot dans une fenêtre d'une taille précise. Nous cherchons les co-occurrents d'un mot dans un espace de 10 mots. Cette définition de contexte considère que tous les mots possèdent la même importance sans tenir compte du fait qu'ils appartiennent à certaines balises html. Le second contexte se base sur notre hiérarchie contextuelle et nous initions le processus avec les termes appartenant aux balises clefs et aux titres. Les co-occurrents des termes sont sélectionnés par rapport à la cooccurrence par voisinage, la cooccurrence par liaison et la hiérarchie contextuelle. Nous utilisons une méthode de clustering non supervisée à savoir les cartes de Kohonen» [KOH 01] avec la distance euclidienne en tant que distance de similarité. En appliquant nos principes algorithmiques pour les deux contextes différents, nous expérimentons différentes alternatives de nombre de classes allant de 20 à 400. Les résultats de nos expérimentations (306 classes) sont évalués par deux experts de domaine puisque certains termes de nos classes n'existent pas dans le thésaurus de l'OMT (Organisation Mondiale du Tourisme). Les experts évaluent les classes de termes. Nous analysons leurs évaluations en tenant compte de différents points de vue comme la distribution des termes, la pondération des paires de termes, la similarité entre deux termes, les concepts extraits, l'interprétation sémantique des classes et le degré de généralité des concepts extraits. Concernant la distribution des

termes, avec le premier contexte, nous obtenons 74 classes parmi lesquelles il existe une classe contenant 55% des termes. Alors que pour le second contexte, seuls 13% des termes initiaux sont regroupés dans la même classe.

**Pondération et Similarité de paires de termes.** La pondération est calculée entre deux termes, le terme à classer et son co-occurrent dans le contexte. Notre contexte structurel permet d'obtenir de meilleurs résultats que le contexte basé sur une fenêtre. Par exemple, le terme « hébergement » et son co-occurrents sémantique « établissement » sont retrouvés avec une pondération plus importante avec notre hiérarchie contextuelle. Nous remarquons également que « loisir » et « oie », qui sont des mots n'ayant aucune relation sémantique, possèdent une faible pondération dans notre contexte en comparaison avec le premier contexte (Fig.1). En ce qui concerne la similarité entre termes, nous remarquons que la similarité calculée dans notre contexte est meilleure que celle obtenue avec le premier contexte. Par exemple, nous retrouvons les termes « archipel » et « île » qui forment le groupe nominal « archipel d'îles » avec une faible similarité dans notre contexte par rapport au premier contexte. Alors que dans ce dernier, nous retrouvons des mots n'ayant pas de relations sémantique comme « bicyclette » et « cuisine » et qui ont une similarité plus faible que celle retrouvé avec le second contexte (Fig.2).

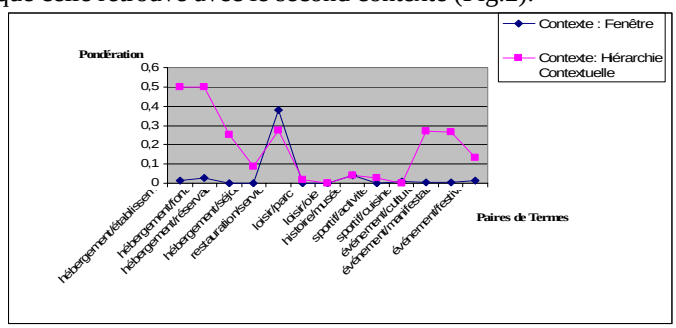



FIG. 1 – *Pondérations de paires de termes avec deux définitions de contextes*

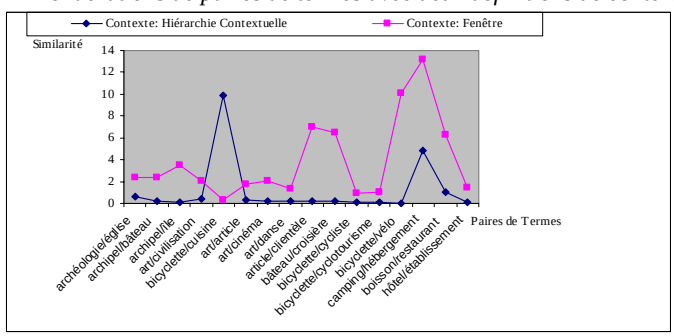



FIG. 2 – *Similarités entre termes avec deux définitions de contextes*

**Interprétation sémantique.** En évaluant les classes, les experts du domaine notent trois types de classes qui sont les classes acceptables, les classes incorrectes et les classes inconnues. Une classe acceptable est une classe que l'expert est capable de labelliser. Une classe incorrecte est une classe qui soit contient des termes qui n'ont pas de relations avec le concept extrait de cette classe, soit elle contient plusieurs concepts clairement identifiés par l'expert. Une classe inconnue est une classe dont les termes n'ont aucune relation sémantique ; l'expert ne peut pas en donner une interprétation sémantique. Dans notre expérimentation, nous obtenons avec notre définition de contexte plus de classes acceptables (53.2%) et moins de classes inconnues (20.51%) et légèrement plus de classes incorrectes (26.28%) en comparaison avec le premier contexte pour lequel nous obtenons respectivement (40.54%), (33.78%) et (25.67%).

**Concepts extraits.** En tenant compte uniquement des classes acceptables, nous calculons la précision. Dans notre étude, « la précision est le ratio des termes pertinents ayant entre eux une importante similarité sémantique par rapport à l'ensemble des termes d'une classe donnée». Comme résultats, nous obtenons respectivement 81.36% et 86.18% pour le premier et le second contexte.

**Degré de généralité des concepts extraits.** Nous établissons une évaluation manuelle en nous basant sur les classes acceptables et sur le thésaurus de l'Organisation Mondiale du Tourisme (OMT). Respectivement pour le premier et le second contexte, nous obtenons 60% et 78.31% de concepts généraux.

## 4 Conclusion et Perspectives

L'acquisition des connaissances est une tâche difficile et lourde au regard de la diversité langagière du Web. Dans cet article, nous nous sommes focalisés sur le processus d'extraction de concepts ontologiques guidée par le contexte. Ce contexte est modélisé par une hiérarchie contextuelle qui représente un contexte structurel. Nous avons également présenté les expérimentations faites sur les premiers niveaux de notre hiérarchie, à savoir sur les mots appartenant aux balises clefs et aux balises titre. Les résultats obtenus ont montré l'importance de la définition du contexte pour améliorer la sélection des co-occurrents sémantiquement proches, la pondération des termes, et par conséquent la pertinence des concepts extraits. Dans les travaux à venir, nous allons poursuivre nos expérimentations concernant les autres niveaux de la hiérarchie contextuelle et nous allons définir un contexte linguistique puis le combiner avec le contexte structurel afin d'améliorer la finesse de décomposition des clusters et de construire une hiérarchie de clusters. Convaincus que l'évaluation et la labellisation sont deux tâches indissociables, nous intégrons cette tâche et de façon incrémentale dans notre processus d'extraction. Nous envisageons de poursuivre notre réflexion par rapport à la découverte des relations et des instances.

## 5 Bibliographie


[BEN 05] N. Bennacer, L. Karoui : "A framework for retrieving conceptual knowledge from Web pages" Semantic Web Applications and Perspectives SWAP, Italy , 2005.

[CHA 00] Chalendar, G and B. Grau. SVETLAN A system to classify nouns in context. Proceedings of the ECAI 2000 Workshop on ontlogy learning, 2000.

[FAU 98] Faure, D., C. Nedellec and C. Rouveirol. Acquisition of semantic knowledge using machine learning methods : the system ASIUM. Technical report number ICS-TR-88-16, Laboratoire de recherche en informatique, inference and learning group, University of Paris-sud, 1998.

[HAH 01] Hahn, U. and M. Romacker. The SYNDIKATE Text Knowledge Base Generator. Proceedings of the 1st International Conference on Human Language Technology Research, San Diego, USA, 2001.

[KOH 01] Kohonen, T. Self organizing Maps. Eds Springer, 2001.

[MEA 01] Meadche, A. and S. Staab. Ontology learning for the semantic Web. IEEE journal on Intelligent Systems, Vol. 16, No. 2, 72-79, 2001.

[MIC 88] Michelet, B. . L'analyse des associations. Thèse de doctorat, Université de Paris VII, UFR de Chimie, Paris, 1988.

[SUG 04] Sugiura, N, N., Izumi and T. Yamaguchi. A supprt environment for domain ontology development with general ontologies and text. IEEE Computational Intelligence Bulletin, February 2004, Vol.3 No.1, 2004.

# Classification contrainte non-supervisée pour le regroupement de modalités

**Aurélie Le Cam**

*France Télécom Recherche et Développement,*
*2 avenue Pierre Marzin*
*22300 LANNION*
*aurelie.lecam@francetelecom.com*

RÉSUMÉ. *L'objectif du travail présenté dans cet article est de limiter le nombre de variables explicatives à tester dans le cadre d'études datamining. On présente une utilisation d'une variante de l'algorithme des K-means : les K-médioïdes [KAU 90]. On applique cet algorithme pour regrouper les modalités des variables explicatives qui nécessitent parfois la prise en compte d'une contrainte de proximité. L'utilisation de cet algorithme permet à l'analyste de tester un ensemble de variables et de croisements de variables attendus par les responsables marketing.*

MOTS-CLÉS : *datamining, modalités, classification, K-medioïdes.*

## 1. Contexte

La division recherche et développement de France Télécom a pour mission, entre autre, d'aider les services marketing à mieux cibler leurs clients dans les actions de vente ou de fidélisation par exemple. Pour cela, les statisticiens cherchent à exploiter le maximum d'informations disponibles au sein des bases de données clients. Au cours de la phase de préparation de données du cycle datamining [CHA 00], on cherche à construire un maximum de variables explicatives pour ensuite sélectionner les plus pertinentes utilisées dans la phase de modélisation. Aujourd'hui, avec les capacités de stockage de données qui augmentent, ce nombre de variables est potentiellement très important et peut vite devenir critique quand on commence à croiser les variables entre elles. C'est le cas dans les études marketing où l'on souhaite par exemple tester la durée d'appels par destination (100 modalités) par tranche horaire (24 modalités) par jour nommé (7 modalités). Ce seul croisement engendre $100 \times 24 \times 7 = 16800$ variables explicatives à tester. Or, les algorithmes d'apprentissage tels que les régressions, les arbres de décisions ou les réseaux de neurones n'autorisent, en entrée, qu'un nombre limité de variables explicatives. L'objectif principal du travail présenté dans cet article est de trouver une méthode statistique qui permet de réduire le nombre de modalités d'une variable descriptive et ainsi permettre de tester un nombre plus raisonnable de variables, tout en explorant l'ensemble de la base de données. On cherche à regrouper les modalités pour lesquelles on observe un comportement semblable. Ce regroupement est réalisé de façon non supervisée, au cours de l'étape de préparation de données, c'est-à-dire quel que soit l'objectif fixé de la modélisation. On présente ici une méthode pour le regroupement ordonné de modalités, c'est-à-dire un regroupement pour des modalités ayant une contrainte de proximité entre elles. La solution choisie est la classification de profils représentant chaque modalité de variables. La méthode de classification est une variante de l'algorithme des k-moyennes [LLO 82, MAC 98].

Dans un premier temps, on présentera la construction des profils que l'on cherche à regrouper, puis on présentera l'algorithme de classification utilisé et on terminera par une application de cet algorithme sur un extrait d'un jeu de données contenant le nombre d'appels par tranche horaire. Le meilleur résultat obtenu est une classification en 4 groupes.

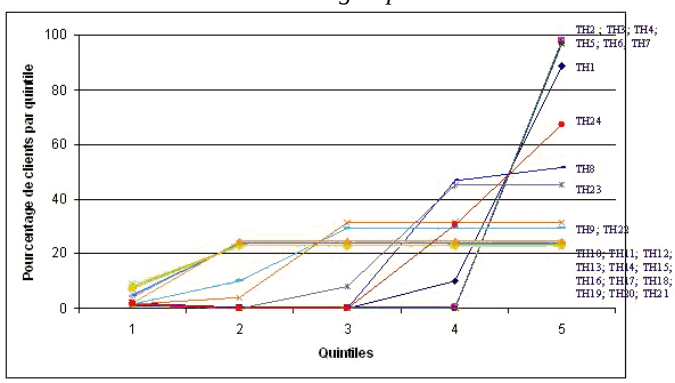
## 2. Eléments à regrouper

### 2.1. *Construction des profils*

Les modalités à regrouper sont représentées par des profils. Ceux-ci sont décrits par 5 points qui sont les quintiles d'une variable caractérisante. Chacun de ces 5 points représente le pourcentage d'individus présents sur le quintile. Dans le cadre des études marketing, la variable caractérisante choisie est le chiffre d'affaire total du client. Cette variable est découpée en quintiles. Un profil est ensuite obtenu en calculant le pourcentage d'individus par quintile possédant la modalité concernée. Les éléments à regrouper sont donc des profils de 5 valeurs.

On donne figure 1 les profils obtenus pour les tranches horaires (24 modalités).

**FIG. 1.** *Profils normalisés des 24 tranches horaires à regrouper*




### 2.2. *Contrainte de proximité*

L'algorithme de regroupement que l'on présente dans cet article s'applique sur des variables pour lesquelles les modalités sont contraintes. C'est le cas par exemple des tranches horaires pour lesquelles il existe une contrainte de proximité entre chaque modalité : la tranche horaire de 12 à 13 heures est proche des tranches horaires de 11 à 12 heures et de 13 à 14 heures. De plus, pour les services marketing, regrouper des éléments contigus constitue une meilleure compréhension et une meilleure utilisation des résultats. On souhaite donc pouvoir intervenir au niveau de l'algorithme de classification pour contraindre certain regroupement. En effet, seules 2 tranches horaires contigües peuvent être regroupées à chaque étape.

## 3. Algorithme utilisé

### 3.1. *Présentation de l'algorithme*

L'algorithme choisi est une variante des k-moyennes : les k-médioïdes [KAU 90]. L'algorithme des k-moyennes permet de faire l'apprentissage non-supervisé de classes. Il existe plusieurs versions de l'algorithme dans la littérature, mais le principe est toujours le même. On doit d'abord assigner une classe aux éléments à regrouper et calculer la moyenne de chaque classe ainsi créée. On assigne ensuite aux éléments la classe dont la moyenne est la plus proche selon une mesure de distance (la distance choisie ici est la distance euclidienne carrée). On continue jusqu'à ce qu'une itération ne provoque aucun changement dans la classification des données. Il existe plusieurs méthodes pour évaluer la solution finale de cet algorithme. Une méthode simple consiste à comparer les erreurs de reconstruction entre les classifications obtenues et à choisir la classification minimisant la somme des distances intra-classes. Pour cela, on classe les objets et on somme leurs distances (euclidiennes) avec la moyenne

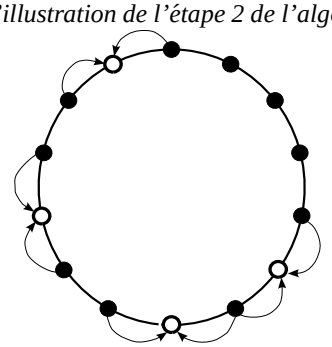
de leur classe respective. Plus cette somme est faible plus les classes sont homogènes et meilleure est la classification. L'algorithme des k-médioïdes a le même principe. La différence est que l'on attribue un élément à une classe en prenant le minimum de distance entre cet élément et un élément de la classe et non la moyenne des éléments de cette classe. Pour le regroupement de modalités des variables à tester dans le cadre de notre étude, les résultats les plus stables ont été obtenus avec l'algorithme des k-médioïdes, en comparaison avec l'algorithme des k-moyennes et la classification ascendante hiérarchique [BRE 84]. Pour prendre en compte la contrainte de proximité, on interviendra directement sur l'algorithme de regroupement.

### 3.2. *Détails de l'algorithme contraint*

$1^{\text{ère}}$ étape : on choisit K objets aléatoirement parmi les n objets à classer.

$2^{\text{ème}}$ étape : on cherche la plus petite distance entre chaque médioïde et ses 2 éléments contigus (distance euclidienne). On agrège l'élément au médioïde. Voir illustration (figure 2).

**FIG. 2.** *Schéma d'illustration de l'étape 2 de l'algorithme*



14 objets à regrouper. 4 médioïdes (en blanc). Les distances à calculer et à comparer sont repérées par les arcs de cercle

Ainsi de suite jusqu'à ce que tous les éléments soient agrégés à un médioïde.

$3^{\text{ème}}$ étape : on choisit les nouveaux médioïdes dans les classes formées après l'étape 2. Le nouveau médioïde dans la classe est l'élément le plus proche (en distance) du centre de gravité de la classe.

$4^{\text{ème}}$ étape : on réitère l'étape 2, jusqu'à stabilité des classes.

### 3.3. *Choix du meilleur regroupement*

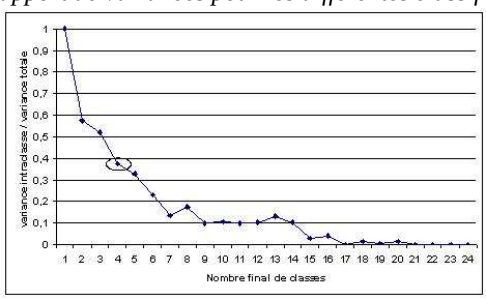
Un tirage aléatoire étant effectué à la première étape de l'algorithme et le choix du nombre de classes se faisant a priori, on réalise plusieurs regroupements à partir de tirages différents. Le nombre de classes finales est testé parmi tous les résultats possibles : entre 1 et le nombre initial de modalités de la variable (24 dans le cas des tranches horaires). Le critère pour le choix du meilleur regroupement est un rapport de variances : variance intraclasse sur variance totale [SAP 90] : $\frac{V_{\text{intra}}}{V_{\text{totale}}}$. La variance intraclasse pour une classe donnée est la somme des distances euclidiennes des éléments de la classe au centre de gravité de la classe. La variance intraclasse totale est la somme de toutes les variances intraclasse : $V_{\text{intra}} = \Sigma_j V_j$ où $V_j = \Sigma_{e_i \in E_i}(e_i - g_j)^2$ avec $e_i$ un individu du nuage et $g_j$ le centre de gravité du groupe $E_j$. La variance totale est la somme des distances euclidiennes de tous les éléments à classer au centre de gravité de tous les éléments : $V_{\text{totale}} = \Sigma_{e_i \in H}(e_i - g)^2$ avec $e_i$ un individu du nuage et $g$ le centre de gravité du nuage. Un premier critère pour choisir ensuite le meilleur regroupement est celui pour lequel le rapport $\frac{V_{\text{intra}}}{V_{\text{totale}}}$ est juste inférieur à 0,5 : c'est le niveau à partir duquel la variance intraclasse devient supérieure à la variance interclasse, ce qui permet d'obtenir un résultat de classification satisfaisant. Un second critère peut

ensuite être le nombre maximal de variables explicatives que l'on peut tester et dépend donc des données dont on dispose. Ce nombre de variables explicatives peut nous permettre de choisir un regroupement donnant un rapport de variances bien en dessous de 0,5 correspondant à une meilleure classification.

## 4. Application sur un jeu d'essai

Le cadre de notre travail est une étude marketing, nous disposons donc de données liées au trafic téléphonique. Dans cette partie, nous allons donner les résultats de la classification des tranches horaires par l'algorithme des K-médioïdes contraint. Les profils à regrouper sont ceux de la figure 1. On réalise la classification en 1-médioïde jusqu'en 24-médioïdes (nombre de modalités des tranches horaires). D'après la figure 3 et le critère du rapport des variances, le meilleur regroupement est le résultat de la classification en 4-médioïdes.

**FIG. 3.** *Evolution du critère du rapport de variances pour les différentes classifications*




Le regroupement en 4 classes minimisant le critère du rapport de variances et satisfaisant la contrainte de proximité des modalités donne les 4 classes suivantes :

- Classe 1 : tranches horaires 1 à 3 (de minuit à 3 heures du matin).
- Classe 2 : tranches horaires 4 à 10 (de 3 heures à 10 heures).
- Classe 3 : tranches horaires 11 à 20 (de 10 heures à 20 heures).
- Classe 4 : tranches horaires 20 à 24 (de 20 heures à minuit).

Ce résultat permet de réduire le nombre de variables obtenues après croisement avec les tranches horaires. En effet, on passe de 24 à 4 modalités. De plus, cet algorithme tient compte de la contrainte et offre au service marketing des regroupement de tranches horaires ayant un sens : on obtient bien des plages horaires (intervalles de tranches horaires) et non pas des regroupements de tranches horaires non contigües, ce qui serait inexploitable du côté marketing.

L'application d'un tel algorithme sur l'ensemble des modalités de variables, contraintes ou non, d'une base de données autoriserait le test d'un maximum de croisements de variables. Cette phase de regroupement de modalités associée ensuite à une sélection de variables permet d'obtenir les indicateurs les plus pertinents en vue d'un ciblage ou d'une fidélisation plus efficace.

## 5. Bibliographie


[BRE 84] BREIMAN L., FRIEDMAN J., OLSHEN R., STONE C., *Classification and regression trees,* Chapman & Hall, 1984.

[CHA 00] CHAPMAN P., CLINTON J., KERBER R., KHABAZA T., REINARTZ T., SHEARER C., WIRTH R., CRISP-DM 1.0 : step-by-step data mining guide, , 2000.

[KAU 90] KAUFMAN L., ROUSSEEUW P., *Finding groups in data : an introduction to cluster analysis,* John Wiley & Sons, New York, 1990.

[LLO 82] LLOYD S., Least squares quantization in PCM, rapport, 1982, Bell Laboratories, IEEE Transactions on Information Theory 28.

[MAC 98] MACQUEEN J., Some methods for classification and analysis of multivariate observations, *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, 1998, p. 281–297.

[SAP 90] SAPORTA G., *Probabilités, Analyse des données et Statistique*, Technip, 1990.

# Filiation de manuscrits sanskrits et arbres phylogénétiques

**M. Le Pouliquen, J.P. Barthélemy, P. Bertrand**

*Département LUSSI*
*ENST Bretagne BP 832*
*29285 Brest Cedex*
*TAMCIC, UMR CNRS 2872*
*marc.lepouliquen@enst-brestagne.fr*

RÉSUMÉ. *La fabrication d'un stemma codicum est l'une des approches les plus rigoureuses de la critique textuelle. Elle exige la reconstruction de l'histoire du texte en classifiant le corpus pour décider si un groupe de manuscrits est engendré par un intermédiaire perdu. Pour classifier notre corpus, nous employons des méthodes de l'analyse textuelle informatisée, de la reconstruction phylogénétique afin d'établir l'arbre de la filiation. Les techniques employées sont dédiées à un corpus de manuscrits sanskrits avec toutes les spécificités de cette langue.*

MOTS-CLÉS : *Critique textuelle, arbre phylogénétique, sanskrit, distances intertextuelles, stemma codicum.*

## 1. Introduction

Dans le cadre de l'édition critique de manuscrits anciens, un des problèmes consiste à trier les différentes versions du texte (appelés *témoins*) afin d'essayer de reconstituer le manuscrit original avec le plus de fidélité. L'analyse de ces différents témoins pour réaliser l'édition critique est un travail colossal et se fait en plusieurs étapes dont l'une consiste à établir un arbre de filiation de ces manuscrits pour savoir lequel à été copié sur l'autre ; c'est l'établissement du *stemma codicum*.

Le projet consiste à utiliser les méthodes de la phylogénétique de l'alignement de corpus et des distances intertextuelles afin de proposer un arbre qui permette d'établir un premier classement automatique de manuscrits sanskrits.

Ces manuscrits en sanskrit, relatifs à la « glose de Bénarès »(*Kasikavritti*), un texte fondamental de la tradition grammaticale indienne, rajoutent d'autres difficultés du fait des particularités propres de langue sanskrite.

## 2. Méthodes philologiques d'établissement du stemma codicum

La plupart des méthodes philologiques utilisées pour la classification des témoins part du constat suivant, à savoir que toutes les copies qui contiennent, aux mêmes endroits, les mêmes fautes, ont été faites les unes sur les autres et donc dérivent toutes d'une copie où ces fautes existaient. Pour classer les témoins, on recourt donc à la méthode de la comparaison des fautes appelées variantes et classées selon leur influence sur l'acte de copie qui permettent de dresser sans trop de peine un arbre généalogique des manuscrits traités. Cette méthode inspirée par Don Quentin [QUE 26] présente l'avantage de préparer le travail de l'édition critique car pour reconstituer le texte le plus proche de l'original, on évalue laquelle des variantes convient le mieux.

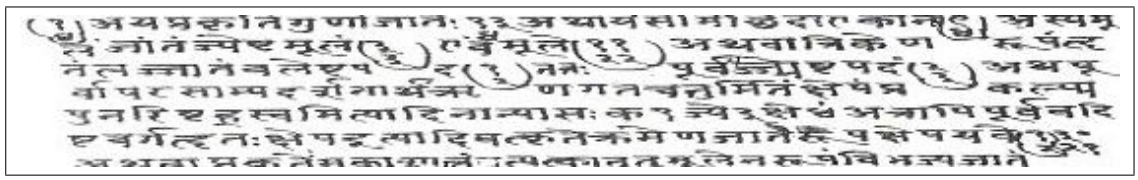


**FIG. 1.** *Exemple de manuscrits sanskrits*

Pour mieux cerner la difficulté du problème, imaginons une oeuvre dont on possède 150 exemplaires non identiques ; que les variantes indépendantes de tel texte se comptent par milliers ; Combien d'années de travail seraient nécessaires à un homme pour la réalisation d'une édition critique ?

L'importance des méthodes informatiques pour aider l'éditeur critique s'impose alors...

## 3. Les problèmes liés au sanskrit.

Le sanskrit est une des grandes langues de l'Asie pratiquée essentiellement en Inde. Son origine remonte à la plus haute antiquité ; son usage, bien qu'en déclin, s'est poursuivi pendant l'ère chrétienne jusqu'à nos jours par des érudits. On imagine difficilement les évolutions dans le temps et l'espace qu'a pu subir le vocabulaire et par conséquent les manuscrits.

Le sanskrit possède un alphabet de 46 lettres ce qui oblige lors de sa translittération à faire correspondre une lettre sanskrite à une séquence de lettres latines et complique par la même la comparaison des textes et la reconnaissances des mots.

Le sanskrit est une langue qui s'écrit généralement sans espace entre les mots, ce qui complique encore les comparaison au niveau des mots. Seul l'utilisation d'un texte lemmatisé comme lexique peut servir à la reconnaissance des mots. L'opération de lemmatisation d'un texte étant à ce jour trop coûteuse pour être étendue à tous les manuscrits, elle n'est réalisée que sur l'un d'eux appelé *Padapatha*.

Pour compliquer davantage nos comparaisons, le *sandhi* est un phénomène désigne les modifications qui se produisent à la rencontre de deux mots dans une phrase. En français le « à le »qui se transforme en « au »est une sorte de sandhi.

Pour en finir avec les spécificités du sanskrit, les spécialistes dont Filliozat [FIL 41] parle de l'orthographe « vicieuse mais traditionnelle des scribes »et d'autres spécialistes nous ont montré que les mots difficiles à comprendre sont ceux où les manuscrits proposent une autre solution : preuve que loin de recopier sans comprendre, le scribe a remplacé le mot qu'il ne comprenait pas...

## 4. Alignement et distances

Pour établir la distance entre les textes, différentes pratiques sont observées. Il convient tout d'abord de décider de la segmentation qui peut être effectuée à plusieurs niveaux : les caractères, les syllabes, les mots, les lemmes, les phrases, les paragraphes... Par expérience, plus la segmentation est riche en sens (des lemmes plutôt que des mots) plus les informations sont pertinentes. Dans un premier temps, à cause des difficultés liées au sanskrit, nous nous contenterons de 2 niveaux de base, les caractères et les mots. Afin de comparer ces derniers, nous avons recours à des techniques d'alignement permettant de mettre en correspondance, par un traitement automatique, les portions de textes qui sont similaires les unes des autres.

Les textes à comparer sont divisés en chapitres, paragraphes, et en séquence de caractères que j'appelle phrase. Seul le premier chapitre est actuellement translitteré, les comparaisons vont donc s'effectuer facilement, paragraphe par paragraphe car ils sont numérotés. Pour réaliser les comparaisons des phrases, un premier alignement de celle-

ci paraît nécessaire. Dans le cadre de l'alignement multilingue, la méthode de Gale et Church [GAL 91] nous donne des résultats pertinents pour l'alignement de nos phrases, enlevant par la même les commentaires et les phrases pas « comparables ».

La distance de Levenshtein [LEV 66] est couramment utilisée dans de nombreuses applications où il faut mesurer la similarité entre deux séquences ici nos phrases. Elle permet de déterminer quelle est la longueur d'une séquence minimale d'opérations pour transformer la première séquence en la seconde. Elle travaille donc au niveau des caractères et a l'avantage de la simplicité

Au niveau des mots, l'usage d'indices tels que, parmi beaucoup d'autres, celui de Jaccard ou la connexion lexicale de Muller [MUL 77] permet de calculer le rapport entre les mots qui sont communs aux deux textes et ceux qui n'appartiennent qu'à l'un des deux. C'est une méthode similaire qui est utilisée en prenant en compte les spécificités de nos manuscrits, à savoir que le seul lexique permettant la comparaison des mots est celui du padapatha.

Une fois une matrice des distances obtenue, nous utilisons des algorithmes de reconstruction phylogénétique pour inférer un arbre et tenter de proposer une racine, c'est à dire le manuscrit original (existant ou non).

## 5. Arbre phylogénétique

La phylogénie peut être considérée comme une représentation de l'histoire évolutive d'un ensemble d'espèces. On choisit alors de représenter les relations qui existent entre elles sous forme d'un arbre phylogénétique (le plus souvent binaire) comme l'a suggéré Buneman [BUN 71]

Le corpus est composée d'un nombre important de manuscrits (environ 120) et de grande taille. Les méthodes choisies ne doivent pas avoir de complexité trop grande et ne s'appliquerons sans doute pas à la totalité du texte. Pour cela, on utilise de préférence les méthodes basées sur les distances.

ADDTREE [SAT 77] et NJ [SAI 87] prolongées par la méthodes des groupements de Barthélemy et Luong [BAR 88] [LUO 88] sont les méthodes les plus fréquemment utilisées pour l'inférence d'arbres phylogénétiques à partir des dissimilarités et sont utilisés ici pour reconstruire l'arbre de la filiation Le principal problème constaté dans les tentatives de « stemmatisation »est celui de l'orientation de l'arbre, c'est à dire la détermination soit du manuscrit original, soit de la liste des parentés entre les manuscrits. Pour tenter de résoudre ce problème, bien que ADDTREE et NJ proposent une racine par construction, nous allons calculer des « coefficients d'intermédiarité »entre 3 témoins. La difficulté provient essentiellement du fait que l'alignement multiple devient NP-difficile à partir de 3. Nous avons alors eu recourt à l'utilisation d'heuristiques pour l'alignement multiple.

## 6. Résultats obtenus

Les algorithmes sont ensuite testés sur différents corpus afin de vérifier et expérimenter les différentes méthodes mises au point. Pour les premiers tests, plusieurs corpus « fictifs »dont on connaît le stemma ont été réalisés. Chacun de ces corpus permet de mettre en évidence un problème particulier (ex : arbre déséquilibré par perte de manuscrits dans une branche). Les méthodes se comportent en général très bien et l'expérimentation permet de les affiner.

Pour ce qui nous concerne, l'ensemble des textes sanskrits déjà collectés (une cinquantaine) a été expérimenté ; la seule preuve d'une convergence vers un arbre « intéressant »est que les différentes méthodes donnent des résultats similaires sur différents corpus constitué avec les manuscrits. Ces premiers résultats sont confiés à des « sanskritistes »pour une possible validation.

Enfin, des corpus déjà étudiés par des philologues qui ont établi l'arbre de filiation sont en cours de constitution et vont être expérimenté prochainement pour voir si les méthodes étudiées sont correctes.

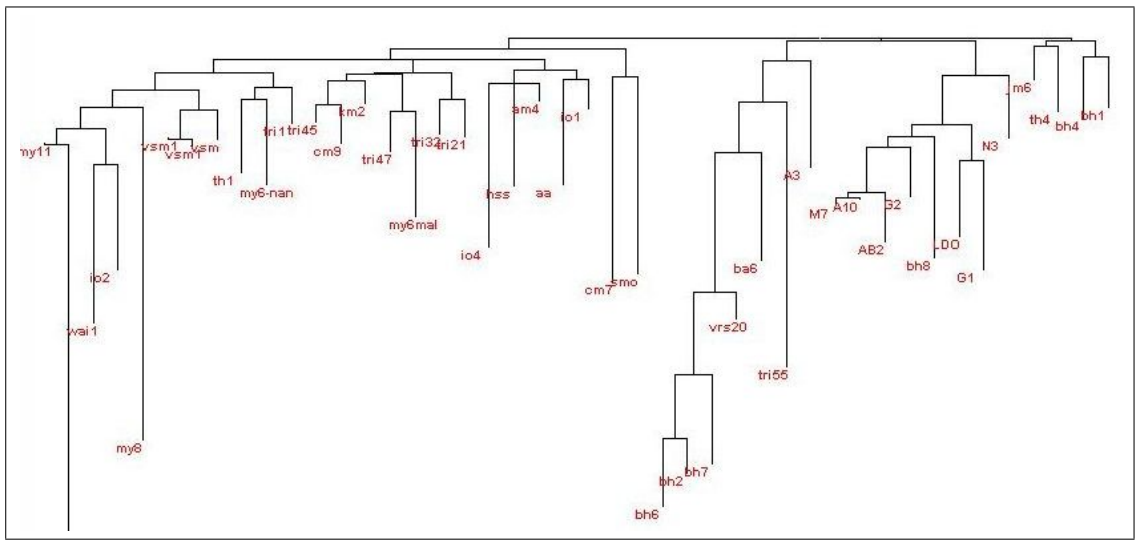



**FIG. 2.** *Exemple d'arbre de la filiation avec les manuscrits sanskrits*

## 7. Conclusion et perspectives

On peut avant tout se demander si, après le nombre important d'opérations effectué sur le corpus, on classe toujours les manuscrits et non pas les copistes ou autres phénomènes ?

Un autre problème est de déterminer au mieux la racine et pour cela il faut sûrement intégrer des informations extérieures comme la paléographie et l'ecdotique comme étudiée à l'Ecole des chartes.

Une étude des règles de l'acte de copie peut permettre d'orienter le graphe et de développer des méthodes d'intermédiarité entre les textes pour reconstruire le stemma. Cette étude peut aussi résoudre le problème de la contamination des textes et choisir leur représentation par un arbre ou un graphe.

Enfin des connaissances supplémentaires sur le sanskrit peuvent être utilisées pour passer d'une analyse au niveau des mots à une analyse au niveau des lemmes sans doute plus riche de sens.

## 8. Bibliographie

[BAR 88] BARTHÉLÉMY J. P., GUÉNOCHE A., *Les Arbres et les Représentations des Proximités*, Masson, 1988.

[BUN 71] BUNEMAN P., Filiations of Manuscripts Mathematics in Archaeological and Historical Sciences, *Edinburgh University Press*, , 1971.

[FIL 41] FILLIOZAT J., Catalogue du fonds sanskrits, paris, adrien maisonneuve édition, 1941.

[GAL 91] GALE W. A., CHURCH K. W., A Program for Aligning Sentences in Bilingual Corpora, *Computational Linguistics*, vol. 19, 1991, p. 75-102.

[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.

[LUO 88] LUONG X., Méthodes d'analyse arborée. Algorithmes. Applications, PhD thesis, Thèse de doctorat, Paris V, 1988.

[MUL 77] MULLER C., *Principes et méthodes de statistique lexicale*, Hachette Paris, 1977.

[QUE 26] QUENTIN H., *Essais de critique textuelle*, Picard, 1926.

[SAI 87] SAITOU N., NEI M., The neighbor-joining method : a new method for reconstructing phylogenetic trees, *Mol Biol Evol*, vol. 4, 1987, p. 406–425.

[SAT 77] SATTAH S., TVERSKY A., Additive similarity trees, *Psychometrika*, vol. 42, 1977, p. 319–345.

# Clustering via DC programming and DCA

**Le Thi Hoai An**[a] **Belghiti M. Tayeb** [b]**, Pham Dinh Tao**[c]

[a]*Laboratory of Theoretical and Applied Computer Science,*
*UFR MIM, University of Paul Verlaine - Metz, Ile du Saulcy, 57045 Metz, France.*
*Laboratory of Modelling,*
[b,c]*Laboratory of Modelling, Optimization and Operations Research,*
*National Institute for Applied Sciences - Rouen*
*BP 08, Place Emile Blondel F 76131 Mont Saint Aignan Cedex, France.*
[a]*lethi@sciences.univ-metz.fr,* [b]*belghiti-moulay.tayeb@insa-rouen.fr,* [c]*pham@insa-rouen.fr*

RÉSUMÉ. *The problem of assigning $m$ point in the $n$-dimensional real space $\mathbb{R}^n$ to $k$ clusters is formulated as that of determining $k$ centers in $\mathbb{R}^n$ such that the sum of distances of each point to the nearest center is minimized. If If the $1$-norm is considered for the distance, then the problem can be formulated in the form of minimizing a piecewise-linear concave function on a polyhedral set. A so-called DCA method based on a DC (Difference of Convex functions) programming approach has been developed for solving this problem. Preliminary numerical solutions on real-world databases show the efficiency and the superiority of the appropriate DCA with respect to the standard K-means algorithm.*

MOTS-CLÉS : *clustering problem, K-median problem, K-median algorithm, K-means algorithm, DC programming, DCA, nonsmooth nonconvex programming, global optimization.*

## 1 Introduction

Clustering is a fundamental problem in unsupervised learning which has many applications in various domains. In recent years, there has been significant interest in developing clustering algorithms to the massive data sets ([1] - [14], [20] - [22], [26], [28], [31] - [34] and reference therein). Two main approaches have been studied for clustering : the first one is the statistical and machine learning based on learning mixture models (see e.g. [1], [2], [22], [26]) and the second is the mathematical programming approach that considers clustering as an optimization problem (see e.g. [3], [4], [20], [28], [32] - [34]). The general term "clustering" covers many different types of problems. All consist of subdividing a data set into groups of similar elements, but there are many measures of similarity, many ways of measuring, and various concepts of subdivision.

An instance of the partitional clustering problem consists of a data set $\mathcal{A} := \{a^1, ...a^m\}$ of $m$ points in $\mathbb{R}^n$, a measured distance, and an integer $k$ ; we are to choose $k$ members $x^\ell$ $(l = 1, ...k)$ (in $\mathcal{A}$ and/or $\mathbb{R}^n$) as **"**centroid**"** (or "median") and assign each member of $\mathcal{A}$ to its closest centroid. The assignment distance of a point $a \in \mathcal{A}$ is the distance from $a$ to the centroid to which it is assigned, and the objective function, which is to be minimized, is the sum of assignment distances. If the centroids are not necessarily in $\mathcal{A}$, then the problem can be formulated as a unconstrained optimization problem. In the contrary case, we are faced with a discrete optimization problem. In both cases, different objective functions corresponding to the distance metric being considered are possible. Two models widely studied

in the literature are the cases where the points come from a real space $\mathbb{R}^n$, and the assignment distance of a point is defined as the squared Euclidean distance (2-norm), and/or the 1-norm. If the squared Euclidean distance is used and the centroids are not necessarily in $\mathcal{A}$, then the corresponding optimization problem can be expressed as ($\|.\|$ denotes the Euclidean norm)

$$\min\left\{\sum_{i=1}^{m}\min_{\ell=1,\ldots,k}\left\|x^\ell-a^i\right\|^2 : x^\ell\in\mathbb{R}^n, \ell=1,\ldots,k\right\}. \tag{1}$$

If the 1-norm is considered instead of the squared Euclidean distance, then the problem can be written as

$$\min\left\{\sum_{i=1}^{m}\min_{\ell=1,\ldots,k} e^T D^{i\ell} : -D^{i\ell}\le x^\ell-a^i\le D^{i\ell}, D^{i\ell}, x^\ell\in\mathbb{R}^n\right\}, \tag{2}$$

where $D^{i\ell}\in\mathbb{R}^n$ is a dummy variable that bounds the components of the difference $x^\ell-a^i$ and $e\in\mathbb{R}^n$ denotes the vector of ones. Both (1) and (2) are nonsmooth nonconvex programs for which there are rarely efficient solution algorithms, especially in the large scale setting.

In this work we develope a new algorithm based on DC programming and DCA for solving the clustering problem when the 1-norm is used, namely Problem (2).

A general DC program is of the form

$$\alpha := \inf\{f(x) := g(x) - h(x) \ : \ x\in\mathbb{R}^n\}, \tag{3}$$

with $g, h$ being lower semicontinuous proper convex functions on $\mathbb{R}^n$. It should be noted that in (3) the convex constraint set $C$ is incorporated in the convex DC component $g$ with the help of its indicator function $\chi_C$ ($\chi_C(x) := 0$ if $x\in C, +\infty$ otherwise). The dual of (3) is the DC program

$$\alpha := \inf\{h^*(y) - g^*(y) \ : \ y\in\mathbb{R}^n\},$$

where $g^*$ is the conjugate function of $g$ :

$$g^*(y) := \sup\{\langle x, y\rangle - g(x) \ : \ x\in\mathbb{R}^n\}.$$

The DCA (see [15] - [19] and references therein) is an efficient method which has been successfully applied to a lot of various large-scale nonconvex programs. It is a descent method without linesearch, consisting of the construction of the two sequences $\{x^k\}$ and $\{y^k\}$, (candidates for being primal and dual solutions, respectively), such that their corresponding limit points $x^\infty$ and $y^\infty$ satisfy local optimality conditions.
Recall that the optimality conditions for DC programming are given by

(i) If $x^*$ is a local minimizer for (3) then

$$\emptyset\ne\partial h(x^*)\subset\partial g(x^*). \tag{4}$$

where $\partial h(x^*) := \{y^*\in\mathbb{R}^n : h(x)\ge h(x^*)+\langle x-x^*, y^*\rangle\ \forall x\in\mathbb{R}^n\}$ is the subdifferenrial of $h$ at $x^*$. Recall that $\partial h(x^*)$ is the extension of the derivative to nondifferentibale convex function and each element $y^*\in\partial h(x^*)$ is a subgradient of $h$ at $x^*$.

Its converse is true for some class of DC programs, in particular for polyhedral ones in which the second DC component $h$ is a polyhedral convex function ([15], [16], [19])

(ii) $x^*$ is called a critical point of $g-h$ or for (3) if

$$\emptyset\ne\partial g(x^*)\cap\partial h(x^*) \tag{5}$$

There are two forms of DCA : the simplified DCA (or simply DCA) and the complete DCA ([16], [17],[19]). In practice the first is preferred to the latter because it is less expensive.

**DCA (DC Algorithm) ([15] - [19] and references therein**

1. Let $x^1 \in \mathbb{R}^n$. Set $k = 1$ and let $\epsilon_1$, $\epsilon_2$ be sufficiently small positive numbers.

2. Compute $y^k \in \partial h(x^k)$.

3. Compute $x^{k+1} \in \partial g^*(y^k)$, i.e., $x^{k+1}$ is a solution of the convex program

$$\min\{g(x) - \langle x, y^k\rangle \ : \ x \in \mathbb{R}^n\}.$$

4. If either $\left\|x^{k+1} - x^k\right\| \leq \epsilon_1 \ (\left\|x^k\right\| + 1)$ or $\left|f(x^{k+1}) - f(x^k)\right| \leq \epsilon_2(\left|f(x^k)\right| + 1)$, then stop and $x^k$ is the computed solution, otherwise, set $k = k + 1$ and go to Step 2.

According to the theory of DC programming, it is easy to show that the clustering problem (2) is a DC program. We can then use DCA for solving it.

## 2 Solving the clustering problem (2) by DCA

First we formulate Problem (2) as a DC program. Fir this we write (2) in the form

$$\begin{cases} \min \quad \sum_{i=1}^m \min\limits_{\ell=1,\ldots,k} \| \ y_{i\ell} \ \|_1 \\ s.t \quad x_\ell - y_{i\ell} \leq a^i \quad i = 1,\ldots,m. \ \ell = 1,\ldots,k \\ \quad -x_\ell - y_{i\ell} \leq -a^i \quad i = 1,\ldots,m. \ \ell = 1,\ldots,k. \end{cases}$$

Let $Z_{i\ell} = (x_\ell, y_{i\ell}) \in \mathbb{R}^{2n}$, $C = (0_n, e_n) \in \mathbb{R}^{2n}$, where $e_n = (1,\ldots,1) \in \mathbb{R}^n$, $0_n = (0,\ldots,0) \in \mathbb{R}^n$ and $i = 1,\ldots,m$ ; $\ell = 1,\ldots,k$.

Let $A = \begin{pmatrix} e_n & -e_n \\ -e_n & -e_n \end{pmatrix}$ and $b_i = \begin{pmatrix} a^i \\ -a^i \end{pmatrix}$.

The problem (2) can be expressed as :

$$\begin{cases} \min \quad \sum_{i=1}^m \min\limits_{\ell=1,\ldots,k} \| \ Z_{i\ell}C \ \|_1 \\ s.t \quad AZ_{i\ell} \leq b^i, \quad i = 1,\ldots,m. \ \ell = 1,\ldots,k \\ \quad Z_{i\ell} \in \mathbb{R}^{2n}, \quad i = 1,\ldots,m. \ \ell = 1,\ldots,k \end{cases}$$

Let $K := \{Z \in \mathbb{R}^{2n\cdot m\cdot k} |\quad AZ_{i\ell} \leq b^i \quad \text{and } i = 1,\ldots,m. \ \ell = 1,\ldots,k\}$. We can advantageously express (2) in the matrix space $\mathbb{R}^{2n\cdot m\times k}$ as follows :

$$(2) \Leftrightarrow \min \quad \{F(Z) := G(Z) - H(Z) : \ Z \in \mathbb{R}^{2n\cdot m\times k}\}, \tag{6}$$

where the DC components $G$ and $H$ are given by

$$G(Z) = \chi_K(Z) \tag{7}$$

$$\text{and} \quad H(Z) = -\sum_{i=1}^m \min_{\ell=1,\ldots,k} \| \ Z_{i\ell}C \ \|_1 \tag{8}$$

where $\chi_K$ denotes the indicator function on $K$, i.e., $\chi_K(Z) = 1$ if $Z \in K$, 0, otherwise. In the matrix space $\mathbb{R}^{2n\cdot m\times k}$, the DC program (6) then is minimizing the difference of the simplest convex function and the nonsmooth convex one (8). This nice feature is very convenient for applying DCA.

According to the previous section, determining the DCA scheme applied to (6) amounts to computing the two sequences $\{Z^{(p)}\}$ and $\{T^{(p)}\}$ in $\mathbb{R}^{2n\cdot m\times k}$ such that

$$T^{(p)} \in \partial H(Z^{(p)}), Z^{(p+1)} \in \partial G^*(T^{(p)}).$$

$$Z^{(p+1)} \in \partial G^*(T^{(p)}) := \arg\min\{G(Z) - \langle Z, T^p\rangle : Z \in \mathbb{R}^{2n_cdotm\times k}\}. \tag{9}$$

We shall present below the computation of $\partial H(Z)$.

## 2.1 Calculation of $\partial H(Z)$

The probem (8) can be write as :

$$H(Z) = -\sum_{i=1}^{m} \min_{\ell=1,\ldots,k} \parallel Z_{i\ell}C \parallel_1 \ = \sum_{i=1}^{m} \max_{\ell=1,\ldots,k} - \parallel Z_{i\ell}C \parallel_1$$

$$\text{for } i = 1,\ldots,m\ H_i(Z) = \max_{\ell=1,\ldots,k} H_{i\ell}(Z) \text{ where } H_{i\ell}(Z) := - \parallel Z_{i\ell}C \parallel_1 .$$

Let $I_i(Z) := \{\ell = 1,\ldots,k : H_{i\ell}(Z) = H_i(Z)\}$. Then we have ([12]) :

$$\partial H_i(Z) = co\{\cup_{\ell\in I_i(Z)}\partial H_{i\ell}(Z)\}, \tag{10}$$

where $co$ stands for the convex hull. Let $T_{i\ell} \in \partial H_{i\ell}(Z)$ then :

$$T_{i\ell} := \frac{-Z_{i\ell}C}{\parallel Z_{i\ell}C \parallel_1} \qquad \text{if} \quad Z_{i\ell} \neq 0 \quad ,0 \quad \text{otherwise.} \tag{11}$$

We can choose the particular subgradient of $H_i$.

$$T_i \in \partial H_i(Z) \Leftrightarrow T_i = T_{i\ell} \qquad \text{with} \qquad \ell \in I_i(Z) \tag{12}$$

The calculation of $\partial H(X)$ is then immediate from the relations (10), (12) and

$$\partial H(Z) = \sum_{i=1}^{m} \partial H_i(Z). \tag{13}$$

Hence, according to (10) and (13) we get the following simpler matrix formula for computing $\partial H$ :

$$T \in \partial H(Z) \Leftrightarrow T = \sum_{i=1}^{m} T_i \text{ with } T_i \in \partial H_i(Z) \text{ for } i = 1,...,k, \tag{14}$$

and the corresponding $T \in \partial H(Z)$ is defined by

$$T = \sum_{i=1}^{m} T_{i\ell} \qquad \text{with } \ell \in I_i(Z). \tag{15}$$

# 3 Description of DCA to solve the problem (2)

We are now in a position to describe the DCA for solving problem(2) via the DC decomposition (6).

### 3.1 Algorithm

Let $\epsilon > 0$ be small enough and $Z^0$ be given. Set $p \leftarrow 0$ ; $er \leftarrow 1$.
**while** $er > \epsilon$ **do**
Compute $T^p \in \partial H(Z^p)$. via (15).
Solve the linear program : $\min\{-\langle Z, T^p\rangle | Z \in K\}$ to obtain $Z^{p+1}$.
$er \leftarrow \| Z^{p+1} - Z^p \|$
$p \leftarrow p + 1$.
**endwhile**.

## 4 Numerical experiments

We have coded the algorithm in C++, and run on a Pentium $2.930GHz$ of 1024 DDRAM. To solve the linear Programming, we used the software CPLEX. Our code are tested on the real data sets given in [4], [7], [32], [33], [34].

In Table 1 we present the comparative numerical results provided by our algorithm and K-means which is available on the web site : http ://www.fas.umonteral.ca/biol/legendre/.

| Problems | DCA | K-means |
|---|---|---|
| ADN | 94.52 | 81.19 |
| IRIS | 99.33 | 97.82 |
| GENE | 91.55 | 80.00 |
| GLASS | 82.78 | 72.62 |
| LENSES | 91.01 | 68.22 |
| LYMPHO | 89.41 | 54.88 |
| PAPILLON | 100.00 | 82.77 |
| PIMAR | 91.04 | 85.66 |
| TITANIC | 97.89 | 90.25 |
| VOTE | 98.34 | 86.05 |
| WINE | 91.61 | 84.87 |
| Average | 93.40 | 80.39 |

**TAB. 1.** *Clustering accuracy (%) achived by each algorithm*

**Conclusion**

We have proposed a new approach for assigning points to clusters based on DC programming and DCA, for solving clustering problems. They have been formulated clustering problem using norm 1 as DC programs in the suitable matrix space in order that DCA be easily computed. It turns out to be a quite simple algorithm which requires only matrix-vector products and solving a linear programming using CPLEX. Preliminary numerical simulations on ten real- world database show the robustness, the efficiency and the superiority of DCA with respect to the k-Mean Algorithm in both running time and quality of solutions.

## References


[1] Alon, N., & Spencer, J. H. (1991). *The probabilistic method.* New York, NY : Wiley.

[2] Arora, S., & Kannan, R. (2001). Learning mixtures of arbitrary Gaussians. In *Proceedings of the 33rd Annual ACM Symposium on Theory of Computing*, pp. 247-257.

[3] K. Al-Sultan (1995), A Tabu search approach to the clustering problem, *Pattern Recognition*, 28(9), 1443-1453.

[4] P.S. Bradley, O.L. Mangasarian and W.N. Street (1997), *Clustering via concave minimization*, Technical Report 96-03, Computer Sciences Department, University of Wisconsin, Madison, Wisconsin May 1996. Advances In Neural Information processing Systems 9, MIT Press, Cambridge, MA, 368-374, M.C. Mozer, M.I. Jordan and T. Petsche, editors. Available by ftp ://ftp.cs.wisc.edu/math-prog/tech-trports/96-03.ps.Z

[5] Charikar, M., & Guha, S. (1999). Improved combinatorial algorithms for facility location and k-median problems. In *Proceedings of the 40th Annual IEEE Symposium on Foundations of Computer Science*, pp. 378-388.

[6] Charikar, M., Guha, S., Tardos, ´E., & Shmoys, D. B. (1999). A constant-factor approximation algorithm for the k-median problem. In *Proceedings of the 31st Annual ACM Symposium on Theory of Computing*, pp. 1-10.

[7] Dhilon, I,S. and Korgan, J. and C. Nicholas, *Feature Selection and Document Clustering*, In M.W. Berry, editor, A Comprehensive Survey of Text Mining, pages 73-100, Springer-Verlag,2003.

[8] Duda, R. O., & Hart, P. E. (1972). *Pattern classification and scene analysis*. Wiley.

[9] Feder, T., & Greene, D. (1988). Optimal algorithms for approximate clustering. In Proc. STOC.

[10] D. Fisher (1987), Knowledge acquisition via incremental conceptual clustering, *Machine Learning*, 2, 139-172.

[11] K. Fukunaga, *Statistical Pattern Recognition.* Academic Press, NY, 1990.

[12] Convex Analysis and Minimization Algorithms, Springer Verlag berlin Heidelberg, 1993

[13] A.K Jain and R.C. Dubes (1988), *Algorithms for clustering Data*, Prentice-Hall Inc, Englewood Cliffs, NJ.

[14] Hartigan, J. A. (1975). *Clustering algorithms*. John Wiley and Sons.

[15] Le Thi Hoai An and Pham Dinh Tao, Solving a class of linearly constrained indefinite quadratic problems by DC algorithms, *Journal of Global Optimization*, Vol 11, No 3, pp 253-285, 1997.

[16] Pham Dinh Tao and Le Thi Hoai An, *Convex analysis approach to DC programming : Theory, Algorithms and Applications*, Acta Mathematica Vietnamica, dedicated to Professor Hoang Tuy on the occasion of his 70th birthday, Vol.22, Number 1 (1997), pp. 289-355.

[17] Pham Dinh Tao and Le Thi Hoai An, DC optimization algorithms for solving the trust region subproblem. *SIAM J. Optimization, Vol. 8, Number 2 (1998), pp. 476-505.*

[18] Le Thi Hoai An and Pham Dinh Tao, Large Scale Molecular Optimization from distances matrices by a DC optimization approach, *SIAM Journal of Optimization*,Volume 14, Number 1, 2003, pp.77-116.

[19] Le Thi Hoai An and Pham Dinh Tao, The DC (difference of convex functions) Programming and DCA revisited with DC models of real world nonconvex optimization problems , *Annals of Operations Research* 2005, Vol 133, pp. 23-46.

[20] O.L. Mangasarian (1997), Mathematical Programming in Data Mining, *Data Mining and Knowledge Discovery* 1, 183-201 .

[21] J. B. MacQueen (1967) : "Some Methods for classification and Analysis of Multivariate Observations, *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability"* , Berkeley, University of California Press, 1 :281-297.

[22] A. Meyerson, L. O'Callaghan, S. Plotkin, A k-Median Algorithm with Running Time Independent of Data Size, *Machine Learning*, 56, 6187, 2004.

[23] Julia Neumann, Christoph Schnörr, Gabriele Steidl, SVM-based Feature Selection by Direct Objective Minimisation, *Pattern Recognition, Proc. of 26th DAGM Symposium, LNCS, Springer* , August 2004.

[24] Pham Dinh Tao and Le Thi Hoai An, Convex analysis approach to d.c. programming : Theory, Algorithms and Applications, *Acta Mathematica Vietnamica, dedicated to Professor Hoang Tuy on the occasion of his 70th birthday* , Vol.22, Number 1 (1997), pp. 289-355.

[25] Pham Dinh Tao and Le Thi Hoai An, DC optimization algorithms for solving the trust region subproblem, *SIAM J. Optimization*, Vol. 8, pp. 476-505 (1998).

[26] M.R Rao(1971), Cluster analysis and mathmatical programming, *Journal of American Statistical Association*, 66, 622-626.

[27] R.T. Rockafellar, *Convex Analysis*, Princeton University, Princeton, 1970.

[28] S.Z. Selim and M.A. Ismail, K-means-Type algorithms : a generalized convergence theorem and characterization of local optimilaty, I*EEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6 : 81-87, 1984.

[29] Stefan Weber, Christoph Schnörr, Thomas Schüle, Joachim Hornegger, Discrete Tomography by Convex-Concave Regularization and D.C. Programming, *Technical Report 15/2003, Computer Science Series*, December 2003.

[30] Stefan Weber, Christoph Schnörr, Thomas Schüle, Joachim Hornegger, *Binary Tomography by Iterating Linear Programs*, R. Klette, R. Kozera, L. Noakes and J. Weickert (Eds.), *Computational Imaging and Vision - Geometric Properties from Incomplete Data,* Kluwer Academic Press 2005.

[31] Tina Wong, Randy Katz, Steven McCanne, A Preference Clustering Protocol for Large-Scale Multicast Applications, *Proceedings of the First International COST264 Workshop on Networked Group Communication,* 1999, pp 1-18.

[32] W.H. Wolberg, W.N. Street and O.L. Mangasarian (1995a), Image analysis and machine learning applied to breast cancer diagnosis and prognosis, *Analytical and Quantitative Cytology Histology*, 17 No. 2, 77-87.

[33] W.H. Wolberg, W.N. Street, D.M. Heisey and O.L. Mangasarian (1995b), Computerzied breast cancer diagnosis and prognosis from fine-needle aspirates, *Archives of Surgery*, 130, 511-516.

[34] W.H. Wolberg,W.N. Street, D.M. Heisey and O.L. Mangasarian(1995c), Computer-derived nuclear features distinguish malignant from benign breast cytology, *Human Pathology,* 26, 792-796.

# Une approche en programmation DC pour la Classification floue

**LE THI Hoai An**[1], **LE Hoai Minh**[2], **PHAM DINH Tao**[3]

[1,2] *Laboratoire d'Informatique Théorique et Appliqué LITA*
*Département Informatique UFR MIM - Université de Metz,*
*Ile de Saulcy - 57045 Metz Cedex*
[1] *lethi@sciences.univ-metz.fr* [2] *lehoai@sciences.univ-metz.fr*
[3] *Laboratoire de Mathématiques de l'Institut National des Sciences Appliquées , Place Emile Blondel - BP08, Mont Saint Aignan*
[3] *pham@insa-rouen.fr*

RÉSUMÉ. *Le problème de la classification (clustering) de données est identifié comme une des problèmatiques majeures en extraction des connaissances à partir de données. Dans cet article, nous nous intéressons à Fuzzy C-Means (FCM), une technique très connue parmi les techniques de Fuzzy Clustering. Nous adaptons une approche basée sur la programmation DC (Différence de Convexe) et DCA (DC Algorithme) pour résoudre ce problème. Les simulations numériques que nous avons réalisées montrent la qualité des solutions obtenues par notre algorithme, sa robustesse et sa performance par rapport aux méthodes existantes.*

MOTS-CLÉS : *Programmation DC, DCA, Classification floue, FCM.*

## 1. Introduction

Le problème de la classification automatique (clustering) est considéré comme une des problèmatiques majeures en extraction des connaissances à partir de données. La classification implique la tâche de classer des points dans les classes homogènes de telle sorte que les points dans la même classe soient aussi semblables que possible et les points dans différentes classes soient aussi différentes que possible. De nombreux sous-problèmes ont été identifiés, comme par exemple la sélection des données ou des descripteurs, la variété des espaces de représentation, la nécessité de découvrir des concepts, d'obtenir une hiérarchie, etc. La popularité, la complexité et toutes ces variantes du problème de la classification ont donné naissance à une multitude de méthodes de résolution.

Parmi les techniques de classification, on peut distinguer les méthodes de classification dure et floue (fuzzy). Dans le premier cas, chaque point est classé dans une et une seule classe. Dans le deuxième cas, à chaque point est associée une probabilité d'appartenance à une classe donnée. Cette approche a été appliquée avec succès dans plusieurs problèmes (diagnostic médical [3], classification de textes [4], est de plus en plus utilisé dans le domaine du datamining.

Dans cet article, nous nous intéressons à Fuzzy C-Means (FCM), une technique très connue parmi les techniques de Fuzzy Clustering. Nous adaptons une approche basée sur la programmation DC (Différence de Convexe) et DCA (DC Algorithme) pour résoudre ce problème. Le papier est organisé de la façon suivante. Dans la deuxième partie, nous allons introduire la formulation du problème FCM, ce qui en fait un problème d'optimisation. En suite, nous mettons en évidence la résolution du problème par une méthode basée sur la programmation DC. Dans la dernière partie, nous montrerons quelques résultats numériques pour comparer la performance de notre algorithme avec K-means, une méthode classique en classeing.

## 2. Fomulation du problème FCM

On note $X = \{x_1, x_2, ..., x_n\}$, l'ensemble de $n$ points à classer. Chaque point $x_i$ est un vecteur dans l'espace $I\!R^p$. Nous avons à classer ces $n$ points dans $c$ classes différents. Nous définissons une matrice de pondération $U$ de taille $(c, n)$ où chaque élément $u_{i,k}$ définit le degré d'appartenance d'un point $x_k$ à la classe $c_i$. Nous avons donc :

$$u_{i,k} \in [0,1] \text{ for i=1..c, k=1..n ; } \sum_{i=1}^{c} u_{i,k} = 1 \text{ k=1..n.} \tag{1}$$

Une partition de $n$ points donnés dans $c$ classes n'est rien d'autre que chercher la matrice de pondération correspondante. Nous définissons la fonction objectif de FCM comme suit :

$$J_m(U,V) = \sum_{k=1}^{n}\sum_{i=1}^{c} u_{i,k}^m ||x_k - v_i||^2 \tag{2}$$

où V est un matrice de taille (c,p) dont chaque ligne $v_i$ correspond au centre de la classe $c_i$. Finalement le problème FCM peut être formulé par le problème d'optimitation suivant :

$$(P) \qquad \begin{cases} \min J_m(U,V) \\ s.t \quad u_{i,k} \in [0,1] \text{ for i=1..c, k=1..n} \\ \sum\limits_{i=1}^{c} u_{i,k} = 1 \text{ k=1..n} \end{cases} \tag{3}$$

## 3. Programmation DC - DCA pour résoudre le problème

Avant de passer à la résolution du problème, nous allons introdure les principes de la programmation DC et DCA.

### 3.1. *Introduction à la programmation DC - DCA*

La programmation DC joue un rôle central en optimisation non convexe et optimisation globale car la quasi totalité des problèmes d'optimisation de la vie courante est de nature DC. Elle connaît des développements spectaculaires au cours de cette dernière décennie. DCA est une méthode de descente (primale-duale) pour la résolution d'un programme DC, qui est la minimisation d'une fonction DC de la forme (les contraintes convexes peuvent être incorporées à la fonction objectif à l'aide de la fonction indicatrice) :

$$\alpha := \inf\{f(x) := g(x) - h(x) \ : \ x \in \mathbb{R}^n\}, \tag{4}$$

où $g, h$ sont les fonctions convexes semi-continues et propres sur $\mathbb{R}^n$. Une telle fonction $f$ est appelée fonction DC et les fonctions convexes $g$ et $h$, composantes DC de $f$. Une fonction DC admet une infinité de décomposition DC. La dualité DC est définie via la conjugaison de fonction convexe ( la conjuguée de $g$, notée $g^* : g^*(y) := \sup\{\langle x, y\rangle - g(x) \ : \ x \in \mathbb{R}^n\}$ et le programme dual de (4) est donné par

$$\alpha := \inf\{h^*(y) - g^*(y) \ : \ y \in \mathbb{R}^n\}, \tag{5}$$

(l'espace dual de $\mathbb{R}^n$ est identifié à lui-même). On rappelle la relation suivante (transport des solutions optimales globales en programmation DC) entre l'ensemble des solutions optimales $\mathcal{P}$ de (4) et celui de (5) noté $\mathcal{D}$ ([6] − [5])

$$\cup\{\partial h(x^*) : x^* \in \mathcal{P}\} \subset \mathcal{D} \text{ et } \cup\{\partial g^*(y^*) : y^* \in \mathcal{D}\} \subset \mathcal{P} \tag{6}$$

où les inclusions deviennent égalités sous des hypothèses techniques.

En analyse convexe, $\partial h(x^0) := \{y \in \mathbb{R}^n : h(x) \geq h(x^0) + \langle x - x^0, y\rangle, \forall x \in \mathbb{R}^n\}$ est appelé le sous-différentiel de $h$ au point $x^0$. Tout élément de $\partial h(x^0)$ est appelé gradient de $h$ en $x^0$. Le sous-différentiel $\partial h(x^0)$ est une partie convexe fermée qui coincide avec le gradient $\nabla h(x^0)$ si et seulement $h$ est différentiable en $x^0$. La relation (6) indique que la résolution d'un programme DC implique celle de son dual.

Basé sur les conditions d'optimalité locale et la dualité DC, DCA consiste en la construction de deux suites $\{x^k\}$ et $\{y^k\}$, candidats respectifs aux solutions des problèmes primal et dual que l'on améliore à chaque itération (les deux suites $\{g(x^k) - h(x^k)\}$ et $\{h^*(y^k) - g^*(y^k)\}$ sont décroissantes) et qui convergent vers des solutions primale et duale $x^*$ et $y^*$verifiant des conditions d'optimalité locale et

$$x^* \in \partial g^*(y^*),\ y^* \in \partial h(x^*). \tag{7}$$

Cette relation (7) implique que $x^*$ est une solution optimale du programme convexe

$$\inf\{f(x) + h(x) - [h(x^*) + \langle x - x^*\rangle] : x \in \mathbb{R}^n\} \tag{8}$$

Le schéma général de DCA prend la forme :

$$y^k \in \partial h(x^k);\ x^{k+1} \in \partial g^*(y^k). \tag{9}$$

La première interprétation de DCA est simple : à chaque itération on remplace dans le programme DC primal la deuxième composante DC $h$ par sa minorante affine $h_k(x) := h(x^k) + \langle x - x^k, y^k\rangle$ au voisinage de $x^k$ pour obtenir le programme convexe suivant

$$\inf\{g(x) - h_k(x)\ :\ x \in \mathbb{R}^n\} \tag{10}$$

dont l'ensemble des solutions optimales n'est autre que $\partial g^*(y^k)$.

De manière analogue, la deuxième composante DC $g^*$ du programme DC dual (5) est remplacée par sa minorante affine $(g^*)_k(y) := g^*(y^k) + \langle y - y^k, x^{k+1}\rangle$ au voisinage de $y^k$ pour donner naissance au programme convexe.

$$\inf\{h^*(y) - (g^*)_k(y)\ :\ y \in \mathbb{R}^n\} \tag{11}$$

dont $\partial h(x^{k+1})$ est l'ensemble des solutions optimales. DCA opère ainsi une double linéarisation à l'aide des sous-gradients de $h$ et $g^*$. Il est à noter que DCA travaille avec les composantes DC $g$ et $h$ et non pas avec la fonction $f$ elle-même. Chaque décomposition DC de $f$ donne naissance à un DCA. Pour un programme DC donné, la question de décomposition DC optimale reste ouverte, en pratique on cherche des décompositions DC bien adaptées à la structure spécifique du programme DC étudié pour lesquelles les suites $\{x^k\}$ et $\{y^k\}$ sont faciles à calculer, si possible explicites pour que les DCA correspondants soient moins coûteux en temps et par conséquent capables de supporter de très grandes dimensions.

### 3.2. *Résolution du problème*

Le premier pas est de décomposer la fonction objectif du problème. En appliquant la formule : $2f_1f_2 = (f_1 + f_2)^2 - (f_1^2 + f_2^2)$, nous obtenons :

$$\begin{aligned} J_m(U,V) &= \sum_{k=1}^{n}\sum_{i=1}^{c} u_{i,k}^m \|x_k - v_i\|^2 = \tfrac{1}{2}\sum_{k=1}^{n}\sum_{i=1}^{c}(u_{i,k}^m + \|x_k - v_i\|^2)^2 - \tfrac{1}{2}((u_{i,k}^{2m} + \|x_k - v_i\|^4) \\ &= G(U,V) - H(U,V) \end{aligned}$$

avec

$$G(U,V) = \frac{1}{2}\sum_{k=1}^{n}\sum_{i=1}^{c}(u_{i,k}^m + \|x_k - v_i\|^2)^2 + \chi_K(U,V);\ H(U,V) = \frac{1}{2}\sum_{k=1}^{n}\sum_{i=1}^{c}((u_{i,k}^{2m} + \|x_k - v_i\|^4)$$

**Théorème** : $h(x) = f(x)^p$ est convexe pour $p > 1$ si f est convexe non négative.

Utilisant ce théorème, nous pouvons démontrer facilement que G(U,V) et H(U,V) sont convexes. Nous avons donc une décomposition DC du problème. Nous sommes en face maintenant au calcul du $\partial H(U^l, V^l)$ et $\partial G^*(Y^l, Z^l)$.

*Calcul de* $(Y^l, Z^l) \in \partial H(U^l, V^l)$

$$\begin{aligned} \partial H(U,V) &= (\partial_U H(U,V), \partial_V H(U,V)) \\ &= (mU^{2m-1}, 2 * \sum_{k=1}^{n} (\|x_k - v_i\|^2 (v_i - x_k))_{i=1..c}) \end{aligned} \tag{12}$$

*Calcul de* $(U^{l+1}, V^{l+1}) \in \partial G^*(Y^l, Z^l)$

Rappelons que $(U^{l+1}, V^{l+1}) \in \partial G^*(Y^l, Z^l)$ si et seulement si $(U^{l+1}, V^{l+1})$ est une solution du problème convexe suivant :

$$\begin{cases} \min G(U,V) - \langle (U,V), (Y^l, Z^l) \rangle \\ u_{i,k} \in [0,1] \text{ for i=1..c, k=1..n} \\ \sum_{i=1}^{c} u_{i,k} = 1 \text{ k=1..n} \end{cases} \tag{13}$$

La solution du problème (13) peut être donnée par la méthode Gradient Projeté.

3.2.1. *Schéma DCA*

**Initialisation :** Choisir $U^0 \in I\!R^{c,p}$ et $V^0 \in I\!R^{c,n}$, une tolérance $\epsilon > 0$.
**Répeter** $l = 0, 1, 2, ...$
- Calculer $(Y^l, Z^l) \in \partial H(U^l, V^l)$ à l'aide de (12)
- Calculer $(U^{l+1}, V^{l+1}) \in \partial G^*(Y^l, Z^l)$ en utilisant la méthode Gradient Projeté
**Jusqu'à** $\|(U^{l+1}, V^{l+1}) - (U^l, V^l)\| \leq \epsilon(\|(U^{l+1}, V^{l+1})\|)$
**Construction des classes** Soient $(U^*, V^*)$ la solution calculée par DCA. Le point $x_i$ appartient à la classe $c_j$ telque $u_{ij} = \max u_{i,k}, k = 1..c$

## 4. Expériences numériques

Pour comparer la performance de notre algorithme, nous avons réalisé les tests numériques de la façon suivante. Deux ensembles de jeux de tests sont choisis : les jeux de tests sur des données réelles (Table 1) et le jeux de tests générés aléatoirement (Table 2). Pour les données réelles, nous avons choisi 4 exemples très connus et beaucoup utilisés dans le domaine de classification pour l'évaluation des algorithmes :

- **IRIS** : 150 objets sont classées dans 3 classes différentes.
- **VOTE** : Congressional Votes dataset (Congressional Quarterly Almanac, 98th Congress, 2nd session 1984, Volume XL : Congressional Quarterly Inc. Washington, D.C., 1985).
- **GENE** : L'ensemble de gènes récupérées sur (http ://faculty.washington.edu/kayee/cluster/)
- **ADN** : L'ensemble de 3186 gènes, chaque gène est présentée par une séquence de 60 éléments (ftp ://genbank.bio.net)

La méthode K-means (avec la métrique du Chi2, pour les données **VOTE** et **ADN**) est choisie pour comparer avec notre méthode et la fonction de comparaison est la fonction coût :

$$\sum_{k=1}^{n} \min_{i=1..c} \|x_k - v_i^*\|^2 \tag{14}$$

où $v_i^*$ sont les centres des classes.

## 5. Conclusion

Nous avons présenté la formulation du problème de FCM dans le cadre de la programmation DC et le DCA pour sa résultion. Les simulations numériques que nous avons réalisées sur des données réelles et des données synthétiques montrent la qualité des solutions obtenues par DCA, sa robustesse et sa performance par rapport aux méthodes existantes.

| Data | | | | DCA-Fuzzy | | | | K-means | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Name | n | p | c | Cost | N$^o$it | Time | POMC | Cost | N$^o$it | Time | POMC |
| IRIS | 150 | 4 | 3 | 202.12 | 4 | 0.11 | 3 | 289.39 | 4 | 0.81 | 8 |
| VOTE | 435 | 2 | 2 | 2032.46 | 4 | 0.03 | 11 | 2307.46 | 4 | 0.03 | 18 |
| GENE | 384 | 17 | 5 | 53316x$10^4$ | 15 | 0.80 | 15 | 61303x$10^4$ | 25 | 0.73 | 32 |
| ADN | 3186 | 60 | 3 | 4601x$10^2$ | 8 | 2.00 | 8 | 4686x$10^2$ | 15 | 1.95 | 21 |

POMC = Pourcentage de points mal-classés

| Data | | | DCA-Fuzzy | | | K-means | | |
|---|---|---|---|---|---|---|---|---|
| n | p | c | Cost | N$^o$it | Time | Cost | N$^o$it | Time |
| 100 | 2 | 5 | **298** | 4 | 0.002 | 313 | 8 | 0.005 |
| 500 | 2 | 8 | **304** | 10 | 0.022 | 331 | 12 | 0.030 |
| 1000 | 8 | 10 | **180** | 5 | 0.023 | 202 | 5 | 0.030 |
| 2000 | 3 | 20 | **1973** | 5 | 0.051 | 2098 | 5 | 0.050 |
| 5000 | 20 | 6 | **18641** | 19 | 1.10 | 19665 | 28 | 1.20 |
| 10000 | 20 | 10 | **42291** | 29 | 1.99 | 44031 | 62 | 6.20 |

## 6. Bibliography


1.Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithm*. New York, NY. Plenum Press. 1981.

2.Susana Nascimento, Boris Mirkin, and Fernando Moura-Pires, *Modeling Proportional Membership in Fuzzy Clustering*. IEEE Transactions on Fuzzy Systems, Vol. 11, N˚ 2, April 2003

3.Xiao Ying Wang, Glenn Whitwell, Jonathan M. Garibaldi, *The Application of a Simulated Annealing Fuzzy Clustering Algorithm for Cancer Diagnosis*, SIP 2005, Japan

4.M.E.S. Mendes Rodrigues and L. Sacks. *A scalable hierarchical fuzzy clustering algorithm for text mining.* In : Proc. of the 4th International Conference on Recent Advances in Soft Computing, RASC'2004, pp.269-274, Nottingham, UK, Dec. 2004

5.Le Thi Hoai An and Pham Dinh Tao, *The DC (difference of convex functions) Programming and DCA revisited with DC models of real world nonconvex optimization problems,* Annals of Operations Research 2005, Vol 133, pp. 23-46.

6.Pham Dinh Tao and Le Thi Hoai An, *Convex analysis approach to d.c. programming : Theory, Algorithms and Applications,* Acta Mathematica Vietnamica, dedicated to Professor Hoang Tuy on the occasion of his 70th birthday, Vol.22, Number 1 (1997), pp. 289-355.

7.Le Thi Hoai An and Pham Dinh Tao, *Solving a class of linearly constrained indefinite quadratic problems by DC algorithms,* Journal of Global Optimization, Vol 11, No 3, pp 253-285, 1997.

8.Le Thi Hoai An, Pham Dinh Tao, Le Dung Muu, *Exact penalty in DC programming.* Vietnam Journal of Mathematics, 27 :2 (1999), pp. 169-178.

9.Le Thi Hoai An, Pham Dinh Tao, Huynh Van Ngai, *Exact penalty techniques in DC programming.* Submitted

# Mesures de proximité entre des objets décrits par des histogrammes

**Yves Lechevallier[1], Rosanna Verde[2], Antonio Irpino[2]**

[1]*INRIA - Institut National de Recherche en Informatique et en Automatique*
*Domaine de Voluceau - Rocquencourt B.P. 105 - 78153 Le Chesnay Cedex, France- yves.lechevallier@inria.fr*

[2]*Facoltà di Studi Politici, Seconda Università di Napoli,*
*Via del Setificio, 15 – Belvedere di San Leucio, 81020 Caserta, Italie*
*rosanna.verde@unina2.it – irpino@unina.it*

*RÉSUMÉ. Cet article propose de comparer des distributions empiriques ou des histogrammes, construits à partir d'échantillons de différentes tailles. Afin de prendre en compte la taille de ces échantillons dans la mesure de proximité, nous associons à chaque probabilité un intervalle de confiance estimé à partir de ces échantillons. Nous proposons donc d'utiliser la distance de Hausdorff et une distance basée sur la métrique de Wasserstein pour mesurer la proximité entre deux intervalles de confiance.*

*MOTS-CLÉS : Classification dynamique, distribution empirique, intervalle, distance.*

## 1 Introduction

Le groupement des observations est un procédé commode et économique de représentation de l'information contenue dans un échantillon. Par exemple si nous avons un échantillon de taille n=10000 dont les valeurs observées sur l'intervalle [0,1[ ne peuvent être mesurées qu'au dixième près, il est clair qu'il est inutile de connaître les 10000 valeurs et qu'il suffit d'indiquer les comptages #1, …,#10 associés aux intervalles Ii=[(i-1)/10,i/10[ c'est-à-dire connaître seulement *l'histogramme de l'échantillon.*
Par exemple nous avons un ensemble de capteurs qui comptent le nombre d'impacts observés dans leur zone d'observation, dans ce cas nous n'avons pas la localisation précise de l'impact mais uniquement l'information que tel capteur a eu cet impact dans sa zone. Nous pouvons aussi prendre comme exemple l'analyse des sources d'émission acoustique décrite dans H. Hamdan [HAM 05], dans ce cas nos données sont les données incertaines discrétisées décrites dans cette thèse.
Ainsi à chaque expérience $s$ notre échantillon de taille $n_s$ sera représenté par son histogramme défini sur l'ensemble des comptages effectués par notre ensemble de capteurs. Chaque expérience a un support $U(s)$ qui est un intervalle semi-ouvert qui est décomposé en un ensemble d'intervalles semi-ouverts disjoints dont l'union est égale à ce support. Pour chaque intervalle $B_l^i$ semi-ouvert on associe une variable aléatoire égale à $I(B_l^i) = \sum_{h=1}^{n_i} I_{x_h}(B_l^i)$ où $I_{x_h}(B_l^i) = \begin{cases} 1 \text{ si } x_h \in B_l^i \\ 0 \text{ sinon} \end{cases}$

## 2 Mesures de proximités

Soit une variable continue $Y$, à chaque échantillon $s$ de taille $n_s$ issu de cette variable aléatoire on a l'histogramme de l'échantillon $((I_1,n_{s1}),\ldots, (I_h,n_{sh}), \ldots,(I_H,n_{sH}))$ où ces $H$ couples sont construits sur un ensemble $I=\{I_1,...,I_h,...,I_H\}$ de $H$ intervalles semi-ouverts disjoints, dits *élémentaires* , vérifiant les propriétés suivantes :

i) $\bigcup_{h=1}^{H} I_h \subset [\min(y_i),\max(y_i)]$;

ii) $I_h \cap I_{h'} = \varnothing$ si $h \neq h'$;

avec des poids $(n_{s1},...,n_{sH})$ vérifiant les propriétés suivantes : $n_{sh} \geq 0$ et $\sum_{h=1}^{H} n_{sh} = n_s$ où $n_{sh}$ est le nombre de réalisations mises dans l'intervalle $I_h$.

Pour comparer de manière efficace deux histogrammes il est indispensable de prendre en compte la taille de chacun des échantillons associés à ces histogrammes. Une solution est d'associer au couple $(s,h)$ un intervalle de confiance $CI(p_{sh})$ de sa probabilité $p_{sh}$ qui est naturellement estimée à partir de la fréquence empirique qui est égale à $q_{sh} = n_{sh}/n_s$. Cet intervalle de confiance $CI(p_{sh})$ est calculé en fonction du risque $\alpha$ (en général 0.05, 0.01 ou 0.001) par :

$$\overset{\circ}{CI}(p_{sh}) = \left[ q_{sh} - |z_{\alpha/2}| \sqrt{\frac{q_{sh}(1-q_{sh})}{n_s}}; q_{sh} + |z_{\alpha/2}| \sqrt{\frac{q_{sh}(1-q_{sh})}{n_s}} \right]$$

avec $P[U > |z_{\alpha/2}|] = \alpha/2$ et $U$ est la loi normale centrée et réduite.

Donc il est possible de décrire un histogramme comme un vecteur de couples $((I_1,CI(p_{s1})),\ldots, (I_h,CI(p_{sh})), \ldots,(I_H,CI(p_{sH})))$. Chaque couple est constitué d'un intervalle élémentaire $I_h$ et d'un intervalle de confiance $CI(p_{sh})$.

En utilisant la notation milieu et demi-longueur pour décrire cet intervalle nous avons $q_{sh}$ comme milieu de l'intervalle et $l_{sh} = |z_{\alpha/2}| \sqrt{\frac{q_{sh}(1-q_{sh})}{n_s}}$ comme moitié de sa longueur. Cet intervalle peut être aussi décrit de manière ensembliste de la façon suivante :

$$CI(p_{sh}) = \{x \in [0,1] / x = q_{sh} + (2t_h - 1) l_{sh} \text{ avec } t_h \in [0,1]\}$$

Ainsi notre histogramme n'est plus uniquement décrit par une liste de fréquences mais aussi par une liste d'intervalles qui mesurent la qualité de l'estimation de ces fréquences.

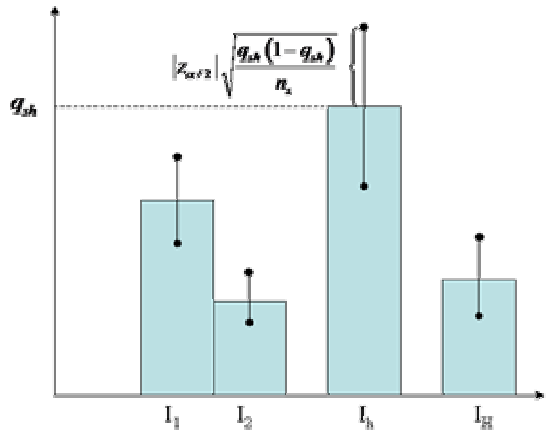



*Fig 1. Histogramme avec l'intervalle de confiance associé à chaque fréquence*

## *2.1 Métrique de Wasserstein*

Si les deux échantillons *s* et *r* sont décrits par deux histogrammes ayant le même support il est possible de définir une extension de la distance euclidienne en prenant en compte les différents intervalles de confiance associés aux fréquences en utilisant la métrique de Wasserstein de la façon suivante :

$$d^2(s,r) = \sum_{h=1}^{H} \int_0^1 \left\{ \left[ q_{sh} + |z_{\alpha/2}| \sqrt{\frac{q_{sh}(1-q_{sh})}{n_s}} (2t_h - 1) \right] - \left[ q_{rh} + |z_{\alpha/2}| \sqrt{\frac{q_{rh}(1-q_{rh})}{n_r}} (2t_h - 1) \right] \right\}^2 dt_h =$$

$$= \sum_{h=1}^{H} (q_{sh} - q_{sh})^2 + \frac{1}{3} \sum_{h=1}^{H} (l_{sh} - l_{rh})^2$$

où $\sum_{h=1}^{H} (q_{sh} - q_{sh})^2$ est la distance euclidienne entre les fréquences et $\frac{1}{3} \sum_{h=1}^{H} (l_{sh} - l_{rh})^2$ est une mesure intégrant les différents effectifs des deux échantillons.

**Remarques :**

Il est simple de démontrer que : $\lim_{n_r, n_s \to \infty} (l_{sh} - l_{rh})^2 = \lim_{n_r, n_s \to \infty} \left( |z_{\alpha/2}| \sqrt{\frac{q_{sh}(1-q_{sh})}{n_s}} - |z_{\alpha/2}| \sqrt{\frac{q_{rh}(1-q_{rh})}{n_r}} \right)^2 = 0$.

Ainsi si les fréquences sont estimées à partir d'échantillons de grandes tailles alors cette distance est identique à la distance euclidienne entre les distributions.
Si les deux histogrammes sont égaux alors la distance proposée est égale à :

$$d^2(s,r) = \frac{1}{3} \sum_{h=1}^{H} \left( |z_{\alpha/2}| \sqrt{\frac{q_h(1-q_h)}{n_s}} - |z_{\alpha/2}| \sqrt{\frac{q_h(1-q_h)}{n_r}} \right)^2 = \frac{1}{3} \sum_{h=1}^{H} |z_{\alpha/2}|^2 \, q_h(1-q_h) \left( \frac{1}{\sqrt{n_s}} - \frac{1}{\sqrt{n_r}} \right)^2$$

ainsi elle ne dépend plus que des effectifs des échantillons associés aux distributions *s* et *r*.

### *2.2 Distance de Hausdorff*

Comme à chaque intervalle de nos histogrammes nous avons associé un intervalle nous pouvons utiliser la distance de Hausdorff d'où :

$$\begin{aligned} \mathrm{d}_{Haus}(r,s) &= \sum_{h=1}^{H} \mathrm{d}_{Haus}(CI(p_{rh}), CI(p_{sh})) \\ &= \sum_{h=1}^{H} \max\left( \left|(q_{sh} - l_{sh}) - (q_{rh} - l_{rh})\right|, \left|(q_{sh} + l_{sh}) - (q_{rh} + l_{rh})\right| \right) \\ &= \sum_{h=1}^{H} \left( |q_{sh} - q_{rh}| + |l_{sh} - l_{rh}| \right) \end{aligned}$$

Si la taille des échantillons tend vers l'infini alors cette distance de Hausdorff tend vers la distance de la norme L1.

## 3 Utilisation des mesures proposées dans le cadre de la classification

Le choix du prototype doit être adapté à la mesure de distance utilisée. Avec la distance de Hausdorff ces histogrammes sont modélisés comme des variables de type intervalle donc nous pouvons utiliser la méthode de classification de type Nuées Dynamique [CHL 02] basée sur la recherche de la meilleure partition $P^*$ d'un ensemble d'objets $E$ en $k$ classes non vides, au sens d'un critère $\Delta$ qui mesure l'adéquation entre les représentations des classes $C_1, .., C_k$ de la partition $P$ en $k$ classes non vides, avec les prototypes $G=(G_1, ..., G_k)$:

$$\Delta(P^*, G^*) = Min\{\Delta(P, G)\}.$$

Une extension de cette méthode de classification a été proposée par [IRP 06] dans le cas de la métrique de Wasserstein. Dans notre contexte, le prototype est aussi un histogramme avec des intervalles de confiance associés. Ce prototype est simplement un histogramme si la taille des échantillons est grande.

## 4 Bibliographie


[CHL 02] CHAVENT M. et LECHEVALLIER Y ; (2002). "Dynamical Clustering of interval data. Optimization of an adequacy criterion based on Hausdorff distance", *Classification, Clustering and Data Analysis,* K. Jaguga et al. (Eds.), Springer, pp 53-60.

[BAR 99] DEL BARRIO E, GINE E et MATRAN C.; (1999,2003) "Central limit theorems for the Wasserstein distance between the empirical and the true distributions". Ann. Probab. (1999),vol 27, no.2, pp 1009–1071 (2003) vol. 31, no. 2, pp.1142-1143

[GIB 02] GIBBS, A.L. et SU, F.E.; (2002): "On choosing and bounding probability metrics", *International Statistical Review*, 70.

[HAM 05] HAMDAN H. ; (2005) Développement de méthodes de classification et de discrimination pour le contrôle par émission acoustique d'appareils à pression, Thèse à Université de Technologie de Compiègne.

[MAL 72] MALLOWS, C. L.; (1972): "A note on asymptotic joint normality". *Annals of Mathematical Statistics*, 43(2), pp 508-515.

[IRP 06] IRPINO, A., VERDE R. et LECHEVALLIER, Y; (2006) "Dynamical clustering of quantitative symbolic data using Wasserstein metric", Compstat 2006, Rome, Italie

[VER 05] VERDE, R. et IRPINO, A. (2005): "A New Distance for Symbolic Data Clustering", *CLADAG 2005, Book of short papers*, MUP, pp 393-396.

# Consensus par groupements fréquents


## Bruno Leclerc

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



*RÉSUMÉ. Soit* $\mathbf{\mathcal{D}} = (\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_k)$ *un profil de classifications sur un ensemble E, que l'on veut agréger en une classification unique* $\mathcal{D}$*. Les classifications considérées ici sont des ensembles de classes deux à deux incomparables pour l'inclusion. Pour un entier p compris entre* 1 *et k, on définit un* consensus par groupements fréquents *en considérant les classes maximales incluses dans des éléments d'au moins p des* $\mathcal{D}_i$*. On étudie les propriétés de cette règle de consensus et on la caractérise en termes d'emboîtements.*



*MOTS-CLÉS : Classe, Consensus, Emboîtement, Famille de Sperner, Implication, Motif fréquent.*


### 1 Introduction

Soit $E$ un ensemble fini, et $R \subseteq (\mathcal{P}(S))^2$ une relation binaire sur l'ensemble des parties de $E$. Dans des articles et communications antérieurs [DOM 04b, LEC 04, LEC 05], nous avons montré l'unicité d'une classification $\mathcal{M}$ (sous la forme d'une famille de Moore) vérifiant par rapport à $R$ deux conditions remontant à Adams [ADA 86] ; ces conditions garantissent, en un sens, que la relation d'emboîtement (voir la section 5 ci-dessous) de $\mathcal{M}$ s'ajuste bien à $R$. Il se pose alors un problème d'existence, car si le consensus d'Adams réalise bien un tel ajustement dans le cas des hiérarchies, il est facile de trouver des relations $R$ pour lesquelles on aboutit à une impossibilité. Dans la communication présentée l'an dernier aux journées de Montréal, nous avons mis en évidence des situations de consensus par emboîtements où, non seulement la classification $\mathcal{M}$ existe, mais de plus son obtention est proche de celle des "motifs fréquents" pour la recherche de règles d'association en fouille des données. L'objet de cet exposé est de systématiser cette observation dans un cadre suffisamment général.

### 2 Définitions

On considère ici des classifications (non hiérarchiques) consistant en un ensemble $\mathcal{D}$ de parties (classes) d'un ensemble donné $E$ à $n$ éléments. La classification $\mathcal{D}$ est de plus supposée être une *famille de Sperner propre*, c'est-à-dire que ses classes sont deux à deux incomparables pour l'inclusion, avec $\mathcal{D} \neq \{E\}$. C'est un *recouvrement* de $E$ si l'union de ses classes est $E$ et une *partition de E* si, de plus, ses classes sont deux à deux disjointes. Nous notons $\mathbf{S}$ l'ensemble des familles de Sperner propres sur $E$. Pour toute partie $A$ de $E$, nous disons que $A$ est un *groupement de* $\mathcal{D}$ s'il existe au moins une classe $C$ de $\mathcal{D}$ contenant $A$.

Soit un profil $\mathcal{D} = (\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_k)$ de telles classifications que nous cherchons à agréger en une seule famille de Sperner $\mathcal{D}$. Posons $K = \{1, 2, \ldots, k\}$. Nous associons au profil $\mathcal{D}$ un *indice de groupement* $g_{\mathcal{D}}$ sur l'ensemble $\mathcal{P}(E)$ des parties de $E$, en posant, pour toute partie $A$ de $E$,

$$g_{\mathcal{D}}(A) = \#\{i \in K : A \subseteq C \text{ pour au moins une classe } C \text{ de } \mathcal{D}_i\}.$$

Le nombre $g_{\mathcal{D}}(A)$ est donc celui des classifications de $\mathcal{D}$ dont $A$ constitue un groupement.

Nous associons une fonction de consensus $F_p : \mathbf{S}^k \rightarrow \mathbf{S}$ à l'indice $g_{\mathcal{D}}$ et à un entier $p \in K$. Nous disons qu'une partie $A$ de $E$ est un *groupement p-fréquent* si $g_{\mathcal{D}}(A) \geq p$, et nous définissons le *consensus par groupements p–fréquents* de $\mathcal{D}$, $F_p(\mathcal{D})$ comme étant la famille de Sperner des groupements $p$-fréquents maximaux. Notons que $F_k(\mathcal{D})$ est l'ensemble des parties $C$ non vides de $E$ de la forme $C = \bigcap_{1 \leq i \leq k} C_i$, avec $C_i \in \mathcal{D}_i$ pour tout $i \in K$, et maximales avec cette propriété. Ainsi, si $\mathcal{D}$ est un profil de partitions, on retrouve le croisement des partitions de $\mathcal{D}$. De son côté, $F_1(\mathcal{D})$ contient ceux des éléments de $\bigcup_{1 \leq i \leq k} \mathcal{D}_i$ qui sont maximaux pour l'inclusion.

Bien qu'assez naturels, de tels groupements fréquents semblent avoir été assez peu étudiés (avec l'exception notable des "formes fortes" de Diday [DID 71]).

## 3 Obtention

La présentation d'un algorithme n'entre pas dans les objectifs de cet exposé. On observe simplement que les groupements fréquents généralisent les *motifs fréquents* recherchés pour l'extraction de règles en fouille de données : les motifs fréquents correspondent au cas où chacune des familles $\mathcal{D}_i$ est réduite à une seule classe $C_i$. Certains des nombreux algorithmes qui ont été proposés dans la littérature pour l'obtention des motifs fréquents sont directement généralisables au problème ci-dessus. Ainsi l'algorithme "prototypal" Apriori (Agrawal et Srikant 1994) procède par exploration arborescente de $\mathcal{P}(E)$, l'élagage de nombreuses branches permettant l'examen de données de grande taille. Cet élagage correspond à la sélection de "motifs" potentiellement fréquents. Ensuite, pour un tel motif $B$, on parcourt la base de données $\mathcal{D}$ pour déterminer si le nombre de ses éléments contenant $B$ atteint ou non l'entier $p$. L'adaptation aux groupements fréquents décrits ci-dessus est immédiate : on parcourt de même les familles $\mathcal{D}_i$ à tour de rôle, en passant à la famille $\mathcal{D}_{i+1}$ dès que l'on a trouvé dans $\mathcal{D}_i$ une classe contenant $B$.

De nombreux algorithmes ont suivi Apriori, dont ils sont souvent des améliorations (cf., e.g., [HIP 00], [BEN 04]). Leur adaptation aux groupements fréquents doit être examinée au cas par cas.

## 4 Quelques propriétés

Nous donnons ici quelques propriétés de la fonction d'agrégation $F_p$. Nous avons déjà vu que $F_p$ est une fonction de $\mathbf{S}^k$ dans $\mathbf{S}$, qui associe la famille de Sperner $F_p(\mathcal{D})$ à tout profil $\mathcal{D}$ de familles de Sperner. On montre facilement que, de plus :

- si les $\mathcal{D}_i$ sont tous des recouvrements de $E$, alors $F_p(\mathcal{D})$ est un recouvrement de $E$,
- si les $\mathcal{D}_i$ sont tous des ensembles d'intervalles d'un ordre total fixé $L$ sur $E$, alors $F_p(\mathcal{D})$ est un ensemble d'intervalles de $L$,
- si les $\mathcal{D}_i$ sont tous des partitions de $E$, et si $p = k$, alors $F_p(\mathcal{D})$ est une partition de $E$.

En revanche, pour $p < k$, $F_p(\mathcal{D})$ n'est pas toujours une partition lorsque $\mathcal{D}$ est un profil de partitions. Prenons par exemple $E = \{a, b, c, d\}$, $k \geq 3$, et un profil $\mathcal{D}$ de partitions, dont $k$-2 égales à $\{\{a, b, c\}, \{d\}\}$, les deux autres étant $\{\{a, b\}, \{c\}, \{d\}\}$ et $\{\{a\}, \{b, c\}, \{d\}\}$. On obtient $F_{k-1}(\mathcal{D}) = \{\{a, b\}, \{b, c\}, \{d\}\}$, qui n'est pas une partition.

Observons que le fait qu'une partie $A$ est ou non un groupement de $F_p(\mathcal{D})$ ne dépend que de la valeur de l'indice $g_{\mathcal{D}}(A)$, et est donc indépendant des éléments ou parties de $E\backslash A$. On en déduit que la fonction d'agrégation $F_p$ vérifie les deux propriétés suivantes, de type "arrowien" (cf., e.g., [DAY 03]), la seconde étant particulièrement forte.

*Unanimité pour les groupements :*

$$[A \subseteq E \text{ et } g_{\mathcal{D}}(A) = k] \Rightarrow [A \text{ est un groupement de } F_p(\mathcal{D})]$$

*Neutre-monotonie pour les groupements* :

$$[\mathcal{D}, \mathcal{D}' \in \mathbf{S}^k, A, A' \subseteq E \text{ et } g_{\mathcal{D}}(A) \leq g_{\mathcal{D}}(A')] \Rightarrow [A \text{ est un groupement de } F_p(\mathcal{D}) \Rightarrow A' \text{ est un groupement de } F_p(\mathcal{D}')]$$

On montre facilement que la réunion de ces deux propriétés constitue une caractérisation des fonctions $F_p$. Nous donnons aussi ci-dessous une caractérisation de ces fonctions d'un tout autre type.

## 5 Une caractérisation en termes d'emboîtements

A une famille $\mathcal{D}$ de parties de $E$, on associe (classiquement) une relation binaire d'implication, et aussi une relation d'emboîtement.

- La relation d'*implication* $I$ sur $\mathcal{P}(E)$ associée à $\mathcal{D}$ correspond à l'idée que la partie $B$ est systématiquement associée à la partie $A$ dans $\mathcal{D}$, en ce sens que $A$ *implique* $B$ (ce qui est noté $A \rightarrow B$, ou $(A, B) \in I$, ou $A\ I\ B$ – on parle aussi de dépendance fonctionnelle ou de règle d'association) si toute classe de $\mathcal{D}$ contenant $A$ contient aussi $B$ (cf. Caspard et Monjardet [CAS 03] pour des résultats et références sur ces implications).
- La relation d'*emboîtement* $\mathcal{E}$ sur $\mathcal{P}(E)$ associée à $\mathcal{D}$ correspond à l'idée que la partie $B$ est plus générale que la partie $A$ par rapport à $\mathcal{D}$, en ce sens que $A$ *est emboîtée dans* $B$ (ce qui est noté $(A, B) \in \mathcal{E}$, ou $A\ \mathcal{E}\ B$) si $A \subset B$ (inclusion stricte) et s'il existe une classe de $\mathcal{D}$ contenant $A$ et ne contenant pas $B$ (cf. [DOM 04a] sur ces emboîtements).

Les relations $I$ et $\mathcal{E}$ se déduisent l'une de l'autre ; par exemple, on a $\mathcal{E} = \{(A, B) \in \mathcal{P}(E)^2 : A \subset B \text{ et } (A, B) \notin I\}$. S'il est plus commode d'énoncer le théorème suivant sous forme d'emboîtements, il a donc une contrepartie en termes d'implications. Dans la suite, nous notons $\mathcal{E}_i$ la relation d'emboîtement associée à la classification $\mathcal{D}_i$, et, pour $p \in K$, $\mathcal{E}^{(p)} = \bigcup_{J \subseteq K,\ |J| \geq p} \bigcap_{i \in J} \mathcal{E}_i$ est l'ensemble des couples $(A, B) \in (\mathcal{P}(E))^2$ qui appartiennent à au moins $p$ des $\mathcal{E}_i$ (notons que $\mathcal{E}^{(p)}$ n'est pas nécessairement une relation d'emboîtement)

**Théorème.** Soit $\mathcal{D} = (\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_k) \in \mathbf{S}^k$ un profil de familles de Sperner propres sur $E$. Alors, pour tout $p \in K$, la famille $\mathcal{D} = F_p(\mathcal{D})$ est l'unique famille de Sperner sur $E$ vérifiant les deux conditions :

1. pour tout $C \in \mathcal{D}$, $(C, E) \in \mathcal{E}^{(p)}$,

2. $\mathcal{E}^{(p)} \subseteq \mathcal{E}$.

Preuve. Montrons que $F_p(\mathcal{D})$ vérifie les conditions 1 et 2. Par définition, il existe pour toute classe $C$ de $\mathcal{D}$ une partie $J$ de $K$, de cardinal au moins $p$, telle que $C \subseteq C_i \in \mathcal{D}_i$ pour tout $i \in J$. Ceci entraîne $(C, E) \in \mathcal{E}_i$ pour tout $i \in J$, d'où $(C, E) \in \mathcal{E}^{(p)}$. La fonction $F_p$ satisfait donc la condition 1.

Soient $A, B \subseteq E$ tels que l'on a $(A, B) \in \mathcal{E}^{(p)}$. On a donc $A \subset B$ et il existe une partie $J$ de $K$ de cardinal au moins $p$ telle que, pour tout $i \in J$, on a une classe $C_i$ de $\mathcal{D}_i$ avec $A \subseteq C_i$ et $B \not\subseteq C_i$. Prenons $J$ telle que $C =$

$\bigcap_{i \in J} C_i$ est maximale avec ces propriétés, ce qui entraîne $C \in F_p(\boldsymbol{\mathcal{D}})$. On a alors $A \subseteq C$ et $B \not\subseteq C$, d'où $(A, B) \in Œ$, ce qui correspond à la propriété 2.

Il reste à montrer l'unicité. Celle-ci découle de travaux antérieurs qui ne sont pas détaillés ici. On observe qu'à une famille de Sperner $\mathcal{D}$ correspond une unique famille de Moore $\mathcal{M}$ (on obtient $\mathcal{M}$ en faisant toutes les intersections possibles d'éléments de $\mathcal{D}$ et en ajoutant la classe $E$). Par construction, $\mathcal{D}$ et $\mathcal{M}$ ont les mêmes relations d'implication et d'emboîtement. On utilise alors les résultats portant sur l'unicité d'une famille de Moore vérifiant 1. et 2. qui ont été donnés dans [DOM 04b] et [LEC 04].

## 6 Conclusion

Nous avons défini une famille de règles de consensus par groupements fréquents s'appliquant à tout profil $\boldsymbol{\mathcal{D}} = (\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_k)$ de familles de Sperner propres, et nous avons caractérisé ces règles. Il reste à généraliser ces résultats en les étendant à des modèles de classification plus généraux.

## 7 Bibliographie


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

[AGR 94] AGRAWAL R., SRIKANT R., " Fast algorithms association rules *", Proceedings of the 20th VLDB Conference*, Santiago, Chile, 1994, p. 1-7.

[BEN 04] BEN YAHIA S., MEPHU NGUIFO E., " Approches d'extraction de règles d'association basées sur la correspondance de Galois ", *RSTI-ISI* 9 (n° 3-4), 2004, p. 23-55.

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

[DAY 03] DAY W.H.E., McMORRIS F.R., *Axiomatic Consensus Theory in Group Choice and Biomathematics*, SIAM, Philadelphia, 2003.

[DID 71] DIDAY E., " Une nouvelle méthode en classification automatique et reconnaissance des formes : la méthode des nuées dynamiques ", *Revue de Statistique Appliquée*, vol XIX, 1971, p. 19-33.

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

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

[HIP 00] HIPP J., GÜNTZER U., NAKHAEIZADEH G., " Algorithms for association rule mining – a general survey and comparison ", *SIGKDD Explorations* 2 (n°1), 2000, p. 58-64.

[LEC 04] LECLERC B., " On the consensus of closure systems ", *Annales du LAMSADE* 3, 2004, p. 237-247.

[LEC 05] LECLERC B., " Implications, emboîtements et ajustements de classifications ", *in* V. Makarenkov, G. Cucumel, F.-J. Lapointe (directeurs), *Comptes rendus des 12èmes rencontres de la Société Francophone de Classification*, Montréal, UQAM, 2005, p. 17-20.

# Analyse des groupes de gènes co-exprimés (AGGC) : un outil automatique pour l'interprétation des expériences de biopuces

**Ricardo Martinez[1], Nicolas Pasquier[1], Claude Pasquier[2], Martine Collard[1], Lucero Lopez[3]**

*[1] Laboratoire I3S,*
*Université de Nice Sophia Antipolis,*
*2000, route des lucioles,*
*06903 Sophia-Antipolis, France.*

*[2] Laboratoire Biologie Virtuelle,*
*Université de Nice Sophia Antipolis*
*Centre de Biochimie, Parc Valrose,*
*06108 Nice cedex 2, France.*

*[3] Projet Odyssee*
*INRIA Sophia Antipolis,*
*2004 route des Lucioles*
*06905 Sophia Antipolis, France.*

*RÉSUMÉ. La technologie des biopuces permet de mesurer les niveaux d'expression de milliers de gènes dans différentes conditions biologiques générant ainsi des masses de données à analyser. De nos jours, l'interprétation de ces volumineux jeux de donnés à la lumière des différentes sources d'informations est l'un des principaux défis dans la bio-informatique. Nous avons développé une nouvelle méthode appelée AGGC (Analyse des Groupes de Gènes Co-exprimés) qui permet de constituer de manière automatique des groupes de gènes à la fois fonctionnellement riches, i.e. qui partagent les mêmes annotations fonctionnelles, et co-exprimés. AGGC intègre l'information issue des biopuces, i.e. les profils d'expression des gènes, avec les annotations fonctionnelles des gènes obtenues à partir des sources d'informations génomiques comme* Gene Ontology. *Les expérimentations menées avec cette méthode ont permis de mettre en évidence les principaux groupes de gènes fonctionnellement riches et co-exprimés dans des expériences de biopuces*[1].

*MOTS-CLÉS : biopuces, ontologie, co-expression, gène et annotation.*

## 1 Introduction

L'analyse de données de biopuces en utilisant les diverses sources d'informations génomiques, continuellement alimentées par des volumes croissants de données, représente un challenge important. Ces sources d'informations sont sémantiques (taxonomies, thésaurus et ontologies), littéraires et bibliographiques (articles, librairies en ligne, etc.), et constituées de bases de données d'expériences et de nomenclatures. L'un des défis majeurs actuels dans ce domaine est l'intégration automatique des connaissances biologiques issues des sources d'informations mentionnées ci-dessus avec les données d'expression de gènes [ATT 01]. Un premier bilan des méthodes développées pour répondre à ce défi a été fait par Chuaqui [CHU 02].

Nous ciblons ici l'enrichissement de deux axes de recherche récemment développés, *séquentiel* et *a priori*, qui exploitent de multiples sources d'annotations telles que *Gene Ontology* (GO)[2]. Ces annotations sont des informations fonctionnelles, relationnelles et syntaxiques sur les gènes.

Dans l'axe séquentiel, partant des clusters de gènes co-exprimés (groupes de gènes qui ont un profil d'expression similaire), des sous-ensembles de gènes co-annotés (partageant la même annotation) sont détectés. Ensuite, la significativité statistique de ces sous-ensembles de gènes co-annotés est testée. Parmi les méthodes dans cet axe citons Onto Express [DRA 03], EASE [HOS 03] et THEA [PAS 04].

Dans l'axe a priori, partant des groupes fonctionnellement riches (GFR), i.e. des groupes de gènes co-annotés, l'information contenue dans les profils d'expression est intégrée. La significativité statistique des

[1] Informations supplémentaires et programme exécutable AGGC : *http://www.i3s.unice.fr/~rmartine/AGGC*

[2] Ontologie d'annotation de gènes *Gene Ontology project* : *http://www.geneontology.org/*

GFR est ensuite testée en utilisant un test basé sur un score enrichi [MOO 03], un test issu d'un z-score [KIM 05] ou un test basé sur une pc-value (distribution hypergéométrique) [BRE 04].
Notre approche, appelée AGGC (Analyse des Groupes de Gènes Co-exprimés), est inspirée de l'axe a priori : les GFR sont d'abord formés à partir de la GO est une fonction qui synthétise l'information contenue dans les données d'expression est appliquée afin d'obtenir une liste ordonnée de gènes [BRE 04]. Dans cette liste, les gènes sont triés par variabilité d'expression décroissante. La significativité statistique des GFR obtenus est alors testée à l'aide d'une preuve d'hypothèse de manière similaire à *Onto express* [DRA 03]. Finalement, nous obtenons des GFR co-exprimés et statistiquement significatifs. La méthode AGGC est une extension de la méthode IGA permettant d'obtenir tous les sous-ensembles possibles de GFR de gènes co-exprimés, sans se limiter au GFR constitué des gènes les plus exprimés.
Cet article est organisé de la manière suivante : dans la section 2 nous décrivons les données de validation ainsi que les outils utilisés ; l'algorithme AGGC est décrit dans la section 3 ; les résultats obtenus sont présentés dans la section 4 ; la section 5 conclu l'article.

## 2 Données et Méthodes

### *2.1 Jeux de données et prétraitement*

Afin d'évaluer notre approche, l'algorithme AGGC a été appliqué à des jeux de donnés dérivés de celui de DeRisi [DER 97] qui est l'un des plus étudiés dans ce domaine. Ce jeu mesure la variation d'expression des gènes durant le processus cellulaire de « diauxic shift » pour la levure *Saccharomyces Cerevisiae*. Ce processus correspond à la transition de la phase de fermentation du sucre en éthanol (croissance anaérobique) vers la phase de respiration aérobie de la levure.
Ces données indiquent les niveaux d'expression des 6199 ORF's (Opening Reading Frame) de la levure, qui est un organisme entièrement séquencé, pour 7 points temporels durant le processus. Les données ont été prétraitées en prenant le $\log_2$ des ratios (pour considérer les inductions et les répressions cellulaires de façon numériquement égale) et en appliquant l'algorithme d'imputation des K plus proches voisins [LIT 02] afin de traiter les valeurs manquantes (1.9% du total).

### *2.2 Groupes de gènes fonctionnellement riches (GFR)*

Nous avons généré une base de données (SGOD) contenant toutes les annotations GO pour chacun des gènes de la levure à partir de GO et SGD. Pour chaque gène sont stockées toutes les annotations du gène et de ses parents. L'ensemble des GFR a été construit à partir de requêtes exécutées sur le SGOD : chaque GFR correspond a un couple constitué d'une annotation GO (*go-term*) et de la liste des gènes annotés par celle-ci.

### *2.3 Mesure des profils d'expression des gènes*

Afin d'incorporer les profils d'expression des gènes, nous nous sommes servi d'une mesure de variabilité d'expression, le *F-score*, qui est plus robuste que d'autres mesures telles que l'*anova*, le *fold change* ou les statistiques *t-student* [RIV 05]. Cette mesure nous permet d'établir une liste des gènes, *g-rank*, ordonnés par variabilités d'expression décroissantes. Nous avons utilisé le programme SAM [TUS 01] pour calculer le *F-score* associé à chaque gène.

## 3 Analyse des groupes de gènes co-exprimés (AGGC)

AGGC est basé sur l'idée que tout changement affin (co-expression) d'un sous-ensemble de gènes appartenant à une GFR est physiologiquement important. Nous disons que deux gènes sont co-exprimés s'ils sont proches par rapport à la métrique de variabilité d'expression (*F-score*). L'algorithme AGGC permet de déterminer pour chaque GFR la *pc-value* qui estime sa cohérence (à partir de *g-rank*) et donc de détecter les groupes statistiquement significatifs.

### *3.1 Algorithme AGGC*

AGGC commence par déterminer la liste *g-rank* à partir des niveaux d'expression et les GFR à partir de la SGOD. Pour chaque GFR constitué de $n$ gènes, l'algorithme détermine les $n(n+1)/2$ sous-ensembles de gènes dont nous voulons tester la co-expression. Pour chacun de ces sous-ensembles nous calculons sa *pc–value* à partir du test suivant décrit ci-dessous.

$H_0$ : probabilité que les $x$ gènes d'un de ces sous-ensembles aient été associés par hasard.

Cette probabilité correspond à la distribution hyper-géométrique suivante :

$$p(X = x \mid N, \mathrm{R}_{g(x)}, n) = \frac{\binom{R_{g(x)}}{x}\binom{N - R_{g(x)}}{n - x}}{\binom{N}{n}} \quad où \quad p(X = 0 \mid N, \mathrm{R}_{g(x)}, n) = 0$$

$N$ : nombre total de gènes dans le jeu de données.
$n$ : nombre de gènes dans le GFR.
$x$ : position (n° d'ordre) du gène dans le GFR.
$\mathrm{r}_{g(x)}$ : rang du gène de position $x$ dans *g-rank*.
$R_{g(x)}$ : nombre de rangs qui séparent le gène $x$ de son prédécesseur (dans le GFR) dans le *g-rank*.

$R_{g(x)}$ est obtenu par : $R_{g(x)} = \mathrm{r}_{g(x)} - \mathrm{r}_{g(x-1)} + 1$ où $R_{g(0)} = \mathrm{r}_{g(0)} = 1$.

La *pc–value* correspondant à cette preuve d'hypothèse est [DRA 03] :

$$pc - value(x) = 1 - \sum_{k=1}^{x} p(X = k \mid N, R_{\mathrm{g(k)}}, n)$$

Afin d'accepter ou rejeter l'hypothèse $H_0$ nous utiliserons comme seuil de significativité : $p–value = Min\ \{N^{-1}, |\Omega|^{-1}\}$ où $|\Omega|$ est la cardinalité de l'ensemble de tous les annotations fonctionnelles. Ainsi pour chaque GFR, si $pc–value(x) < p–value$ alors on rejet $H_0$, i.e. le GFR est statistiquement significatif.

## 4 Résultats

Afin d'évaluer notre méthode, nous avons comparé les résultats obtenus par DeRisi, par IGA et AGGC. Les résultats obtenus avec AGGC pour les gènes sur-exprimés sont présentés dans le tableau 1. Les groupes identifiés par AGGC et DeRisi sont en **gras**, les groupes identifiés seulement par AGGC sont en *italique*, et le seul groupe identifié par AGGC et IGA est souligné. AGGC a permis de retrouver sept des neuf groupes de gènes obtenus manuellement par DeRisi. Les deux groupes annotés « glycogen metabolism » et « glycogen synthase » n'ont pas été identifiés par AGGC car ils s'expriment uniquement dans la phase initiale du processus et que nous n'avons pas intégré les informations sur les voies métaboliques. Toutefois AGGC a identifié huit groupes statistiquement significatifs et cohérents vis a vis du processus étudié.

| Groupe GO fonctionnellement riche | *n* gènes | *x* gènes sur-exprimés | *pc-value* |
|---|---|---|---|
| *proton-transporting ATP synthase complex* | 2 | 2 | 4.38E-06 |
| *invasive growth (sensu Saccharomyces)* | 5 | 3 | 6.13E-06 |
| *signal transduction during filamentous growth* | 2 | 2 | 8.77E-06 |
| **respiratory chain complex II** | 4 | 4 | 3.75E-05 |
| **succinate dehydrogenase activity** | 4 | 4 | 3.75E-05 |
| **mitochondrial electron transport** | 4 | 4 | 3.75E-05 |
| *aerobic respiration* | 36 | 10 | 3.30E-05 |
| **tricarboxylic acid cycle** | 14 | 5 | 5.09E-05 |
| **tricarboxylic acid cycle** | 14 | 5 | 6.54E-05 |
| *gluconeogenesis* | 12 | 2 | 9.64E-05 |
| *response to oxidative stress* | 10 | 3 | 1.55E-06 |
| *filamentous growth* | 8 | 4 | 9.06E-05 |
| *vacuolar protein catabolism* | 4 | 2 | 2.63E-05 |
| **respiratory chain complex IV** | 8 | 2 | 4.05E-04 |
| **cytochrome-c oxidase activity** | 8 | 2 | 4.05E-04 |

**Tableau 1 : GFR sur-exprimés obtenus par AGGC avec une *p-value* de $7X10^{-4}$.**

Des résultats similaires, accessibles sur la page du projet, ont été obtenus pour les GFR sous-exprimés.

## 5 Conclusion

L'algorithme AGGC présenté dans cet article permet d'identifier automatiquement les groupes de gènes co-exprimés significatifs et fonctionnellement riches sans avoir de connaissance a priori des résultats. Il est extensible aux annotations biologiques de toutes natures et aux diverses mesures de variabilité proposées dans le domaine.
AGGC analyse tous les sous-ensembles possibles de chaque GFR, accroissant ainsi la sensibilité de la détection des groupes de gènes co-exprimés, même en présence de données très bruitées. A l'extrême il peut produire des résultats statistiques significatifs sans avoir besoin de répliquer les expériences. Il est également robuste contre les mauvaises assignations lors de la création des groupes fonctionnels à partir des sources publiques (annotations erronées) ou bien de processus automatiques (erreurs de nommage, fautes d'orthographe, etc.).
Les résultats expérimentaux ont montré la validité de l'approche et ont permis d'identifier des informations pertinentes sur les processus biologiques étudiés. Afin d'identifier les groupes de gènes s'exprimant seulement dans certaines phases du processus, nous prévoyons ultérieurement d'intégrer les informations concernant les voies métaboliques.

## 6 Bibliographie


[ATT 01] ATTWOOD T., MILLER C.J, *Which craft is best in bioinformatics?* Compute. Chem., 25, 2001, p. 329-339.

[BRE 04] BREITLING R., AMTMANN A., HERZYK P., *IGA : A simple tool to enhance sensitivity and facilitate interpretation of microarray experiments*, BMC Bioinformatics, 5:34, 2004.

[CHU 02] CHUAQUI R., *Post-analysis follow-up and validation of microarray experiments*. Nature Genetics, 32, 2002, p. 509 – 514.

[DER 97] DERISI J., IYER L., BROWN V., *Exploring the metabolic and genetic control of gene expression on a genomic scale*. Science, n° 278, 1997, p. 680-686.

[DRA 03] DRAGHICI S., KHATRI P., et al. *Global functional profiling of gene expression,* Genomics, 81, 2003, p. 1-7.

[HOS 03] HOSACK D., DENNIS G., et al., *Identifying biological themes within lists of genes with EASE,* Genome Biology, 4, R70, 2003.

[KIM 05] KIM S., VOLSKY D., *PAGE : Parametric Analysis of Gene Set Enrichment*, BMC Bioinformatics, 6:144, 2005.

[LIT 02] LITTLE R. et RUBIN D., *Statistical Analysis with Missing Data,* John Wiley & Sons, 2002.

[MOO 03] MOOTHA V., LINDGREN C., ERIKSSON K., SUBRAMANIAN A., et al., *PGC-l'alpha-reponsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes,* Nat Genet., 34(3), 2003, p. 267-273.

[PAS 04] PASQUIER C., GIRARDOT F., JEVARDAT K., CHRISTEN R., *THEA : Ontology-driven analysis of microarray data*. Bioinformatics, vol.20, issue 16, 2004.

[RIV 05] RIVA A., CARPENTIER A., TORRESANI B., HENAUT A., *Comments on selected fundamental aspects of microarray analysis*, Computational Biology and Chem. 29, 2005, p. 319-336.

[ROB 02] ROBINSON M., et al., *FunSpec : a Web based cluster interpreter for yeast*. BMC Bioinformatics, 3, 35, 2002.

[TUS 01] TUSHER V., TIBSHIRANI R., CHU G., *Significance analysis of microarrays applied to the ionizing radiation response*, Proc. Nat. Acad. Sci. USA, 98 (9), 2001, p. 5116-21.

# Réflexions sur l'extraction de motifs rares

**Sandy Maumus**[1,2], **Amedeo Napoli**[1], **Laszlo Szathmary**[1], **Yannick Toussaint**[1]

[1] *LORIA, 54506 Vandoeuvre-lès-Nancy*
{*maumus, napoli, szathmar, yannick*}*@loria.fr*
[2] *INSERM U525, 54000 Nancy*
*Sandy.Maumus@nancy.inserm.fr*

RÉSUMÉ. *Les études en fouille de données se sont surtout intéressées jusqu'à présent à l'extraction de motifs fréquents et à la génération de règles d'association à partir des motifs fréquents. L'algorithme le plus célèbre ayant permis d'atteindre ces objectifs est Apriori, qui a été suivi par toute une famille d'algorithmes mis au point par la suite et possédant tous la caractéristique d'extraire l'ensemble des motifs fréquents ou un sous-ensemble de ces motifs (motifs fermés fréquents, motifs fréquents maximaux, générateurs minimaux). Dans cet article, nous posons le problème de la recherche des motifs rares ou non fréquents, qui se trouvent dans le complémentaire de l'ensemble des motifs fréquents. Ce type de motif n'a jamais vraiment fait l'objet d'une étude systématique, malgré l'intérêt et la demande existant dans certains domaines d'application. Ainsi, en biologie ou en médecine, il peut se révéler très important pour un praticien de repérer des symptômes non habituels ou des effets indésirables exceptionnels se déclarant chez un patient pour une pathologie ou un traitement donnés.*

MOTS-CLÉS : *Fouille de données, extraction de motifs, motifs fréquents et rares.*

## 1. Introduction

La fouille de données a pour objectif d'identifier des motifs et des associations implicites dans de grandes bases de données [HAN 01]. Un motif est un ensemble de propriétés ou attributs tandis qu'une association est de la forme `A` $\longrightarrow$ `B` où `A` et `B` sont des motifs. La recherche de motifs fréquents et de règles d'association sont parmi les tâches les plus importantes en fouille de données. Les études en fouille de données se sont surtout intéressées jusqu'à présent à l'extraction de motifs fréquents — motifs dont la fréquence d'apparition parmi les individus d'une population donnée est supérieure à un seuil donné — et à la génération de règles d'association dérivant des motifs fréquents. L'algorithme le plus célèbre permettant d'extraire des motifs et des règles est Apriori, qui a été suivi par toute une famille d'algorithmes mis au point par la suite et possédant tous la caractéristique d'extraire l'ensemble des motifs fréquents ou un sous-ensemble de ces motifs (motifs fermés fréquents, motifs fréquents maximaux, générateurs minimaux [BAS 02]).

Dans cet article, nous posons le problème de la recherche de *motifs rares* ou *non fréquents*, qui se trouvent dans le complémentaire de l'ensemble des motifs fréquents. Les problèmes de la fouille de motifs rares et de la génération des règles d'association rares qui en dérivent n'ont pas encore été traités en détail dans la littérature (sachant que cet article est une version abrégée, revue et corrigée de [SZA 06] et qu'il existe aussi une étude théorique sur la complexité de la recherche des motifs fréquents et non fréquents dans [BOR 02]). Dans la suite, nous expliquons d'abord l'intérêt que peuvent revêtir les motifs et règles rares, puis nous donnons les définitions et les grandes lignes de la recherche de motifs et de règles rares. L'article se termine par une discussion sur les motifs et règles rares, accompagnée d'une comparaison avec la recherche de motifs fréquents et l'extraction de règles d'association, ainsi que d'une série de questions qui sont en cours d'investigation.

## 2. Motivations

La découverte de motifs rares peut se révéler très intéressante en médecine et en biologie. Considérons d'abord une base de données médicales et le problème de l'identification de la cause de maladies cardio-vasculaires (MCV). Une règle d'association fréquente (extraite d'un motif fréquent) comme "{niveau élevé de cholestérol} $\longrightarrow$ {MCV}" permet de faire émerger l'hypothèse que les individus ayant un fort taux de cholestérol ont un risque élevé de MCV. À l'opposé, s'il existe un nombre conséquent de végétariens dans la base de données, alors une règle d'association rare comme "{végétarien} $\longrightarrow$ {MCV}" permet de faire émerger l'hypothèse qu'un végétarien a un risque faible de contracter une MCV. Dans un tel cas, les motifs {végétarien} et {MCV} sont tous deux fréquents, mais le motif {végétarien, MCV} est lui-même rare.

Le deuxième exemple, qui s'appuie sur les données réelles de la cohorte STANISLAS [MAU 05], montre l'intérêt de l'extraction des motifs rares pour la fouille de cohortes supposées saines. La cohorte STANISLAS est composée d'un millier de familles françaises présumées saines. L'objectif principal de l'étude de la cohorte est de mettre en évidence l'influence des facteurs génétiques et environnementaux sur la variabilité des risques cardio-vasculaires. Parmi les informations intéressantes à extraire de cette base de données figurent les profils associant des données génétiques à des valeurs extrêmes ou limites des paramètres biologiques. Cependant, ces associations sont plutôt rares dans les cohortes supposées saines. Dans ce contexte, l'extraction de motifs rares peut s'avérer très utile pour étudier les variations dans les profils — les profils rares pouvant conduire à des problèmes néanmoins — et ainsi avoir une idée plus complète des associations entre paramètres, ce que ne permet pas la seule recherche de motifs fréquents.

Le troisième exemple est en rapport avec la pharmacovigilance, qui est une branche de la pharmacologie dédiée à la détection et l'étude des effets indésirables des médicaments. L'extraction des motifs rares dans une base de données des effets indésirables de médicaments peut contribuer à un suivi plus efficace des effets indésirables graves et servir ensuite à prévenir les accidents mortels qui aboutissent au retrait de certains médicaments (comme par exemple le retrait de la cérivastatine, médicament hypolipémiant, en août 2001).

## 3. La recherche de motifs rares

Une méthode générique pour retrouver les motifs rares est présentée ci-après. Dans un premier temps, la méthode identifie un ensemble générateur minimal appelé *ensemble des motifs rares minimaux* ou MRMs. Dans un second temps, les MRMs sont utilisés pour retrouver tous les motifs rares. Avant d'arriver aux détail de la méthode, un rappel des définitions classiques est proposé.

- Une base de données formelle s'appuie sur le produit cartésien $\mathtt{O} \times \mathtt{A}$ associé à une relation R, où $\mathtt{O} = \{\mathtt{o}_1, \mathtt{o}_2, \ldots, \mathtt{o}_m\}$ est un ensemble d'objets, $\mathtt{A} = \{\mathtt{a}_1, \mathtt{a}_2, \ldots, \mathtt{a}_n\}$ est un ensemble d'attributs et $\mathtt{R} \subseteq \mathtt{O} \times \mathtt{A}$ est une relation telle que $\mathtt{R}(\mathtt{o}_i, \mathtt{a}_j)$ signifie que l'objet $\mathtt{o}_i$ possède l'attribut $\mathtt{a}_j$.
- Un ensemble d'attributs forme un *motif* dont la taille est le nombre d'attributs qui le composent. Le *support* d'un motif P correspond au nombre d'objets contenant le motif et un motif est *fréquent* si son support est supérieur ou égal à un seuil de fréquence minimum donné (noté minsupp).
- La recherche de motifs fréquents consiste à engendrer tous les motifs dont le support est supérieur ou égal au seuil minsupp, en appliquant les principes suivants [AGR 96] :
  (i) "*la recherche des motifs fréquents commence par traiter les motifs de longueur minimale ; le support des motifs est calculé après un accès à la base données formelle ; les motifs fréquents sont conservés et les motifs non fréquents sont élagués*",
  (ii) "*tous les sous-motifs d'un motif fréquent sont fréquents*",
  (iii) "*tous les super-motifs d'un motif non fréquent sont non fréquents*".
  De plus, un motif P est *fermé* s'il n'existe aucun super-motif Q de P ($\mathtt{P} \subseteq \mathtt{Q}$) de même support.

Cela étant, un motif est dit *rare* ou *non fréquent* si son support est inférieur ou égal à un *support maximum*, noté maxsupp. Dans ce qui suit, la valeur de maxsupp se calcule à partir de celle de minsupp, à savoir maxsupp =

minsupp − 1 (ici minsupp et maxsupp sont donnés en valeur absolue). La recherche de motifs rares consiste à engendrer tous les motifs dont le support est inférieur ou égal au seuil *maxsupp*.

Il peut exister un intervalle de valeurs entre *minsupp* et *maxsupp*. Mais dans cet article, nous avons travaillé avec un cas particulier sans intervalle, c'est à dire que pour nous un motif est rare s'il n'est pas fréquent. Cela implique l'existence d'une seule frontière entre motifs rares et fréquents. Une telle frontière est étudiée et discutée dans [BOU 03, CAL 05].

L'ensemble des motifs rares et l'ensemble des motifs fréquents ont tous deux un sous-ensemble minimal générateur. Dans le cas des motifs fréquents, ce sous-ensemble est l'ensemble des *motifs fréquents maximaux* (MFMs). Un motif est un motif fréquent maximal s'il est fréquent et si tous ses super-motifs ne sont pas fréquents.

De façon complémentaire, un *motif rare minimal* (MRM) est un motif rare dont tous les sous-motifs ne sont pas rares. L'ensemble des motifs rares minimaux forme un ensemble générateur minimal à partir duquel tous les motifs rares peuvent être retrouvés, comme tous les motifs fréquents peuvent être retrouvés à partir des motifs fréquents maximaux. Pour les motifs fréquents maximaux, tous les sous-motifs possibles des MFMs sont considérés et leur support est calculé après un passage sur la base de données. De façon duale, tous les super-motifs des motifs rares minimaux sont considérés, puis le calcul du support de ces motifs se fait grâce grâce à un passage sur la base de données.

Parmi les motifs rares se distinguent les motifs rares de support 0 (zéro), appelés *motifs zéros*, et les motifs rares de support non nul, ou *motifs non zéros*. Le nombre de motifs rares zéros peut être très élevé. De façon analogue à un motif rare minimal, un motif est *générateur zéro minimal* (GZM) si c'est un motif zéro et si tous ses sous-motifs sont des motifs non zéros (tous ses super-motifs sont bien sûr des motifs zéros).

Les motifs rares minimaux peuvent être retrouvés simplement à l'aide de l'algorithme Apriori de la façon suivante : quand un motif non fréquent donc rare P est détecté — son support est inférieur ou égal à maxsupp (ou strictement inférieur à minsupp) — aucun des super-motifs de P n'est considéré par la suite, car ces super-motifs sont de manière sûre non fréquents. Puisque l'algorithme Apriori explore le treillis des motifs niveau par niveau — du "bas vers le haut" ou des tailles minimales aux tailles maximales —, il calcule nécessairement le support des motifs rares minimaux. Les motifs rares minimaux sont élagués et l'algorithme Apriori construit ensuite les motifs candidats de longueur k dont tous les sous-motifs de longueur $(k-1)$ sont fréquents. Si, pour un candidat P de longueur k, un des sous-motifs de P, soit Q, de longueur $(k-1)$, n'est pas fréquent, alors P est rare ; et en plus cela signifie que Q est un sous-motif rare minimal. Ainsi, l'espace de recherche dans le treillis des motifs est réduit de façon significative.

Une légère modification d'Apriori suffit pour conserver les motifs rares minimaux : dès que le support d'un motif candidat P est inférieur au support minimum, alors P est enregistré dans l'ensemble des motifs rares minimaux. Ensuite, tous les motifs rares sont retrouvés à partir des motifs rares minimaux. Pour cela, il faut engendrer tous les super-motifs possibles des motifs rares minimaux. Les générateurs zéros minimaux permettent de filtrer les motifs zéros pendant la génération des super-motifs rares.

## 4. Synthèse et questions

Dans cet article, une méthode pour extraire les motifs rares dans une base de données a été présentée. La méthode s'appuie sur l'algorithme de recherche de motifs fréquents Apriori et se compose de deux parties : (i) recherche d'un sous-ensemble générateur minimal des motifs rares (MRMs), (ii) recherche à partir des MRMs des motifs rares dont le support n'est pas nul. Ce travail de recherche est l'un des premiers à s'intéresser de façon systématique et spécifique aux motifs rares. L'algorithme Apriori a été le premier algorithme de recherche de motifs fréquents et il a été suivi de nombreux autres algorithmes plus efficaces et performants. De manière similaire, il ne fait aucun doute que la méthode de recherche de motifs rares présentée ici sera dans un avenir proche améliorée et les auteurs de l'article s'y emploient. Ainsi, des sous-ensembles utiles pour la recherche de motifs fréquents ont été découverts, comme les motifs fermés fréquents, les motifs fréquents maximaux, les générateurs (clés) minimaux, etc. De façon duale, de tels sous-ensembles doivent pouvoir être définis pour les motifs rares, puisque

par exemple le complémentaire des motifs fréquents maximaux est l'ensemble des motifs rares minimaux. Une autre question intéressante est la suivante : comme les motifs fermés fréquents déterminent sans ambiguïté tous les motifs fréquents et leur support, existe-t-il un sous-ensemble analogue qui déterminerait les motifs rares ? En outre, va suivre l'exploitation de la recherche des motifs rares pour la génération de règles d'association rares.

Pour terminer, il faut évoquer une série de questions plus théoriques qui se posent également :

- Des représentations condensées des motifs fréquents sont introduites dans [BOU 03, CAL 05], ainsi qu'une bordure négative et une bordure positive entre motifs fréquents et rares, et des ensembles disjonctifs libres et $\delta$-libres ($\delta$ mesure le nombre de contre-exemples à une règle d'association). Quel est le dual des ces ensembles pour les motifs rares ?
- La base de Duquenne-Guigues permet d'engendrer les règles d'association exactes ou de confiance 1 [GUI 86] (ces règles sont aussi celles qui peuvent être extraites du treillis de concepts associé à la base de données formelle). Cette base peut être calculée à l'aide des motifs dits *pseudo-fermés* qui se définissent comme suit : un motif $\mathtt{P}$ est pseudo-fermé s'il n'est pas fermé et si tous les sous-motifs pseudo-fermés $\mathtt{Q} \subset \mathtt{P}$ qu'il contient strictement ont une fermeture contenue dans $\mathtt{P}$ (voir par exemple [GAN 99, STU 01]). Il est intéressant d'étudier le rapport exact existant entre les motifs pseudo-fermés et les motifs rares : est-ce que l'un se dérive de l'autre et est-ce que l'un permet de calculer l'autre.

## 5. Bibliographie


[AGR 96] AGRAWAL R., MANNILA H., SRIKANT R., TOIVONEN H., VERKAMO A. I., Fast Discovery of Association Rules, FAYYAD U., PIATETSKY-SHAPIRO G., SMYTH P., UTHURUSAMY R., Eds., *Advances in Knowledge Discovery and Data Mining*, Menlo Park, California, 1996, AAAI Press / MIT Press, p. 307–328.

[BAS 02] BASTIDE Y., TAOUIL R., PASQUIER N., STUMME G., LAKHAL L., Pascal : un algorithme d'extraction des motifs fréquents, *Technique et science informatiques*, vol. 21, n$^o$ 1, 2002, p. 65–95.

[BOR 02] BOROS E., GURVICH V., KHACHIYAN L., MAKINO K., On the Complexity of Generating Maximal Frequent and Minimal Infrequent Sets, *Symposium on Theoretical Aspects of Computer Science*, 2002, p. 133-141.

[BOU 03] BOULICAUT J.-F., BYKOWSKI A., RIGOTTI C., Free-Sets : A Condensed Representation of Boolean Data for the Approximation of Frequency Queries, *Data Mining and Knowledge Discovery*, vol. 7, n$^o$ 1, 2003, p. 5–22.

[CAL 05] CALDERS T., RIGOTTI C., BOULICAUT J.-F., A survey on condensed representations for frequent sets, BOULICAUT J.-F., RAEDT L. D., MANNILA H., Eds., *Constraint-based mining and Inductive Databases*, Lecture Notes in Computer Science 3848, Springer-Verlag, Berlin, 2005, p. 64–80.

[GAN 99] GANTER B., WILLE R., *Formal Concept Analysis*, Springer, Berlin, 1999.

[GUI 86] GUIGUES J., DUQUENNE V., Familles minimales d'implications informatives résultant d'un tableau de données binaires, *Mathématiques et Sciences Humaines*, vol. 95, 1986, p. 5–18.

[HAN 01] HAN J., KAMBER M., *Data Mining : Concepts and Techniques*, Morgan Kaufmann Publishers, San Francisco, 2001.

[MAU 05] MAUMUS S., NAPOLI A., SZATHMARY L., VISVIKIS-SIEST S., Exploitation des données de la cohorte STANISLAS par des techniques de fouille de données numériques et symboliques utilisées seules ou en combinaison, *Atelier Fouille de Données Complexes dans un Processus d'Extraction des Connaissances - EGC 2005, Paris, France*, 2005, p. 73–76.

[STU 01] STUMME G., TAOUIL R., BASTIDE Y., PASQUIER N., LAKHAL L., Intelligent structuring and reducing of association rules with formal concept analysis, BAADER F., BREWKA G., EITER T., Eds., *KI 2001*, Lecture Notes in Artificial Intelligence 2174, Springer-Verlag, Berlin, 2001, p. 335–350.

[SZA 06] SZATHMARY L., MAUMUS S., PETRONIN P., TOUSSAINT Y., NAPOLI A., Vers l'extraction de motifs rares, RITSCHARD G., DJERABA C., Eds., *Extraction et gestion des connaissances (EGC'2006), Lille*, RNTI-E-6, Cépaduès-Éditions Toulouse, 2006, p. 499–510.

## 6. Annexe : un exemple de recherche de motifs fréquents et rares

Dans cet article, la base de données formelle suivante, reprise de [BAS 02], est utilisée (notée $\mathcal{D}$, table 1). Le seuil de fréquence (absolu), `minsupp`, est fixé à 3, et donc le seuil de non fréquence, `maxsupp`, à 2. Les motifs rares et fréquents issus de cette base de données formelle peuvent être visualisés sur la figure 1.

| | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | x | | x | x | |
| 2 | | x | x | | x |
| 3 | x | x | x | | x |
| 4 | | x | | | x |
| 5 | x | x | x | | x |

**TAB. 1.** *La base de données formelle $\mathcal{D}$ prise en exemple.*

Étant donné un seuil de fréquence, les motifs peuvent être classifiés en deux catégories : les motifs rares, dont la fréquence est en-dessous du seuil, et les motifs fréquents, dont la fréquence est au-dessus du seuil. Une frontière existe entre ces deux catégories, qui peut être visualisée sur le treillis des parties de l'ensemble des attributs considérés (voir figure 1). En bas du treillis se trouve le plus petit motif, ou motif de longueur nulle, qui correspond à l'ensemble vide. À chaque niveau se situent les motifs de même taille. Au sommet du treillis se trouve le motif le plus long qui contient tous les attributs. Le support de chaque motif par rapport à la base de données formelle $\mathcal{D}$ est indiqué dans le coin en haut à droite à côté de chaque motif.

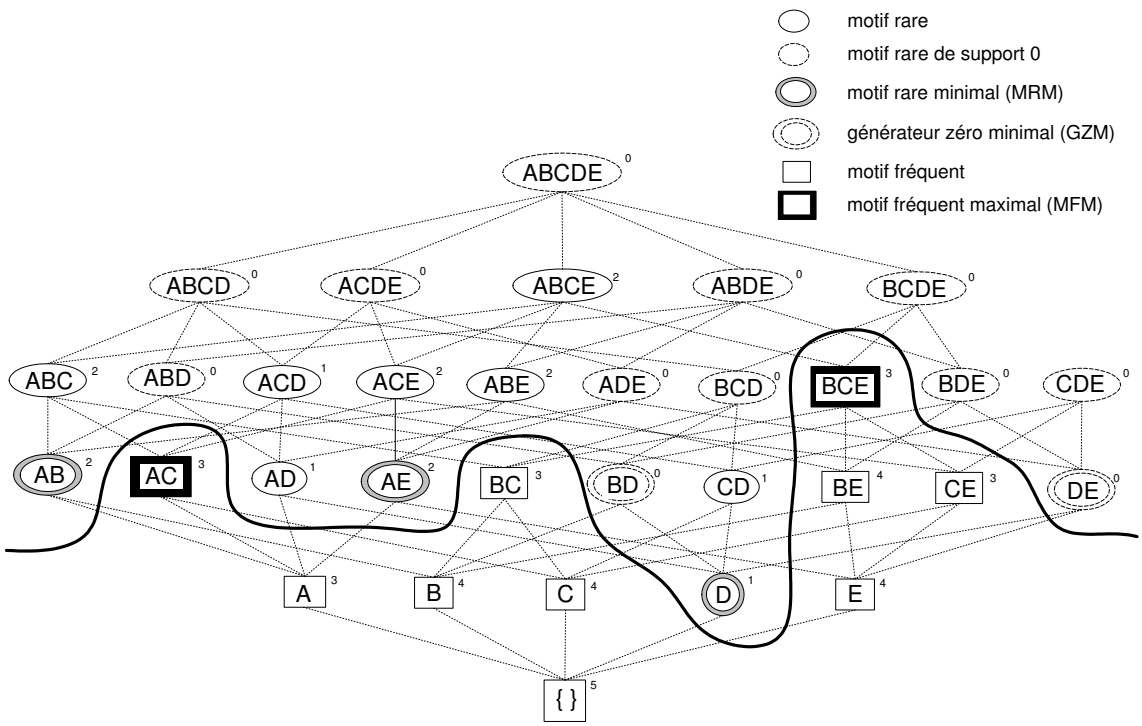



**FIG. 1.** *Le treillis des parties de la base de données formelle $\mathcal{D}$ avec les motifs fréquents et les motifs rares.*

L'ensemble des motifs rares forme un sup-demi-treillis car il est est fermé pour l'opération sup — tout sup de deux rares est rare, mais il ne forme pas un inf-demi-treillis, car l'inf de deux rares n'est pas forcément rare. De façon duale, les motifs fréquents forment un inf-demi-treillis mais pas un sup-demi-treillis.

Sur la figure 1, les deux générateurs zéros minimaux sont {BD} et {DE}. Les GZMs forment une représentation condensée et sans perte d'information des motifs zéros : à partir des GZMs, tous les motifs zéros peuvent être retrouvés — avec leur support, qui est toujours 0 — ; pour cela, il suffit d'engendrer tous les super-motifs possibles des GZMs en utilisant les attributs de la base de données. Mais, cette génération n'est pas effectuée à cause du trop grand nombre de motifs zéros mais aussi parce que seuls les GZMs sont utiles. La seconde partie de la méthode de recherche permet de retrouver tous les motifs rares non-zéros à partir des MRMsà l'aide d'une approche par niveau. Si un candidat intègre un sous-motif GZM, alors ce candidat est de manière sûre un motif zéro et peut donc être élagué. Les GZMs permettent ainsi de réduire l'espace de recherche lors de la recherche des motifs rares.

# DC Programming and DCA for Diversity Data Mining

**NGUYEN CANH Nam[a], LE THI Hoai An[b] and PHAM DINH Tao[c]**

[a,c] *Laboratoire Modélisation, Optimisation et Recherche Opérationnelle*
*LMI - Institut National des Sciences Appliquées ,*
*Place Emile Blondel - BP08, Mont Saint Aignan*
*{nguyencn, pham}@insa-rouen.fr*
[b] *Laboratoire d'Informatique Théorique et Appliqué LITA*
*Département Informatique UFR MIM - Université de Metz,*
*Idle Saulcy - 57045 Metz Cedex*
*lethi@univ-metz.fr*

RÉSUMÉ. *Dans ce travail, nous présentons une nouvelle approche basée sur la DC programmation et DCA pour une classe de problèmes très importants en Data Mining. Une reformulation DC est établie grâce à la pénalité exacte et la technique de relaxation SDP est développée pour initialiser DCA. Nous illustrons notre algorithme sur des données fournies dans [1] et des données de grande dimension générées selon le schéma indiqué dans [1].*

MOTS-CLÉS : *Maximum Diversity, Data Mining, DC programming, DCA, SDP relaxation.*

## 1 Introduction

Data mining is one of important domains in the real world, especially in recent years. In this, many optimization problems are proposed which need to be solved definitively. This work is interested in one model, which gives rise to the realms, called Diversity Data Mining, or precisely, the maximum diversity problem is concerned. This problem arises in many realms of business and government, e.g. Environmental Balance, Medical Treatment, Genetic Engineering, Molecular Structure Design, Agricultural Breeding Stocks, Right Sizing the Firm, Composing Jury Panels (see [10] for more detail).

The model of problem can be presented as follows. Let $N = \{1, 2, \ldots, n\}$ and $E = \{e_i : i \in N\}$ be a population of $n$ elements and $e_{ik}$ with $k \in R = \{1, 2, ..., r\}$ the $r$ values of attributes of each element. The objective is to select a subset of size $m < n$ to maximize the *diversity* of the chosen elements.

To express formally the objective function, we associate a measure of diversity $d_{ij}$ with each pair of element $e_i$ and $e_j$ ; that is, $d_{ij}$ is some function of the elements $e_{ik}$ and $e_{jk}$ ($k \in R$). An example, $d_{ij}$ is the Euclidean distance between two elements, $d_{ij} = \sqrt{\sum_{k=1}^{r} (e_{ik} - e_{jk})^2}$.

Let $M$ be a subset of $N$ and the overall diversity will be $\alpha(M) = \sum_{i<j:i,j\in M} d_{ij}$. The Diversity Maximization problem consists of maximizing the function $\alpha(M)$ subject to $|M| = m$.

Let $x_j$ be a binary variable denoting whether or not element $e_j$ is chosen to be a member of the selected subset, then its overall diversity can be written as $\alpha(M) = \sum_{i=1}^{n} \sum_{j=1}^{i-1} d_{ij} x_i x_j$, or equivalently $\frac{1}{2} x^T D x$ with $d_{jj} = 0$ for all $j = 1, 2, \ldots, n$.

Finally, we can model the Diversity Maximization problem in the following form

$$\alpha(M) = \max \quad f(x) = \frac{1}{2} x^T D x \qquad (\mathbf{DDM})$$

$$\text{subject to} \qquad \sum_{j=1}^{n} x_j = m$$

$$x_j \in \{0, 1\} \quad \forall j = 1, 2, \ldots, n$$

The maximum diversity problem is known to be NP-hard. In literature, exact solution methods are presented in the work of Glover et al. [3], and Kuo et al [11]. But these methods have been limited to instances with very small dimension. So these approaches are not viable for problems large enough to be of significant practical interest.

Many heuristic methods have been proposed in recent years, see [2, 10, 9, 1] which are designed to approximate large scale problems as they appear in real world application. Nevertheless, it is a great challenge to develop deterministic methods for this problem with large scale dimension. The approach presented in this article is certainly a step in this direction.

DCA has been introduced by T. PHAM DINH in 1985, as an extension of his subgradient algorithms for convex maximization programs, and extensively developed by H.A. LE THI and T. PHAM DINH since 1994 to solve DC programs [5, 12, 13, 8]. The DCA has been successfully applied in solving real world nonconvex programs to which it quite often gave global solutions and proved to be more robust and more efficient than related standard methods, especially in large scale setting. DCA is actually one of the rare algorithms for nonsmooth nonconvex programming which allows solving large-scale DC programs.

A so-called DC program is that of minimizing a DC function over a convex set. According to the theory of DC programming and via the well-known result concerning exact penalty, we can easily reformulate the original problem as DC programs. We then suggested using DC programming approach and DCA for their solution.

## 2 DC programming and DCA

Consider the general DC program

$$\alpha = \inf\{f(x) := g(x) - h(x) \ : \ x \in \mathbb{R}^n\} \quad (P_{dc})$$

with $g, h \in \Gamma_0(\mathbb{R}^n)$ (the convex cone of all lower semicontinuous proper convex functions on $\mathbb{R}^n$). A such function $f$ is called DC function, and $g - h$, DC decomposition of $f$ while the convex functions $g$ and $h$ are DC components of $f$.

We have the dual program of $(P_{dc})$

$$\alpha = \inf\{h^*(y) - g^*(y) : y \in \text{dom } h^*\}$$

that is written, in virtue of the natural convention in DC programming, say $+\infty - (+\infty) = +\infty$ :

$$\alpha = \inf\{h^*(y) - g^*(y) : y \in Y\}. \quad (D_{dc})$$

where

$$g^*(y) := \sup\{\langle x, y\rangle - g(x) \ : \ x \in \mathbb{R}^n\}$$

is the conjugate function of $g$ We observe the perfect symmetry between primal and dual DC programs : the dual to $(D_{dc})$ is exactly $(P_{dc})$.

Based on local optimality conditions and duality in DC programming, the DCA consists in the construction of two sequences $\{x^k\}$ and $\{y^k\}$, candidates to be optimal solutions of primal and dual programs respectively, such that the sequences $\{g(x^k) - h(x^k)\}$ and $\{h^*(y^k) - g^*(y^k)\}$ are decreasing, and $\{x^k\}$ (resp. $\{y^k\}$) converges to a primal feasible solution $\widetilde{x}$ (resp. a dual feasible solution $\widetilde{y}$) verifying local optimality conditions and

$$\widetilde{x} \in \partial g^*(\widetilde{y}), \quad \widetilde{y} \in \partial h(\widetilde{x}). \tag{1}$$

These two sequences $\{x^k\}$ and $\{y^k\}$ are determined in the way that $x^{k+1}$(resp. $y^k$) is a solution to the convex program $(P_k)$ (resp. $(D_k)$) defined by

$$\inf\{g(x) - h(x^k) - \langle x - x^k, y^k\rangle : x \in \mathbb{R}^n\}, \quad (P_k)$$

$$\inf\{h^*(y) - g^*(y^{k-1}) - \langle y - y^{k-1}, x^k\rangle : y \in \mathbb{R}^n\} \quad (D_k).$$

The *first interpretation* of DCA is simple : at each iteration one replaces in the primal DC program $(P_{dc})$ the second component $h$ by its affine minorization $h_k(x) := h(x^k) + \langle x - x^k, y^k\rangle$ at a neighbourhood of $x^k$ to give birth to the convex program $(P_k)$ whose the solution set is nothing but $\partial g^*(y^k)$. Likewise, the second DC component $g^*$ of the dual DC program $(D_{dc})$ is replaced by its affine minorization $(g^*)_k(y) := g^*(y^k) + \langle y - y^k, x^{k+1}\rangle$ at a neighbourhood of $y^k$ to obtain the convex program $(D_k)$ whose $\partial h(x^{k+1})$ is the solution set. DCA performs so a double linearization with the help of the subgradients of $h$ and $g^*$ and the DCA then yields the next scheme :

$$y^k \in \partial h(x^k); \quad x^{k+1} \in \partial g^*(y^k). \tag{2}$$

For a complete study of DC programming and DCA the reader is referred to [5, 8, 12, 13] and references therein.

## 3 DCA for solving DDM problem

Using the exact penalty in DC programming [6], Problem (DDM) can be reformulated as a DC program and solved by DCA.

The choice of initial point for DCA has an important impact on the quality of solutions computed by DCA. We suggest using DCA applied to the concave program [7]

$$0 = \min\{\sum_{i=1}^{n} \min\{x_i, 1 - x_i\} : x \in K\}.$$

with initial point given by the SemiDefinite Relaxation technique (see [4] ).

**TAB. 1.** *Numeric results for the Diversity Data Mining Problem*

| Problem | $n$ | $m$ | DCA | | Heuristic | |
|---|---|---|---|---|---|---|
| | | | Val | time (s) | Val | time (s) |
| 1∗ | 100 | 40 | 4074 | 6.04 | 4142 | 234.00 |
| 2∗ | 200 | 80 | 15769 | 40.09 | 16225 | 4487.60 |
| 3∗ | 300 | 120 | 35369 | 114.51 | 35881 | 17562.01 |
| 4∗ | 400 | 160 | 61284 | 276.09 | 62454 | 52056.20 |
| 5∗ | 500 | 200 | 95941 | 601.65 | 97320 | 138031.30 |
| 1 | 120 | 42 | 4561 | 8.02 | | |
| 2 | 150 | 55 | 7974 | 12.19 | | |
| 3 | 180 | 65 | 10631 | 30.88 | | |
| 4 | 240 | 90 | 19944 | 45.80 | | |
| 5 | 320 | 115 | 32198 | 124.19 | | |
| 6 | 360 | 140 | 47150 | 188.50 | | |
| 7 | 450 | 172 | 70634 | 424.55 | | |
| 8 | 480 | 180 | 78336 | 544.25 | | |
| 9 | 500 | 180 | 78407 | 626.42 | | |
| 10 | 560 | 205 | 100329 | 1002.69 | | |
| 11 | 600 | 220 | 114775 | 1308.14 | | |
| 12 | 630 | 220 | 116091 | 1534.86 | | |
| 13 | 700 | 240 | 136632 | 2247.59 | | |
| 14 | 720 | 235 | 131222 | 2488.36 | | |
| 15 | 800 | 275 | 178269 | 3681.81 | | |
| 16 | 850 | 275 | 178361 | 4657.91 | | |
| 17 | 900 | 325 | 250012 | 5787.52 | | |
| 18 | 910 | 310 | 226504 | 6129.28 | | |
| 19 | 970 | 340 | 272687 | 7916.39 | | |
| 20 | 1000 | 370 | 322727 | 8665.41 | | |

## 4 Numerical experiences and Conclusions

We report the numerical results of our algorithm in the table 1. The table is divided into two parts. In the first part, the data is given in [1]. We also generated 20 instances of test problems with large dimension and the result is provided in the second part.

¿From the numerical results, we observe that, the values given by DCA are very good and are quite smaller than the ones given by heuristic methods. However the running time of DCA is much smaller. It is worth noticing that, most of its running time is for solving the SemiDefinite Program (see [4]), the time consumed by DCA is very small. We can also remark that, DCA works well when the dimension of problem is very large for which other methods can not work.

**Conclusions.** We have proposed in this paper a deterministic method based on DC programming for the diversity data mining problem which have many applications in Data Mining. Computational experiments show the efficiency of our algorithm. The algorithm needs to be studied more carefully in the goal of finding global solutions which are often given by DCA. It should be noticed that, **with DCA, we have not the limit in their ability to solve problems of practical size**. Moreover the strategy to find a good initial point for DCA also needs to be studied. These issues are currently in progress.

## References


[1] GEIZA C., LUIZ S., SIMONE L., Experimental Comparision of Greedy Randomized Adaptive Search Procedures for the Maximum Diversity Problem, *to appear in Journal of Heuristic Springer*, , 2006.

[2] GHOSH J., Computational aspects of the maximum diversity problem, *Operations research letter*, 19, 1996, 175–181.

[3] GLOVER F., HERSH G., MCMILLAN C., Selecting subsets of maximum diversity, 77-9, 1977, MSIS Report, University of colorado at Boulder.

[4] HENRY W., ROMESH S., LIEVEN V., *Handbook of SemiDefinite Programming - Theory, Algorithms, and Application*, Kluwer Academic Publisher, 2000.

[5] HOAI AN L. T., TAO P., Solving a class of linearly constrained indefinite quadratic problems by DC algorithms, *Journal of Global Optimization*, 11, 3, 1997, 253–285.

[6] HOAI AN L. T., TAO P., MUU L. D., Exact Penalty in DC Programming, *Vietnam Journal of Mathematics*, 27, 2, 1999, 169–178.

[7] HOAI AN L. T., TAO P., A continuous Approach for Globally Solving Linearly Constrained Quadratic Zero-One Programming Problems, *Optimization*, 50, 2001, 93–120.

[8] HOAI AN L. T., TAO P., The DC (difference of convex functions) Programming and DCA revisited with DC models of real world nonconvex optimization problems, *Annals of Operations Research*, 133, 2005, 23–46.

[9] KATAYAMA K., NARIBISA H., An evolutionary approach for the maximum diversity problem, , 2003, Department of Information and Computer Engineering, Okayama Univesity of Science.

[10] KOCHENBERGER G., GLOVER F., Diversity Datamining, , 1999, University of Mississipi.

[11] KUO C., GLOVER F., DHIR K., Analyzing and Modeling the Maximum Diversity Problem by Zero-One Prgramming, *Decision Sciences*, 24, 1993, 1171–1185.

[12] TAO P., HOAI AN L. T., Convex analysis approach to DC programming : Theory, Algorithms and Applications., *Acta Mathematica Vietnamica, dedicated to Professor Hoang Tuy on the occasion of his 70th birthday*, 22, 1, 1997, 289–355.

[13] TAO P., HOAI AN L. T., DC optimization algorithms for solving the trust region subproblem, *SIAM J. Optimization*, 8, 1998, 476–505.

# Mesure ordinale du caractère arboré d'une dissimilarité

**Christophe Osswald**

*$E^3I^2$ – ENSIETA*
*2 rue François Verny, 29806 Brest Cedex 9*
*Christophe.Osswald@ensieta.fr*

RÉSUMÉ. *Ces travaux exposent le principe d'une mesure d'écart au modèle arboré, et présentent les résultats obtenus sur des dissimilarités aléatoires courantes ainsi que des jeux de données courants. Cette mesure utilise le nombre d'arêtes des graphes de rigidité obtenus sur les restrictions à quatre éléments d'une dissimilarité donnée. Nous l'évaluons sur des données aléatoires de plusieurs types, ainsi que sur trois jeux de données : les données de Henley, réputées résistantes ; des données biologiques, réputées arborées, et des données sonar, dont les paramètres évoluent dans un espace vectoriel.*

MOTS-CLÉS : *Dissimilarités, réalisations de dissimilarités, graphes de rigidité, mesures de structuration*

## 1. Introduction

Lors d'un traitement de données par un processus de classification non-supervisée, il est courant de construire une dissimilarité qui approche au mieux les relations entre les éléments à classer, tout en garantissant que la dissimilarité produite sera plus aisée à interpréter que la mesure originale. La qualité de l'approximation est alors mesurée par une métrique.

Toutefois, l'information portée par la dissimilarité d'origine n'est pas toujours de nature numérique, notamment s'il s'agit de nombres donnés par un opérateur humain, et effectuer un calcul sur celle-ci peut amener à sur-interpréter les données. Ainsi dire que $d(x, y) = 3$ et $d(x, z) = 5$ traduit plutôt le fait que $x$ est plus proche de $y$ que de $z$ qu'une différence de deux unités entre la distance de $x$ à $y$ et de $x$ à $z$.

Dans le cadre de cette stricte notion ordinale, il est notamment possible de construire des systèmes de classes à partir des données, d'en évaluer l'homogénéité, ou de rechercher des structures permettant de les expliquer : les graphes de rigidité [FLA 79][BRU 03b].

Nous nous attachons ici à extraire une information globale sur la capacité d'une dissimilarité à ranger les éléments qu'elle mesure, *via* le comportement local des graphes de rigidité des restrictions de la dissimilarité. Nous créons ainsi deux mesures $\mu_R$ et $\mu_K$, telles qu'une dissimilarité arboricole, dont les classes vont s'attacher à un arbre, vont atteindre une mesure minimale de 0, alors que des dissimilarités non ordonnées vont atteindre une mesure de 1.

## 2. Construction d'un système de classes à partir d'une dissimilarité

Une façon naturelle [JAR 71] de construire une classe d'une dissimilarité $d$ à partir de $x$ et $y$, est de chercher les *cliques maximales* de diamètre $d(x, y)$ contenant $x$ et $y$. L'ensemble de ces cliques maximales (des graphes-seuils de $d$) forment le système de classes associé à $d$ : $\mathcal{K}_d$. Si cette façon de procéder est naturelle, un graphe à $n$ sommets peut avoir $\mathcal{O}(3^{\frac{n}{3}})$ cliques maximales différentes.

Il y a d'autres manières de construire un système de classes à partir d'une dissimilarité :

– boules $B_d(x, d(x, y)) = \{z \mid d(x, z) \leqslant d(x, y)\}$ : $n(n-2)$ classes, calculées en $\mathcal{O}(n^3)$ opérations.
– 2-boules $B_d(x, y) = \{z \mid \max(d(x, z), d(y, z)) \leqslant d(x, y)\}$ : $\frac{n(n-1)}{2}$ classes, calculées en $\mathcal{O}(n^3)$ opérations.
– réalisations $\delta_d(x, y)$ : $\frac{n(n-1)}{2}$ classes, calculées en $\mathcal{O}(n^4)$ opérations [BRU 03a].

$$z \in \delta_d(x, y) \iff z \in \cap\{C \in \mathcal{K} \mid x \in C, y \in C\}$$

## 3. Graphes de rigidité

Une démarche courante en classification non-supervisée consiste à approcher une dissimilarité donnée par une dissimilarité d'un type choisi – ultramétriques, dissimilarités de Robinson, quasi-ultramétriques, distances d'arbre, etc. La mesure de la qualité de l'approximation se fait par une évaluation de la distance entre la dissimilarité d'origine $\delta$ et la dissimilarité approchée $\delta^*$, par exemple par une mesure en norme $L_2$ :

$$d^2(\delta, \delta^*) = \sum_{x,y} (\delta(x, y) - \delta^*(x, y))^2.$$

Les graphes de rigidité fournissent une structure sur laquelle accrocher les classes d'une dissimilarité, afin d'en permettre une meilleure interprétation. Ainsi, si $\mathcal{H}_d$ est un système de classes issu de la dissimilarité $d$, un graphe de rigidité associé $G_{\mathcal{H}_d}$ sera tel que sa restriction à toute classe de $\mathcal{H}_d$ est connexe.

Un graphe $G = (X, E)$ compatible avec une dissimilarité $d$ sur $X$ est un graphe de rigidité pour l'ensemble $\mathcal{K}_d$ de ses classes naturelles. Il a la propriété que pour tous éléments $x$ et $y$, $d(x, y)$ est plus grand que $d(u, v)$ pour $u$ et $v$ sur un chemin de $x$ à $y$, pour au moins un chemin de $x$ à $y$ dans $G$. Comme une classe naturelle [JAR 71] d'une dissimilarité $d$ est une clique maximale d'un graphe-seuil de $d$ ($G_\lambda = (X, E_\lambda)$ avec $E_\lambda = \{(u, v) \in X^2 \mid d(u, v) \leqslant \lambda\}$), le graphe $G$ restreint à toute classe naturelle de $d$ is connexe. Nous choisissons des graphes minimaux en terme de nombre d'arêtes pour cette propriété, afin d'obtenir une structure aussi simple que possible : $G_d$, un graphe de rigidité minimum de la dissimilarité $d$.

La figure suivante montre comment construire un graphe de rigidité minimum à partir d'une dissimilarité exemple. Les classes naturelles de $d$ sont $xy$ et $zt$ (diamètre 1), $xz$, $yt$ et $zut$ (diamètre 2) et $xyztu$ (diamètre 3). La dissimilarité $d$ admet deux graphes de rigidité minimum :

| $d$ | $x$ | $y$ | $z$ | $t$ | $u$ |
|---|---|---|---|---|---|
| $x$ | 0 | 1 | 2 | 3 | 3 |
| $y$ | 1 | 0 | 3 | 2 | 3 |
| $z$ | 2 | 3 | 0 | 1 | 2 |
| $t$ | 3 | 2 | 1 | 0 | 2 |
| $u$ | 3 | 3 | 2 | 2 | 0 |

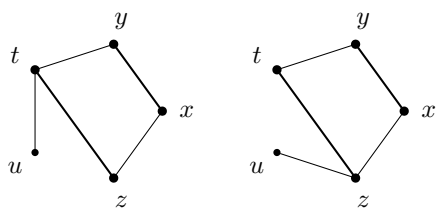



FIG. 1. *Graphes de rigidité minimum des classes naturelles de $d$*

Si $d$ est une dissimilarité arboricole sur $X$, elle admet un graphe de rigidité qui est un arbre. Il est équivalent que l'ensemble $\mathcal{B}_d$, $2\mathcal{B}_d$, $\mathcal{K}_d$ ou $\mathcal{R}_d$ soit rigide avec un arbre. Il s'agit de la stucture la plus simple que l'on puisse trouver, $X$ étant toujours une classe.

## 4. Mesure construite sur quatre éléments

Les dissimilarités sur un ensemble de quatre éléments admettent 720 (6 !) ordres totaux différents pour leurs valeurs de dissimilarités deux à deux. A l'ordre des sommets près, ces ordres totaux se traduisent par sept emboîtements de graphes-seuil possibles, représentés en figure 2. Pour chacun de ces cas, la taille (en nombre d'arêtes) d'un graphe de rigidité minimum est donnée, pour l'ensemble des boules, des 2-boules, des classes naturelles ou des réalisations de la dissimilarité.

Considérer les cas dans lesquels deux paires $(x, y)$ et $(z, t)$ peuvent avoir une même dissimilarité obligerait à passer de l'ensemble des ordres totaux à l'ensembles des ordres faibles sur six élements, ce qui augmente largement le nombre de cas à considérer (4683 au lieu de 720) et obligerait à définir une loi de probabilité sur ces éléments, pour laquelle le modèle uniforme ne s'impose guère.

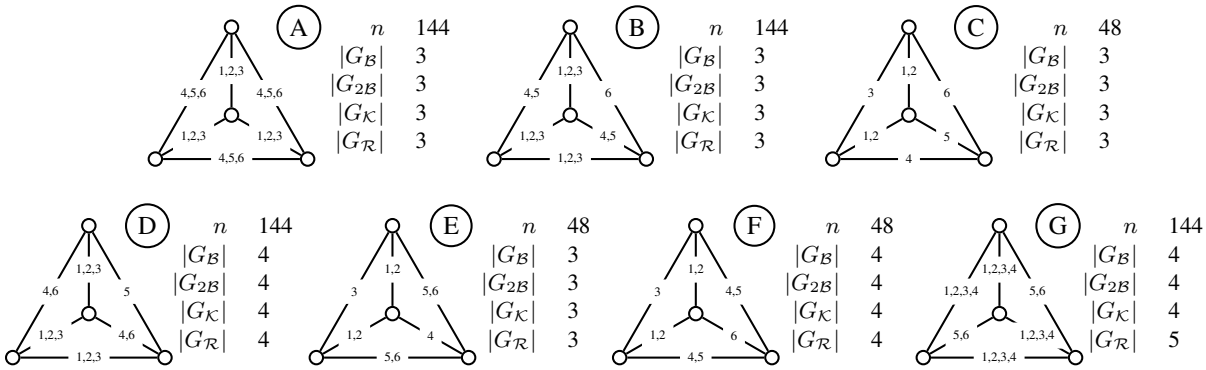



FIG. 2. *Nombre d'arêtes des graphes de rigidité minimum des dissimilarités à quatre éléments*

Les cas A et E ont un graphe de rigidité qui est une étoile. B et C sont des dissimilarités de Robinson. G est une dissimilarité circulaire qui n'est pas une dissimilarité de Robinson ; ce cas peut être considéré comme l'écart le plus important au modèle arboré. A l'exception du cas D, le graphe de rigidité obtenu est unique.

Sur les ensembles à quatre éléments, les ensembles $\mathcal{B}_d$, $2\mathcal{B}_d$ et $\mathcal{K}_d$ sont confondus. A partir de cinq éléments, ces ensembles diffèrent. De façon générale, on a :

$$|G_{\mathcal{B}}| \leqslant |G_{2\mathcal{B}}| \leqslant |G_{\mathcal{K}}| \leqslant |G_{\mathcal{R}}|$$

Si $d$ est une dissimilarité aléatoire sur $X$ un ensemble à quatre éléments dont les six valeurs suivent une même loi continue, les 720 ordres possibles sont équiprobables. Ainsi un graphe de rigidité minimum des classes naturelles de cette dissimilarité aura en moyenne $3 + \frac{7}{15}$ arêtes ; un graphe de rigidité minimum des réalisations de cette dissimilarité en aura en moyenne $3 + \frac{2}{3}$.

## 5. Mesures de structuration pour certaines dissimilarités

Ainsi, nous pouvons analyser les restrictions à quatre éléments d'une dissimilarité $d$ sur $X$ pour déterminer la proximité de celle-ci avec un modèle arboré. Soit $|G^4_{\mathcal{R}_d}|$ le nombre d'arêtes d'un graphe de rigidité minimum de la réalisation de la restriction de $d$ à quatre éléments, et $|G^4_{\mathcal{K}_d}|$ le nombre d'arêtes d'un graphe de rigidité minimum de ses classes naturelles. Nous avons donc deux indices $\mu_K$ et $\mu_R$ issus respectivement des classes naturelles et des réalisations, qui valent 0 pour une dissimilarité arborée, et 1 pour une dissimilarité aléatoire :

$$\mu_K = \frac{15}{7}(E(|G^4_{\mathcal{K}_d}|) - 3) \qquad \mu_R = \frac{3}{2}(E(|G^4_{\mathcal{R}_d}|) - 3)$$

Une mesure globale peut être obtenue à partir d'un graphe de rigidité minimum pour la dissimilarité prise sur $X$ [OSS 04]. Toutefois cette approche est limitée aux réalisations de la dissimilarité – la recherche d'un graphe de rigidité minimum est NP-difficile dans les autres cas [BRU 03b].

Si $x$, $y$, $z$ et $t$ sont quatre points de $\mathbb{R}^2$, dont les coordonnées sont des variables aléatoires uniformes sur [0,1] : $\mu_K = 0.52$ et $\mu_R = 0.43$ pour une norme $L_2$ et $\mu_K = 0.51$ et $\mu_R = 0.47$ pour une norme $L_1$ . Si $x$, $y$, $z$ et $t$

sont quatre points de $\mathbb{R}^3$, dont les coordonnées sont des variables aléatoires uniformes sur [0,1] : $\mu_K = 0.65$ et $\mu_R = 0.55$ pour une norme $L_2$.

Si $x$ $y$, $z$ et $t$ sont quatre points d'un cercle, et que la distance considérée est la longeur d'un arc reliant les deux points, nous obtenons $\mu_K = \frac{5}{7}$ et $\mu_R = \frac{2}{3}$. Dans ce cas, les configurations A et E sont impossibles, et la configuration G, typique des dissimilarités circulaires, a une probabilité de $\frac{1}{9}$.

Les données de Henley [HEN 69] sont issues d'une expérience de rappel libre. Il a été demandé aux sujets d'écrire tous les animaux auxquels ils pouvaient penser en un temps donné. La dissimilarité globale entre deux animaux est la distance moyenne qui les sépare dans ces listes. Ces données ont été abondamment étudiées, par exemple par Barthélemy et Guénoche [BAR 88] et sont réputées résistantes aux méthodes usuelles de classication. Ces données se caractérisent par $\mu_K = 0.81$ et $\mu_R = 0.66$, assez peu éloignés des paramètres de données aléatoires.

Les distances entre codes génétiques de diverses espèces animales sont considérées comme suivant un modèle arboré. La dissimilarité obtenue à partir de l'ADN de gamma-globine sur 36 singes [PAG 99] nous amène $\mu_K = 0.415$ et $\mu_R = 0.297$. Les configurations de type G sont particulièrement rares (0.4% du total).

Une base de données de 26 images sonar (source : GESMA, obtenues sur un Klein 5400 latéral) est découpée en imagettes représentant chacune un peu plus de 1m$^2$ de fond marin. De chaque imagette sont extraites quatre caractéristiques, devant servir à discriminer les types de fonds marins. Les paramètres obtenus sont $\mu_K = 0.40$ et $\mu_R = 0.30$, alors qu'une distance euclidienne sur des points uniformément répartis sur un cube de $\mathbb{R}^4$ a des paramètres de $\mu_K = 0.70$ et $\mu_R = 0.61$. Ici, l'algorithme en $\mathcal{O}(n^4)$ qui parcourt tous les quadruplets des données a été abandonné au profit d'un échantillon représentatif d'un million d'éléments.

## 6. Conclusion

Les deux mesures proposées, $\mu_K$ et $\mu_R$, se rejoignent pour les dissimilarités arboricoles et les dissimilarités aléatoires. Toutefois, sur les jeux de données réelles et les distances aléatoires, nous avons toujours $\mu_K > \mu_R$ : le cas G, typique des dissimilarités circulaires qui ne sont pas de Robinson, est rendu plus rare pour les distances que pour les dissimilarités, en raison de l'application de l'inégalité triangulaire. C'est uniquement sur ces quadruplets que se construit la différence entre le nombre d'arêtes d'un graphe de rigidité minimum des réalisations ou des classes naturelles. Il est donc préférable d'utiliser $\mu_K$ pour obtenir une mesure du caractère arboré d'une dissimilarité, bien que cela empêche d'articuler le résultat obtenu avec une version globale de cette même mesure, le problème associé étant NP-difficile.

## 7. Bibliographie


[BAR 88] BARTHELÉMY J.-P., GUÉNOCHE A., *Les arbres et les représentations de proximité*, Masson, Paris, 1988.

[BRU 03a] BRUCKER F., Réalisations de dissimilarités, *Rencontres de la SFC*, 2003, p. 7-10.

[BRU 03b] BRUCKER F., OSSWALD C., BARTHÉLEMY J.-P., Rigid hypergraphs : combinatorial optimization problem in clust ering and similarity analysis, *Actes de INOC 2003*, 2003.

[FLA 79] FLAMENT C., DEGENNE A., VERGÈS P., Analyse de similitude ordinale, *Informatique et Sciences Humaines*, vol. 40-41, 1979, p. 223-231.

[HEN 69] HENLEY N. M., A Psychological Study of the Semantics of Animal Terms, *Journal of Verbal Learning and Verbal Behavior*, vol. 8, 1969, p. 176-184.

[JAR 71] JARDINE N., SIBSON R., *Mathematical Taxonomy*, Wiley, London, 1971, part II.

[OSS 04] OSSWALD C., Mesure de structuration d'un système de classes, *Comptes-rendus des 11e rencontres de la SFC*, Bordeaux, France, 2004, p. 261-264.

[PAG 99] PAGE S. L., CHIU C. H., GOODMAN M., Molecular phylogeny of old world monkeys (Cercopithecidae) as inferred from gamma-globin DNA sequences, *Molecular Phylogenetics and Evolution*, vol. 13, 1999, p. 348-359.

# Méthodes de classification pour l'extraction de règles

**Marie Plasse* **, Ndeye Niang*, Gilbert Saporta*,**
**Alexandre Villeminot**, Laurent Leblond****

** CNAM Laboratoire CEDRIC*
*292 Rue St Martin Case 441*
*75141 Paris Cedex 03, France*

*** PSA Peugeot Citroën*
*45 rue Jean-Pierre Timbaud*
*78307 Poissy Cedex, France*

*RÉSUMÉ. Cette communication présente une comparaison de la classification de variables et de la classification croisée utilisées préalablement à la recherche de règles d'association sur un jeu de données industrielles.*

*MOTS-CLÉS : Classification de variables, classification croisée, règles d'association, événements rares*

## 1 Introduction

Les données que nous analysons sont issues de la fabrication automobile où plusieurs dizaines de milliers de véhicules sont décrits par plusieurs milliers de variables binaires rares. La recherche de règles d'association entre ces attributs conduit à une profusion de règles. Une classification des variables préalable nous a permis de regrouper les attributs, puis d'orienter la recherche de règles afin d'en diminuer le nombre. Dans cette communication nous étudions, sur un petit échantillon de nos données, les apports d'une classification croisée à cette approche. Après avoir présenté quelques éléments théoriques sur les méthodes utilisées, nous comparons les règles obtenues après classification de variables d'une part et classification croisée d'autre part.

## 2 Quelques éléments théoriques sur les méthodes utilisées

### *2.1 La recherche de règles d'association*

La méthode de recherche de règles d'association est née pour analyser les articles fréquemment achetés ensemble dans les supermarchés. Chaque sortie de caisse correspond à une transaction où plusieurs items ont été achetés simultanément. Une règle d'association est une implication $A \rightarrow C$ où l'antécédent $A$ et le conséquent $C$ sont des ensembles d'items, où $A \cap C = \varnothing$. Une règle repose sur les notions de support et de confiance. Le support est le nombre ou le pourcentage de transactions qui contiennent les items de la règle. La confiance est le pourcentage de transactions qui contiennent les items du conséquent parmi celles qui contiennent l'antécédent. Les algorithmes de recherche de règles d'association, tel que l'algorithme fondateur *Apriori* [AGR 94], procèdent en deux étapes. La première est la recherche des ensembles d'items fréquents dont le support est supérieur à un seuil fixé par l'utilisateur. A partir de ces ensembles, la

Avec tous nos remerciements à M. Nadif pour son aide précieuse sur la classification croisée.

seconde étape est l'extraction des règles dont la confiance est jugée suffisante par l'utilisateur. Le nombre de règles extraites étant souvent important, pour sélectionner les plus intéressantes, il est utile de les classer par ordre décroissant de leur intérêt statistique au sens d'un indice de pertinence. De nombreux indices ont été proposés tels que le lift $\left(P(A\cap C)/P(A).P(C)\right)$ [BRI 97] qui est facilement interprétable. Le choix d'un indice plutôt qu'un autre dépend du contexte ; aussi, dans le cadre de notre application, l'indice de Jaccard $\left(P(A\cap C)/P(A\cup C)\right)$ discrimine le mieux les règles qui nous intéressent [PLA 06]. Il nous est donc possible de sélectionner les règles les plus pertinentes grâce à cet indice.

### *2.2 Classification de variables*

Comme pour la classification d'individus, il existe deux grandes familles de méthodes de classification de variables : des méthodes de partitionnement (telles que *Varcha* [VIG 03]) et des méthodes hiérarchiques. Dans cette seconde famille, la méthode descendante (procédure *Varclus* de SAS) recherche des classes unidimensionnelles décrites par une seule composante principale et les méthodes ascendantes conduisent à une hiérarchie de partitions emboîtées de l'ensemble des variables. Ces dernières reposent sur le choix d'une stratégie d'agrégation et d'un indice de similarité entre les variables. Le $\Phi^2$ de Pearson, l'indice de Jaccard ou encore celui de Russel-Rao sont des indices adaptés au cas binaire. L'utilisation conjointe de la classification de variables et des règles d'association [PLA 05] permet de faire face à la profusion de règles obtenue avec une recherche classique des règles. La classification de variables permet de construire des classes homogènes d'attributs. La recherche de règles d'association à l'intérieur de chacune de ces classes est pertinente car il est facilement possible d'identifier les classes où les attributs sont très corrélés et produisent donc de nombreuses règles. L'ensemble d'associations, plus restreint, est plus simple à analyser.

### *2.3 Classification croisée*

L'objectif de la classification croisée est de trouver une paire de partitions (**z**,**w**), où **z** est une partition de l'ensemble *I* des *n* individus en *K* classes et **w** est une partition de l'ensemble *J* des *m* variables en *H* classes, *K* et *H* étant connus. Ce problème est résolu de manière itérative par une optimisation alternée de la partition des individus en bloquant celle des variables puis de la partition des variables en fixant celle des individus. Plusieurs algorithmes ont été proposés selon le type de données, dont l'algorithme *Crobin* dans le cas binaire [GOV 83] qui propose de maximiser un critère de type inertie. Cet algorithme est rapide et donne de bons résultats lorsque les blocs ont les mêmes proportions et des degrés d'homogénéité semblables. Lorsque ce n'est pas le cas, le problème de la classification croisée peut être traité par l'approche modèle de mélange, où les données sont supposées provenir d'un mélange de plusieurs distributions de probabilité, où chaque composant du mélange correspond à une classe ([GOV 03], [GOV 05]). Le problème consiste alors à retrouver pour chaque objet sa population d'origine la plus probable en fonction du vecteur d'observations qui le caractérise. Les données observées **x** enrichies par les informations manquantes (ici les classes) constituent les données complètes. Ainsi, les données manquantes sont, d'une part le vecteur $\mathbf{z}=(\mathbf{z}_1,...,\mathbf{z}_i,...,\mathbf{z}_n)$ où $\mathbf{z}_i=k$ (avec $k=1...K$) est le numéro *k* de la classe de l'individu *i*, et le vecteur $\mathbf{w}=(\mathbf{w}_1,...,\mathbf{w}_j,...,\mathbf{w}_m)$ où $\mathbf{w}_j=h$ (avec $h=1...H$) est le numéro *h* de la classe de la variable *j*. Le modèle de mélange croisé s'écrit $f(\mathrm{x};\theta)=\sum_{(z,w)\in Z\times W}\prod p_{z_i}\prod q_{w_j}\prod\varphi_{z_iw_j}(x_i^j;\alpha_{z_i}^{w_j})$ où les densités $\varphi_{kh}$ appartiennent à la même famille, les paramètres $p_k$ et $q_h$ sont les probabilités qu'une ligne et une colonne appartiennent respectivement aux $k^{ème}$ et $h^{ème}$ composants du mélange. L'estimation du vecteur $\theta$ des paramètres $(p_1,...,p_K,q_1,...,q_H,\alpha_{11},...,\alpha_{KH})$ de ce modèle est réalisée par la méthode du maximum de vraisemblance grâce à des extensions de l'algorithme *Estimation-Maximisation*. Ainsi, l'algorithme *Bloc-CEM* [GOV 03] propose de maximiser la log-vraisemblance des données complètes. Cette approche fournit des résultats rapidement mais présente certains inconvénients, elle conduit notamment à une estimation biaisée. Plus lent mais plus fiable, l'algorithme *Bloc-EM* [GOV 05] permet de maximiser l'espérance de la log-

vraisemblance des données complètes, conditionnellement aux données observées **x** et à l'estimation courante de $\theta$. Dans le cas des données binaires, la distribution de probabilités est la distribution de Bernoulli $\varphi_{kh}\left(x_i^j;\alpha_k^h\right)=\left(\alpha_k^h\right)^{x_i^j}\left(1-\alpha_k^h\right)^{\left(1-x_i^j\right)}$. Après initialisation, une première étape, où la partition sur les colonnes est fixée, est constituée d'une phase Estimation où sont calculées les probabilités a posteriori qu'un individu *i* appartiennent à une classe *k*. Vient ensuite la phase Maximisation où sont déduites les proportions $p_k$ des composants du mélange et les probabilités $\alpha_k^h$ de prendre la valeur "1" dans le bloc (*k,h*). Une seconde étape, où la partition en ligne est bloquée, estime les probabilités a posteriori qu'une variable *j* soit dans la classe *h*. La phase de maximisation attribue ensuite les proportions $q_h$ de chaque classe *h* ainsi que de nouvelles probabilités $\alpha_k^h$. Ces deux étapes sont répétées jusqu'à la convergence. La recherche de règles d'association dans des blocs homogènes où la plupart des véhicules présentent les mêmes attributs permet en outre de diminuer l'espace de recherche. En effet, les blocs de "0" sont ignorés et l'interprétation des blocs entiers de "1" est triviale et elle ne nécessite pas d'effectuer une recherche d'associations.

## 3 Classification de variables vs classification croisée sur une application

### *3.1 Recherche de règles d'association sans classification préalable*

Les données constituent un échantillon de 727 véhicules décrits par la présence ou l'absence de 109 attributs. La matrice de données binaires est clairsemée puisqu'elle ne comprend que 2,9% de "1". Un véhicule possède en moyenne 3,2 attributs et l'attribut le plus fréquent apparaît sur environ 10% des véhicules mais 80% des attributs apparaissent sur moins de 1% des véhicules.
Les premières règles trouvées ont un support de 50 véhicules mais elles ont des confiances faibles. Les attributs étant rares, il est préférable de fixer un seuil très bas pour le support et d'être plus sévère au niveau de la confiance. De plus, pour sélectionner le plus de règles pertinentes, nous pouvons fixer un seuil minimum pour l'indice de Jaccard. Avec un tel paramétrage nous espérons obtenir des règles fiables sur des événements rares. Le Tableau 1 montre les résultats obtenus avec une confiance de 90% et des seuils différents pour le support et l'indice de Jaccard. Le nombre de règles à analyser est trop important.

| Support minimum | Confiance minimum | Jaccard minimum | Nombre d'ensembles fréquents | Nombre de Règles |
|---|---|---|---|---|
| 30 véhicules | 90% | 0 | 1 230 | 39 867 |
| 30 véhicules | 90% | 0,9 | 1 230 | 21 254 |
| 10 véhicules | 90% | 0 | 65 583 | 26 210 753 |
| 10 véhicules | 90% | 0,6 | 65 583 | 11 839 141 |
| 10 véhicules | 90% | 0,9 | 65 583 | 10 127 600 |

**Tableau 1 : Règles obtenues sans classification préalable**

### *3.2 Classification de variables préalable à la recherche de règles*

Le dendrogramme résultant d'une classification ascendante hiérarchique avec la stratégie de Ward et l'indice de Russel-Rao préconise une partition des variables en 2 ou 5 classes. Nous présentons les résultats obtenus sur deux classes, sachant que la première classe des partitions en 2 et 5 classes est identique et que les autres classes sont regroupées en une seule. Les résultats de la recherche de règles d'association à l'intérieur des deux classes avec un support minimum de 10 véhicules et une confiance minimum de 90% sont présentés dans le Tableau 2.

| Classe | Nombre de variables | Pourcentage de "1" | Nombre d'ensembles fréquents | Nombre de règles | | |
|---|---|---|---|---|---|---|
| | | | | *Jaccard ≥ 0,9* | *Jaccard ≥ 0,6* | *Jaccard ≥ 0* |
| 1 | 16 | 13 | 65535 | 10160318 | 11839140 | 26210797 |
| 2 | 93 | 1 | 36 | 0 | 1 | 2 *(jac ≥ 0,55)* |

**Tableau 2 : Composition des classes et règles produites**

Les règles de la classe 1 concernent 11 attributs très corrélés : il sont présents simultanément sur 16 véhicules, ce qui explique le nombre élevé de règles. Les deux règles isolées dans la classe 2 sont assez

intéressantes du point de vue de l'indice de Jaccard. La classification préalable permet de découvrir des associations sur des items plus rares. En effet, sans classification, les items les plus fréquents créent une profusion des règles qui noie les résultats et empêche de voir les associations intéressantes.

### *3.3 Classification croisée préalable à la recherche de règles*

Une classification hiérarchique sur les individus permet d'avoir une idée du nombre de classes à fixer en ligne. Le Tableau 3 montre les résultats obtenus avec 3 classes en ligne, 2 classes en colonne et le même paramétrage qu'avec la classification de variables (le bloc 6 ne figure pas car il ne contient que des "0").

| Bloc | Pourcentage de "1" | Nombre d'individus | Nombre de variables | Nombre d'ensembles fréquents | Nombre de règles | | |
|---|---|---|---|---|---|---|---|
| | | | | | *Jaccard ≥ 0,9* | *Jaccard ≥ 0,6* | *Jaccard ≥ 0* |
| 1 | 8,8 | 682 | 15 | 29 | 0 | 0 | 0 |
| 2 | 1,2 | 682 | 94 | 30 | 0 | 0 | 1 *(jac=0,55)* |
| 3 | 100 | 29 | 15 | 32767 | 14283372 | 142283372 | 142283372 |
| 4 | 2,86 | 29 | 94 | 1 | 0 | 0 | 0 |
| 5 | 48,3 | 16 | 15 | 63 | 602 | 602 | 602 |

**Tableau 3 : Composition des blocs et règles produites**

Le bloc 3 est intégralement constitué de "1" : les 15 attributs sont présents simultanément sur les 29 véhicules du bloc. Les 14 millions de règles issues de ce bloc sont porteuses d'une seule et même information. Dans le bloc 5, la plupart des règles sont provoquées par la présence de 6 attributs sur 13 des 16 véhicules. Enfin, la règle du bloc 2 avait été détectée grâce à la classification de variables également.

## 4 Conclusion

Les deux approches, simple et croisée, conduisent à une réduction du nombre de règles, une fois les groupes analysés. Elles permettent d'identifier puis d'isoler les groupes d'attributs fortement liés, et enfin d'orienter la recherche de règles vers des groupes moins homogènes où des associations moins évidentes seront découvertes. La classification croisée fournit une partition des données plus fine et intéressante. De plus, elle présente l'avantage de pouvoir exclure les blocs totalement homogènes de la recherche de règles, ce qui peut se révéler très utile sur des données de taille importante. Cette approche, prometteuse sur un échantillon restreint, va donc être utilisée sur la base entière comportant plusieurs dizaines de milliers de véhicules et des milliers de variables.

## 5 Bibliographie


[AGR 94] AGRAWAL R., SRIKANT R. *Fast Algorithms for Mining Association Rules*. Proceedings of the 20th Conference on Very Large Databases, Santiago, Chile, pp. 487-499, 1994.

[BRI 97] BRIN S., MOTWANI R., SILVERSEIN C. *Beyond market baskets: generalizing association rules to correlations*. Proceedings of the ACM-SIGMOD Conference on Management of Data, Tucson, Arizona, USA, 1997.

[GOV 83] GOVAERT G. Classification croisée, Thèse d'Etat, Université Paris 6, France, Juin 1983.

[GOV 03] GOVAERT G., NADIF M. Clustering with block mixture models. Pattern Recognation, 36(2) : pp. 463-473, 2003.

[GOV 05] GOVAERT G., NADIF M. *An EM Algorithm for the Block Mixture Model.* IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 643-647, 2005.

[PLA 05] PLASSE M., NIANG N., SAPORTA G. Utilisation conjointe des règles d'association et de la classification de variables. Journées Françaises de Statistique, Pau, France, 2005.

[PLA 06] PLASSE M., NIANG N., SAPORTA G., LEBLOND L. *Une comparaison de certains indices de pertinence des règles d'association.* Revue des Nouvelles Technologies de l'Information, Actes 6e Conférence Extraction et Gestion des Connaissances, EGC'06, Série E, n°6, Vol.II, pp.561-568, Lille, janvier 2006.

[VIG 03] VIGNEAU E., QANNARI E.M. *Clustering of variables around latent component - application to sensory analysis*. Communications in Statistics , Simulation and Computation, 32(4) : pp. 1131-1150, 2003.

# Définition et illustration du Mélange Tabulaire Gaussien

***Discrétisation probabiliste pour l'analyse exploratoire des données***

**Rodolphe Priam, Mohamed Nadif, Francois-Xavier Jollois**

*INRETS, 2 avenue du Général Malleret-Joinville, 94114 Arcueil*
*LITA, Université de Metz Ile du Saulcy, 57045 Metz*
*CRIP5, Université Paris Descartes 45 rue des Saints-Pères, 75006 Paris*

RÉSUMÉ. *La visualisation des ensembles de données de grande taille fait appel à de nouvelles méthodes afin de révéler rapidement et efficacement leur contenu. La projection point à point sur le plan d'un nuage de données est un paradigme intéressant malgré la difficulté d'exploration du résultat lorsque le nombre de points affichés est important. C'est pourquoi nous étudions une nouvelle manière de montrer une projection bidimensionnelle à partir d'un nuage multidimensionnel : notre modèle génératif construit une vue tabulaire d'un nuage projeté. La méthode révèle le contenu des zones de forte densité par leur discrétisation non équidistribuée. Cette approche est une alternative à une carte auto-organisatrice lorsqu'une projection alternative existe. La visualisation pixellisée résultante est illustrée en projetant un échantillon d'images réelles : il devient possible d'observer comment sont disposés sur le plan les labels de classe ou bien les fréquences d'un ensemble de modalités sans modification de la carte mentale en raison d'un changement de l'agrandissement du zoom par exemple. La conclusion donne les perspectives de notre point de vue original qui aboutit à une représentation bien lisible d'une projection pour une analyse des données des grands échantillons de données.*

MOTS-CLÉS : *Analyse en Composantes Principales, Discrétisation, Modèle de mélange gaussien, Algorithme EM généralisé.*

## 1. Introduction

Les méthodes de projection non linéaires $\mathcal{M}$ telles que [SAM 69] sont proposées dans la littérature afin de projeter un échantillon de données $\{x_i\}_{i=1}^{i=I}$ ; ces méthodes rappellent que le paradigme de la projection $\{y_i = \mathcal{M}(x_i)\}_{i=1}^{i=I}$ d'un nuage de données est une bonne manière de visualiser la structure de la distribution d'un ensemble de données. Une alternative à ce point de vue est la méthode des cartes de Kohonen ou cartes auto-organisatrices (SOM) [KOH 97] qui construit une projection tabulaire d'un nuage de données, avec les cellules voisines -du tableau construit- suffisamment similaires dans l'espace des données pour préserver la structure de la distribution. Cette façon de considérer une représentation visuelle des données est certainement plus aisément lisible qu'un nuage bidimensionnel brut de projection aux nombreux points, car il devient alors difficile d'accéder à un point en particulier quand bien même la distribution projetée apparaît. Pour cette raison, certains outils de type zoom sont généralement nécessaires, mais un zoom peut entraîner une perte de la localisation sur la projection puisque la puissance de l'agrandissement est modifiée pour pallier un changement de densité locale dans l'organisation des points, ce induisant une variation dans les distances relatives entre projetés. Comme ces familles de méthodes projectives sont certainement complémentaires, et que la visualisation tabulaire est aisément compréhensible, le but de ce papier est de proposer une représentation matricielle probabiliste pour un nuage de données 2D.

Nous nous sommes intéressés à la représentation de données dans un tableau qui est organisé de la même façon que sont disposées les données sur le plan, comme proposé dans notre modèle tabulaire : un modèle de mélange de gaussiennes contraintes avec les moyennes disposées le long de lignes et colonnes libres de se déplacer afin de révéler les zones de forte densité. Notre étude permet de construire des tables rapidement lisibles à partir de corpus multimédia. Lorsque les cellules ou classes sont nombreuses comme dans le cas des bases de données massives, une approche basée sur une visualisation au niveau du pixel devient nécessaire apportant une information

pertinente et synthétique. Dans ce papier, nous présentons le modèle, nous donnons les cartes de pixels obtenues pour un ensemble d'images et nous expliquons comment la fouille de données visuelle en bénéficie ; finalement nous concluons par une discussion et des perspectives.

## 2. Le Modèle Tabulaire

Dans cette section, nous supposons qu'une méthode de type *mapping* $\mathcal{M}$ a été utilisée pour obtenir des coordonnées bidimensionnelles $y_i = (y_{is})_{s=1,2}$ qui sont traitées afin d'obtenir une vue tabulaire d'un nuage de données projetées. La base de l'algorithme est un modèle de mélange gaussien qui étend la méthode des K-means [MAC 67] à une expression probabiliste avec une variable cachée classifiante. Un mélange de lois gaussiennes s'écrit $P(y_i) = \sum_{1\leq k\leq K} \pi_k G(y_i; m_k, \sigma)$ avec $K$ facteurs, composants ou clusters, où la k-ième densité $G$ est une loi normale multidimensionnelle de moyenne $m_k$ et de paramètre de la variance sphérique $\sigma$. Le paramètre $\pi_k$ est la probabilité qu'une observation $y_i$ appartienne au k-ième composant, il correspond alors à la proportion de points dans le k-ième cluster. La log-vraisemblance des données observées $\mathcal{D}$ suppose l'échantillon $\{y_i\}_{i=1}^{i=I}$ i.i.d. à partir d'une distribution de probabilité de densité $P(y_i)$ et s'écrit $\mathcal{L}(\theta|\mathcal{D}) = \sum_{1\leq i\leq I} \log\left\{\sum_{1\leq k\leq K} \pi_k\, G(y_i; m_k, \sigma)\right\}$ où $\theta = (m_1, m_2, \cdots, m_K, \pi_1, \pi_2, \cdots, \pi_{K-1}, \sigma)$. L'inférence de ce modèle est réalisée en maximisant la log-vraisemblance dont la solution n'est pas analytiquement calculable. L'algorithme EM [DEM 77] résoud ce problème par le biais de la vraisemblance complétée $\mathcal{L}(\theta, Z|\mathcal{D}) = \prod_{1\leq i\leq I} \pi_{z_i} G(y_i; m_{z_i}, \sigma)$ par la connaissance d'une partition $\mathcal{Z}$, où $z_i$ est la variable latente dont la valeur inconnue discrète est dans l'ensemble des valeurs $\mathcal{K} = \{1, 2, \cdots, K\}$.

L'algorithme procède itérativement en deux pas, E (pour *Expectation*) et M (pour *Maximization*), en maximisant $\mathcal{Q}(\theta|\theta^{(t)})$ l'espérance conditionnelle de $\mathcal{L}(\theta, Z|\mathcal{D})$, étant donnée une estimation courante précédente $\theta^{(t)}$. Dans la suite, nous contraignons les vecteurs $m_k$ afin d'obtenir une représentation tabulaire.

### 2.1. *Le modèle*

Nous construisons un rectangle régulier avec $K_1$ colonnes et $K_2$ lignes, tel que $k$ s'écrit $k = K_1 \times (k_2 - 1) + k_1$, $(k_1 = 1, \ldots, K_1; k_2 = 1, \ldots, K_2)$, et $K = K_1 K_2$. Posons :

$$m_k = \left[\begin{array}{c} m_{k,1} \\ m_{k,2} \end{array}\right] \text{ avec } m_{k,s} = \frac{\sum_{1\leq \ell\leq k} exp(u_{\ell,s})}{\sum_{1\leq \ell\leq K_s} exp(u_{\ell,s}) + 1},$$

où $s \in \{1, 2\}$, et $u_{k,s}$ est un paramètre réel inconnu qui peut être obtenu en maximisant $\mathcal{Q}(\theta|\theta^{(t)})$. Notre paramétrisation est telle que $0 \leq m_{k,s} \leq 1$ afin d'obtenir une plus simple expression dans la suite. En remarque, la constante additive 1 dans le dénominateur de $m_{k,s}$ peut être remplacée par n'importe quel réel positif, et la somme dans le numérateur induit une direction topologique fixée pour les probabilités correspondantes : une variante contrainte de la paramétrisation classique *soft-max*. Finalement, nous normalisons l'étendue des composantes vectorielles en calculant $\tilde{y}_{is} = (y_{is} - min_i(y_{is}))/(max_i(y_{is}) - min_i(y_{is}))$ ; dans la suite, on suppose $y_{is}$ ainsi normalisé. Dans ce cas, avec $q_k^{(t)}(y_i) = q_{(k_1,k_2)}^{(t)}(y_i)$ et $\pi_k = 1/K$ -en résumé afin de placer davantage de centres dans les zones de forte densité-, nous proposons le nouveau critère sensible aux marginales :

$$\mathcal{Q}_m(\theta|\theta^{(t)}) \equiv \sum_{1\leq s\leq 2} \sum_{1\leq i\leq I} \sum_{1\leq k\leq K_s} q_{k_s}^{(t)}(y_i)(y_{is} - m_{k,s})^2,$$

On a noté $q_{k_s}^{(t)}(y_i) = q_{k,s}^{(t)}(y_i)$ la marginale de $q_{(k_1,k_2)}^{(t)}(y_i)$ sur $k_1$ si $s = 2$, et sur $k_2$ si $s = 1$ où $q_{(k_1,k_2)}^{(t)}(y_i) \propto \pi_{(k_1,k_2)}^{(t)} G(y_i; m_{(k_1,k_2)}^{(t)}, \sigma^{(t)})$ dénote une probabilité conditionnelle, étant donné $y_i$ et $\theta^{(t)}$.

Une forme analytique exacte n'existe pas pour maximiser cette quantité, donc nous employons une montée de gradient pour calculer $m^{(t+1)} = \text{argmax}_m \mathcal{Q}_m(\theta|\theta^{(t)})$. En dérivant le critère, nous obtenons le vecteur de gradient $\mathbf{Dm}_s^{(t)}$ avec pour composante de la dérivée relative à $u_{\ell,s}$ :

$$\mathbf{Dm_s}_{\ell}^{(t)} = \sum_{1\leq i\leq I} \sum_{1\leq k\leq K_s} q_{k_s}^{(t)}(y_i)(m_{k,s}^{(t)} - y_{is})(m_{\ell,s}^{(t)} - m_{\ell-1,s}^{(t)})(\delta_{\ell\leq k} - m_{k,s}^{(t)}),$$

avec $m_{0,s} = 0$, et $\delta_{\ell\leq k}$ qui vaut un si $\ell \leq k$ et zéro sinon. Finalement, on pose le pas de gradient d'estimation :

$$u_{\ell,s}^{(t+1)} = u_{\ell,s}^{(t)} - \rho^{(t)}\ \mathbf{Dm_s}_{\ell}^{(t)}$$

avec un $\rho^{(t)}$ décroissant bien choisi afin d'obtenir une solution stable pour nos paramètres $u_{\ell,s}$. Un pas de Newton-Raphson en calculant la Hessienne $\mathbf{Hm_s}^{(t)}$ est une alternative. Finalement, itérer le calcul de $u_{\ell,s}^{(t)}$ et $\sigma^{(t)}$ converge vers un maximum (local). Nous notons avec un chapeau les paramètres finaux. Les centres résultants, dans l'espace des données, sont facilement reconstruits à partir des vecteurs $\hat{m}_k$ par l'inverse de la translation précédent, l'homothétie, et une rotation lorsque nécessaire, comme expliqué ci-dessous.

### 2.2. *La rotation*

Lorsque l'inertie des données est mal orientée vis à vis des axes du repère, on peut préférer mettre en oeuvre une rotation pour obtenir le maximum d'inertie expliquée par le modèle tabulaire. Une première approche dans le choix d'une rotation optimale de la vue tabulaire le long des directions de plus grande variance est d'ajouter des matrices de covariance non sphériques dans les densités gaussiennes. Cette solution non explicite est remplacée ici pas une rotation had hoc sur l'échantillon projeté. Nous ajoutons une transformation matricielle sur $y_i$ telle que nous remplaçons $y_i$ par $\bar{y} + W(y_i - \bar{y})$ où $W$ est une matrice $\mathbb{R}^{2\times 2}$ et $\bar{y}$ est la moyenne empirique des vecteurs $y_i$. Le gradient est modifié en remplaçant $y_{is}$ par $\bar{y}_s + \sum_{s'} w_{ss'}^{(t)} \bar{y}_{is'}$ où $\bar{y}_{is'} = (y_{is'} - \bar{y}_{s'})$. Ajouter une contrainte additive permet d'éviter une solution dégénérée. Une plus élégante transformation est la vraie matrice de rotation $W^{(t)}$, qui s'écrit, pour un angle de rotation $\alpha^{(t)}$, et une mise à l'échelle implicite de facteur $\beta$ pour maintenir les étendues des nouvelles coordonnées des projetés dans [0,1] :

$$W^{(t)} = \begin{bmatrix} cos\,\alpha^{(t)} & -sin\,\alpha^{(t)} \\ sin\,\alpha^{(t)} & cos\,\alpha^{(t)} \end{bmatrix}.$$

Finalement la première dérivée $\mathbf{D}_\alpha{}^{(t)}$ et la seconde $\mathbf{H}_\alpha{}^{(t)}$ fournissent le pas de Newton-Raphson $\alpha^{(t+1)} = \alpha^{(t)} - \mathbf{D}_\alpha{}^{(t)}/\mathbf{H}_\alpha{}^{(t)}$ qui, à la convergence, aboutit à la rotation désirée. Le modèle complet amélioré partage les propriétés intéressantes de plusieurs méthodes d'analyse des données comme expliqué dans la section d'illustration suivante.

## 3. Application à la visualisation

Le modèle tabulaire est similaire à une analyse en composante principale [LEB 84] linéaire discrétisée qui donne une rotation identique : lorsque les centres sont assez nombreux, et avec une affectation aux classes non lissées. La solution correspondante est équivalente à l'ACP car $\sum_i ||y_i - [\bar{y} + \hat{W}^{-1}(\hat{m}_{\hat{z}_i} - \bar{y})]||^2 = \sum_i ||\bar{y}_i - \hat{W}^{-1}\hat{\hat{m}}_{\hat{z}_i}||^2$ est alors minimal. La méthode est également liée à une discrétisation et une construction d'un histogramme pour une variable bivarié : nous obtenons deux nouvelles variables discrètes, le long des lignes et des colonnes, avec un choix d'intervalles sensible à la densité marginale de la projection bivariée. Nous sommes capables de voir le nuage de données d'une façon plus rapide et plus accessible. Il s'agit d'une approche complémentaire aux solutions existantes en visualisation avec la forte perspective d'améliorer leur propre définition avec la contribution de la densité locale du nuage de données.

Nous illustrons le modèle sur 500 images à partir d'un échantillon de 2000 images de chiffres digitalisés en composantes binaires. La méthode de projection est ici LLE (*Locally Linear Embedding*) [ROW 00]. Cette

méthode est très efficace lorsque les images traitées appartiennent à des classes d'objets de formes similaires car des relations linéaires locales existent dans ce cas. Nous obtenons une projection de l'échantillon, les $y_i$ sur la Figure 1. Le modèle tabulaire permet de montrer le plan de projection discrétisé par une visualisation au niveau du pixel qui est alors naturelle et apporte une représentation synthétique et potentiellement interactive. Etant donnée la localisation des dix classes sur le plan discrétisé sur la Figure 1, chaque classe représentée par une carte de pixels, nous obtenons la visualisation bien définie sur la Figure 2.

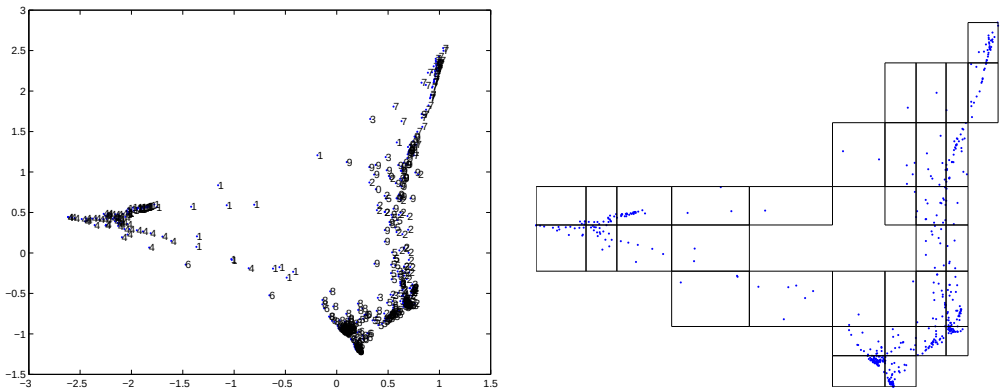


**FIG. 1.** *La projection LLE des 500 chiffres avec un paramètre de 6 plus proches voisins sur la gauche et les clusters rectangulaires du modèle tabulaire résultants sur la droite.*

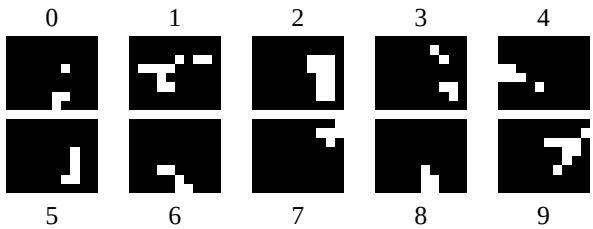



**FIG. 2.** *Les 10 cartes de pixels illustrent le paradigme d'une représentation au niveau du pixel : ces cartes (numérotées par les labels chiffre) montrent d'une façon synthétique les relations pour le LLE entre les 10 classes de chiffres . Une application potentiellement interactive est de sélectionner une portion de la projection 2D originale et de découvrir sur les cartes de pixels la zone correspondant à la portion sélectionnée, avec des intensités de couleur correspondantes aux taux de remplissage des labels associés.*

Il est alors possible de comparer le lieu des classes sur les cartes, sans nécessairement mémoriser précédemment une quelconque carte ou superposer des densités bivariées complexes contrairement au cas de l'usuelle représentation lissée du nuage de données projetées. Une information supplémentaire peut se placer au niveau du pixel. Par exemple, en considérant une carte de pixels unique, il est possible de placer plusieurs classes à l'intérieur d'une unique cellule. Il est également possible de remplacer les labels de classe par un ensemble de variables ou modalités. Cette approche est particulièrement intéressante notamment en fouille de données textuelles car des significations contextuelles existent entre les mots.

## 4. Conclusion

Nous avons proposé une nouvelle manière d'explorer une projection de données par un modèle génératif qui construit un tableau de classes à partir d'un nuage de données qui est généralement difficile à lire lorsque le

nombre de points visualisés est important. Notre méthode présente des propriétés qui lui permettent d'effectuer plusieurs traitements sur la projection afin d'obtenir une représentation finale interprétable : une rotation de type ACP associée à une discrétisation de type percentile. Il s'agit d'une alternative à un découpage régulier du plan de projection afin de représenter davantage d'informations visuelles. La visualisation basée sur les pixels est introduite pour montrer les cartes finales et permettre une comparaison directe des positions des classes toutes à la fois. La méthode est indépendante du choix de la projection pour obtenir les points 2D. Il s'agit d'une alternative aux cartes auto-organisatrices appliquées au données d'origine, lorsqu'une projection bidimensionnelle existe déjà par le biais d'une autre méthode non linéaire. Cependant, une carte auto-organisatrice pourrait être utilisée pour analyser notre projection 2D, au lieu du modèle tabulaire. La différence de l'approche est le risque de perdre le lieu exact des points vis à vis de la projection d'origine. Cette alternative à notre travail doit être étudiée en complément à nos travaux dans le futur. Des contraintes additives peuvent être utilisées pour améliorer le contenu tabulaire. Il semble que la densité de la projection des nuages de données reste encore peu utilisée dans les outils actuels de visualisation des données (massives). Une telle information peut s'utiliser pour mieux visualiser et *zoomer* la densité locale de données 2D ou 3D.

## 5. Bibliographie


[DEM 77] DEMPSTER A., LAIRD N., RUBIN D., Maximum-likelihood from incomplete data via the em algorithm, *J. Royal Statist. Soc. Ser. B., 39*, , 1977.

[KOH 97] KOHONEN T., *Self-organizing maps*, Springer, 1997.

[LEB 84] LEBART L., MORINEAU A., WARWICK K., *Multivariate Descriptive Statistical Analysis*, J. Wiley, 1984.

[MAC 67] MACQUEEN J., Some Methods for Classification and Analysis of Multivariate Observations, *5th Berkeley Symp. Math. Stat. and Proba.*, vol. 1, 1967, p. 281-296.

[ROW 00] ROWEIS S. T., SAUL L. K., Nonlinear Dimensionality Reduction by Locally Linear Embedding, *Science*, vol. 290, n$^o$ 5500, 2000, p. 2323–2326.

[SAM 69] SAMMON J., A nonlinear mapping for data structure analysis, *IEEE Transactions on Computers*, vol. 5, n$^o$ 18C, 1969, p. 401-409.

# Évaluation des méthodes supervisées pour la discrimination de protéines

**Ricco Rakotomalala, Faouzi Mhamdi**

*Laboratoire ERIC – Université Lyon 2*
*69500 BRON*
*ricco.rakotomalala@univ-lyon2.fr*
*URPAH – Université d'El Manar*
*TUNISIE*
*faouzi.mhamdi@ensi.rnu.tn*

RÉSUMÉ. *Nous évaluons différentes méthodes supervisées dans le cadre très particulier de la discrimination de protéines. Les descripteurs étant automatiquement générés, nous avons utilisé les $n$-grammes, nous nous retrouvons face à un double problème : la taille de l'espace de représentation est très élevée par rapport au nombre d'observations, d'un rapport de 4 à 300 selon le type de description choisi ; un grand nombre de descripteurs ne sont pas pertinents pour la discrimination. Il ressort nettement dans ce contexte que les méthodes linéaires telles que les SVM linéaires ou la régression PLS sont celles qui s'avèrent être les plus adaptées. En cela nous rejoignons un très grand nombre de publications où les SVM notamment s'avèrent très performants. Une étude détaillée des résultats montre que ce succès repose essentiellement sur la robustesse face à la dimensionnalité qui devient, de fait, le critère le plus important dès lors que l'on traite des domaines où les descripteurs sont générés automatiquement en grand nombre.*

MOTS-CLÉS : *Bio-informatique, Fouille de données, Apprentissage supervisé, Catégorisation de textes, $n$-grammes*

## 1. Introduction

L'annotation et le classement de protéines est une activité importante du biologiste. L'augmentation du volume de données à traiter rend nécessaire l'automatisation de cette tâche. Ces dernières années, l'extraction de connaissances à partir de données [FAY 96] a permis de dégager un cadre qui rend reproductible le processus de classement automatique des protéines à partir de leur structure primaire. En effet, une séquence de protéine est décrite par une suite de caractères pris dans un alphabet de 20 signes. Le rapprochement avec les nombreux travaux réalisés dans la catégorisation de textes est naturel [SEB 05]. Par rapport à un traitement standard, le traitement de données non-structurées introduit une étape supplémentaire : l'extraction de descripteurs à partir de la description primaire pour aboutir à un tableau de données exploitable par les algorithmes d'apprentissage.

Dans cet article, nous traitons de la discrimination de protéines à partir de leur structure primaire. Les protéines sont regroupés en familles selon leur fonction. Il est admis que les protéines appartenant à la même famille ont des structures identiques, encore faut-il pouvoir caractériser la similarité entre deux protéines. Notre objectif est de construire une fonction de classement qui permet d'associer automatiquement une protéine à sa famille d'appartenance.

Dans la section suivante, nous présentons la démarche de discrimination de protéines. Puis nous présentons l'expérimentation que nous avons mis en place pour évaluer notre démarche. Nous concluons dans la 4ème et dernière section.

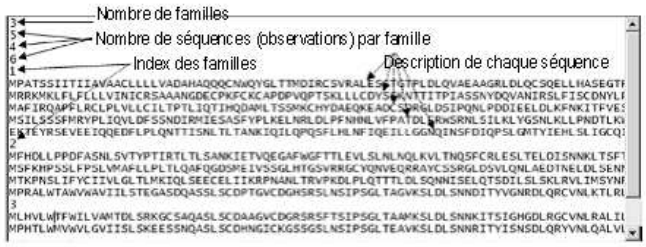



**FIG. 1.** *Description native des séquences de protéines*

## 2. La discrimination de protéines

Les fichiers utilisés dans cet article proviennent de la banque de données SCOP [MUR 95]. Pour chaque famille de protéines, nous disposons d'une série d'observations. Chaque observation est décrite par une chaîne de caractères de longueur variable (Figure 1). Traiter telle quelle cette description native avec les algorithmes usuels de fouille de données n'est pas possible. Il nous faut les transformer de manière à disposer d'un tableau attribut-valeur. En nous inspirant des résultats mis en avant dans la catégorisation de textes, nous nous sommes tournés vers la technique des $n$-grammes.

Un $n$-gramme correspond à une suite de caractères de longueur $n$. La transformation consiste à repérer tous les $n$-grammes possibles dans les fichiers. Chaque $n$-gramme correspond à un nouveau descripteur, nous signalons dans la colonne la présence ou l'absence du $n$-gramme pour chaque observation constituant la base d'apprentissage. Nous aboutissons donc à un tableau de données booléen $(0/1)$. D'autres types de pondération peuvent être mis en oeuvre : l'occurrence du $n$-gramme, sa fréquence, etc. Nos expérimentations montrent que ce choix pèse peu sur les performances.

L'enjeu est de définir la longueur adéquate du $n$-gramme, la valeur de $n$. Si elle est trop faible, par exemple si nous fixons $n = 1$, l'information capturée est trop pauvre, chaque caractère étant de surcroît présent dans quasiment toutes les séquences de protéines, la colonne correspondante sera remplie de $1$ (présence du $1$-gramme). Si la valeur de $n$ est trop élevée, l'information capturée devient trop spécifique, nous nous heurtons à deux problèmes : le nombre de colonnes du tableau d'apprentissage sera colossal, le nombre théorique de colonnes étant égal à $20^n$, le calcul ne sera pas possible dans la plupart des cas ; chaque colonne sera presque toujours remplie de $0$ (absence du $n$-gramme). La longueur $n$ ne peut donc être que la résultante d'un compromis, rien ne nous indique au départ sa valeur adéquate s'agissant de la discrimination de protéines. Notre expérience de la catégorisation de textes nous a tout simplement montré que dans certains cas $n = 3$ donne des résultats intéressants, dans la catégorisation de nouvelles par exemple.

## 3. Expérimentation

### 3.1. *Données et méthodes d'apprentissage*

Pour évaluer notre approche, nous avons extrait $5$ familles de protéines au hasard de la **banque de données** SCOP [MUR 95]. Nous disposons approximativement de $50$ observations par famille. Dans cet article, nous cherchons à discriminer les familles de protéines deux à deux, nous avons donc constitué $10$ fichiers de données constitués d'une centaine d'observations.

A partir de la description native (Figure 1), nous avons construit des tableaux booléens de données en utilisant le principe des $n$-grammes. Nous avons fait varier $n$ de $2$ à $4$. Si le nombre théorique de descripteurs que l'on peut obtenir est égal à $2^n$, dans la pratique, nous en observons largement moins à mesure que $n$ augmente ($400$ en moyenne pour $n = 2$ ; $6500$ pour $n = 3$ ; $30000$ pour $n = 4$).

| Methode | Biais | Variance |
|---|---|---|
| SVM Linéaire | Linéaire | Faible |
| PLS (2 axes) | Linéaire | Faible |
| Bayesien Naïf | Linéaire | Modérément faible |
| CART | Non-linéaire | Élevée |
| 1-PPV | Non-linéaire | Élevée |
| SVM RBF | Non-linéaire | Faible |

**TAB. 1.** *Caractéristiques des méthodes d'apprentissage utilisées*

Concernant les **méthodes d'apprentissage**, il existe une multitude de points de vue. Il est bien souvent difficile d'en discerner clairement les caractéristiques. Dans notre cas, une bonne manière de procéder est de les positionner selon leur mode de représentation d'une part, selon leur préférence d'apprentissage d'autre part (Tableau 1). Il est possible de trouver une vue synthétique de ces algorithmes[1] dans l'ouvrage de Hastie et al. [HAS 01].

Le premier indique la capacité de la méthode à retraduire la "'forme"' d'un concept, nous distinguons les modèles linéaires des modèles non-linéaires. A priori, nous avons toujours intérêt à choisir un modèle non-linéaire : "'qui peut le plus, peut le moins"'. En réalité, la situation est un peu plus compliquée. En effet, sur les données synthétiques le "'concept"' à apprendre existe vraiment puisque artificiellement créé par le chercheur. Il en autrement sur données réelles où la relation entre les descripteurs et la variable à prédire est une vue de l'esprit, nous essayons de retraduire une hypothétique causalité avec une fonction mathématique. Cette (in)capacité à appréhender les concepts se traduit généralement par le terme de "'biais"' dans la littérature.

Le second critère, la préférence d'apprentissage, décrit le mode d'exploration des solutions mise en oeuvre par l'algorithme, elle permet de restreindre la recherche. Bien souvent, mais ce n'est pas toujours le cas, les caractéristiques du mode d'exploration est retraduite par le critère à optimiser lors de l'apprentissage. A priori, nous avons tout intérêt à choisir une méthode qui teste toutes les hypothèses possibles de manière à choisir la meilleure. En réalité, ce n'est pas toujours vrai. En effet, le principal danger est de retrouver des particularités propres au fichier d'apprentissage au détriment de la "'vraie"' relation que nous voulons mettre en évidence. Notre situation est d'autant plus difficile que nous traitons des données où le nombre de descripteurs est très élevé par rapport aux observations, les fonctions de distributions conditionnelles sont mal estimées. Cette (in)stabilité par rapport au fichier d'apprentissage se traduit généralement par le terme de "'variance"' dans la littérature.

Les termes "'biais"' et "'variance"' sont également utilisés en référence aux composantes de l'erreur quadratique, utilisée pour évaluer les performances des méthodes d'apprentissage.

### 3.2. *Résultats*

Les résultats de nos expérimentations sont résumés dans le tableau 2. Nous nous sommes intéressés au taux de mauvais classement, le taux d'erreur, estimé par validation croisée. Deux résultats semblent s'imposer : les méthodes à fort biais (linéaires) sont les meilleures ; les $n$-grammes de longueur $n = 2$ suffisent largement dans la discrimination des protéines.

Une lecture plus attentive des résultats relativise ces premières impressions. Nous pouvons nous attendre en effet qu'un SVM avec un noyau RBF (radial basis function) s'en sorte mieux dans la mesure où il est capable de retraduire des concepts plus complexes. Or nous constatons que c'est la méthode globalement la moins performante. Second point important qui attire notre attention. Dans la plupart des cas, la qualité de l'apprentissage se dégrade à mesure que l'on augmente la dimensionnalité ($n = 3$ et $n = 4$), sauf en ce qui concerne les arbres

1. Toutes ces méthodes sont implémentées et sont disponibles dans le logiciel TANAGRA (http ://eric.univ-lyon2.fr/ ricco/tanagra). Dans le cas des SVM, nous appelons directement des bibliothèques externes (LIBSVM – http ://www.csie.ntu.edu.tw/ cjlin/libsvm/)

| Methode | 2-grammes | 3-grammes | 4-grammes |
|---|---|---|---|
| SVM Linéaire | 0.032 | 0.038 | 0.081 |
| PLS (2 axes) | 0.041 | 0.048 | 0.102 |
| Bayesien Naïf | 0.048 | 0.071 | 0.248 |
| CART | 0.210 | 0.155 | 0.141 |
| 1-PPV | 0.043 | 0.214 | 0.269 |
| SVM RBF | 0.063 | 0.130 | 0.479 |

**TAB. 2.** *Moyenne du taux d'erreur selon les méthodes et le type de représentation*

de décision. CART réalise automatiquement une sélection des attributs pertinents. Il semble que dans certains cas $n = 4$ propose un espace de description intéressant. Certes, CART est en retrait par rapport aux méthodes linéaires, mais il faut surtout y voir un problème lié à la faiblesse des effectifs, les arbres souffrent très vite de la fragmentation des données.

Nous pouvons dès lors nous demander si les bonnes performances pour $n = 2$ n'indiquent pas davantage un apprentissage mieux assuré dans un espace réduit, plutôt que l'hypothétique mise en évidence de la longueur "'optimale"' des $n$-grammes. De la même manière, nous pouvons nous demander si dans ce contexte, le fait que les méthodes linéaires se placent en meilleure position ne soit pas tout simplement la conséquence de leur capacité à résister à un espace sur-dimensionné et très bruité. En effet, en fixant une contrainte forte de représentation, la méthode est moins soumise aux perturbations occasionnées par les descripteurs non-pertinents.

Pour évaluer l'effet de la dimensionnalité lorsque $n = 3$, nous avons introduit une méthode de sélection de variables très simple. Nous avons choisi dans chaque fichier les 30 descripteurs les plus corrélés avec la variable à prédire. Il apparaît dans ce cas que les SVM(RBF) et 1-PPV présentent des performances se rapprochant des méthodes linéaires.

## 4. Conclusion

Nous avons développé une approche inspirée de la catégorisation de textes pour résoudre un problème de discrimination de protéines. Il apparaît que l'utilisation des $n$-grammes permet d'extraire des descripteurs performants. Les méthodes linéaires s'imposent par la suite pour fournir des taux d'erreur de classement faibles. Une étude détaillée des expérimentations semble indiquer que ce résultat repose avant tout sur la capacité de ces méthodes à résister aux espaces de représentation sur-dimensionnés, avec de nombreux descripteurs non-pertinents.

Ces résultats nous suggèrent la voie à suivre pour l'amélioration de notre processus de discrimination des protéines : développer des méthodes efficaces de sélection de variables pour ne pas perturber l'apprentissage ; lever la restriction sur la taille $n$ des $n$-grammes et donc proposer une procédure pour extraire des descripteurs de longueur variable. Ici également, il nous faudra restreindre fortement l'espace de recherche, l'utilisation des techniques fondées sur la recherche des co-occurrences significatives des $n$-grammes semble une piste intéressante.

## 5. Bibliographie


[FAY 96] FAYYAD U., PIATETSKY-SHAPIRO G., SMYTH P., From data mining to knowledge discovery in databases, *Ai Magazine*, vol. 17, 1996, p. 37–54.

[HAS 01] HASTIE T., TIBSHIRANI R., FRIEDMAN J., *The elements of statistical learning*, Springer Series in Statistics, Springer-Verlag, New York, 2001.

[MUR 95] MURZIN G., BRENNER E., HUBBARD T., CHOTHIA C., SCOP : a structural classification of proteins database for the investigation of sequences and structures, *J. Mol. Bio.*, vol. 247, 1995, p. 536–540.

[SEB 05] SEBASTIANI F., Text Categorization, ZANASI A., Ed., *Text Mining and its Applications to Intelligence, CRM and Knowledge Management*, p. 109–129, WIT Press, Southampton, UK, 2005.

# Hiérarchies semi-floues et degré d'imbrication de hiérarchies

**Sahondra Ravonialimanana, Henri Ralambondrainy, Jean Diatta**

*Université de Fianarantsoa, Université de la Réunion*
*rafilipo@wanadoo.mg, {ralambon,jdiatta}@univ-reunion.fr*

RÉSUMÉ. *Nous définissons la notion de hiérarchie semi-floue à partir de celle de hiérarchie classique. Nous proposons des indices mesurant d'une part l'imbrication d'un ensemble flou dans une hiérarchie semi-floue, et d'autre part l'imbrication d'une hiérarchie semi-floue dans une autre. Ces indices sont des outils permettant de comparer deux hiérarchies classiques.*

MOTS-CLÉS : *classification, hiérarchie, ensemble flou*

## 1. Introduction

Les méthodes de Classification Ascendante Hiérarchiques (CAH)[2] structurent des données usuelles en une hiérarchie d'ensembles classiques. Dans ce papier, nous montrons que de tels algorithmes produisent sur un ensemble flou [1] une structure particulière que nous qualifions de hiérarchie semi-floue généralisant celle de hiérarchie classique. Puis, en partant de l'indice de Kosko [3], nous proposons des indices mesurant d'une part l'imbrication d'un ensemble flou dans une hiérarchie semi-floue, et d'autre-part l'imbrication d'une hiérarchie semi-floue dans une autre. Ces indices sont des outils permettant de comparer deux hiérarchies classiques.

## 2. Hiérarchies semi-floues

### 2.1. *Rappels sur les ensembles flous*

On note $X$ le référentiel, un ensemble fini d'observations, de cardinal $n$. Un sous ensemble flou $h$ de $X$ est défini par une fonction d'appartenance $m_h \in [0,1]^X$. Le support de $m_h$ est l'ensemble $supp(m_h) = \{x \in X / m_h(x) \neq 0\}$ ; $\mathcal{F}(X)$ désigne l'ensemble des sous-ensembles flous de $X$. Nous adopterons les opérations ensemblistes usuelles suivantes sur $\mathcal{F}(X)$ : *égalité* : $h = h' \iff m_h = m_{h'}$, *inclusion* : $h \subseteq h' \iff m_h \leq m_{h'}$, *intersection* : $m_{h \cap h'} = m_h \wedge m_{h'}$. Muni de l'ordre d'inclusion $\mathcal{F}(X)$ a un plus petit élément $m_\emptyset$ noté $\mathbf{0}$ et un plus grand élément $m_X$ noté $\mathbf{1}$. Un ensemble flou sera désigné indifféremment par $h$ ou $m_h$.

### 2.2. *Hiérarchies semi-floues*

Soit $\mathcal{H}$ une hiérarchie totale sur $X$. En identifiant les sous-ensembles de $X$ à leurs fonctions caractéristiques respectives, les propriétés de $\mathcal{H}$ peuvent s'exprimer comme suit :
1) $\mathbf{1} \in \mathcal{H}$ ; 2) $\forall H, H' \in \mathcal{H} : \mathbf{1}_{H \cap H'} = \mathbf{1}_H \wedge \mathbf{1}_{H'} \in \{\mathbf{1}_H, \mathbf{1}_{H'}, \mathbf{0}\}$ ; 3) $\forall x \in X : \mathbf{1}_{\{x\}} \in \mathcal{H}$.
Cela nous conduit à la définition suivante d'une hiérarchie semi-floue :

**Définition 2.1** *Une hiérarchie semi-floue $\mathcal{F}$ est une famille de sous-ensembles flous de $\mathcal{F}(X)$ vérifiant :*
*B1)* $\mathbf{1} \in \mathcal{F}$ ; *B2)* $\forall h, h' \in \mathcal{F} : m_{h \cap h'} = m_h \wedge m_{h'} \in \{m_h, m_{h'}, \mathbf{0}\}$ ; *B3)* $\forall x \in X : m_{\{x\}} \in \mathcal{F}$ ; *où pour* $x \in X$, $m_{\{x\}}$ *désignera un ensemble flou dont le support est* $\{x\}$.

La propriété (B2) signifie que deux classes $h, h'$ d'une hiérarchie semi-floue sont telles que $m_h \leq m_{h'}$ ou $m_{h'} \leq m_h$ ou $supp(m_h) \cap supp(m'_h) = \emptyset$. Les partitions associées à une hiérarchie semi-floue ne sont pas des partitions floues au sens de [4]. Un élément appartient à un certain degré à une classe mais les supports des classes sont d'intersection vide comme dans le cas des partitions classiques. Si $\mathcal{F}$ est une hiérarchie semi-floue sur $X$, alors il est facile de voir que $\mathcal{H}_{\mathcal{F}} = \{H = supp(m_h)/h \in \mathcal{F}\}$ est une hiérarchie classique. On dira que $\mathcal{H}_{\mathcal{F}}$ est le support de la hiérarchie et l'on notera $\mathcal{H}_{\mathcal{F}} = supp(\mathcal{F})$.

### 2.3. *Construction d'une hiérarchie semi-floue*

Une manière simple pour construire une hiérarchie semi-floue est d'utiliser un algorithme classique de CAH sur un ensemble flou donné. Soit $A$ un ensemble flou dont le support est $X$, et $\Delta$ un indice de dissimilarité défini sur $\mathcal{P}(X)$, déterminé à partir de la fonction d'appartenance $m_A$ et d'un indice d'agrégation classique $\delta$. Par exemple, l'indice défini pour tout $h, h' \in \mathcal{P}(X)$ par

$$\Delta(h, h') = \frac{\delta(h, h')}{(\vee_{x \in h} m_A(x))(\vee_{x' \in h'} m_A(x'))}$$

agrégera des classes en tenant compte de la dissimilarité entre les classes et du degré d'appartenance des éléments aux classes.

**Proposition 2.1** *Soit $A$ un ensemble flou dont le support est $X$ et $\mathcal{H}$ une hiérarchie classique totale sur $X$. L'ensemble $\mathcal{F} = \{m_H | H \in \mathcal{H}\}$ tel que :*

1. $\forall x, y \in X,\ m_{\{x\}}(x) = m_A(x)$;
2. $H \neq X,\ m_H(y) = \vee_{x \in H} m_{\{x\}}(y)$;
3. $m_X = \mathbf{1}$.

*est une hiérarchie semi-floue dont le support est $\mathcal{H}$.*

**Démonstration.** Par définition de $\mathcal{F}$, il est facile de voir que les propriétés $B1$ et $B3$ sont vérifiées par $\mathcal{F}$. Remarquons que si $y \in H$ alors $m_H(y) = \vee_{x \in H} m_{\{x\}}(y) = m_{\{y\}}(y)$ et que si $y \notin H, m_H(y) = 0$. Autrement dit $H = supp(m_H)$. Etudions la propriété (B2). Soient $m_H, m_{H'} \in \mathcal{F}$. Si $H \cap H' = \emptyset$ alors $m_H \wedge m_{H'} = \mathbf{0}$. En effet, si $y \in X$, alors $m_H \wedge m_{H'}(y) \neq 0$ si et seulement si $m_H(y) \neq 0$ et $m_{H'}(y) \neq 0$ c.-à-d. pour $y \in H$ et $y \in H'$. Si $H \cap H' = \emptyset$ alors il n'existe pas de tel $y$ et donc $m_H \wedge m_{H'} = \mathbf{0}$. Si $H \cap H' \neq \emptyset$, comme $\mathcal{H}$ est une hiérarchie, on peut supposer $H \subset H'$, et clairement $m_H = \vee_{x \in H} m_{\{x\}} \leq m_{H'} = \vee_{x' \in H'} m_{\{x'\}}$ et alors $m_H \wedge m_{H'} = m_H$. □

Il est facile de voir que si $A = X$ c.-à-d. $m_A = \mathbf{1}$, alors la hiérarchie semi-floue $\mathcal{F}$ associée à une hiérarchie classique totale $\mathcal{H}$ sur $X$ est $\mathcal{H}$.

### 2.4. *Imbrication de deux ensembles flous*

Dans cette partie, nous définissons des indices d'imbrication plus généraux à partir de l'indice de Kosko [3] qui mesure l'imbrication d'ensembles flous, .

**Définition 2.2** *La magnitude d'une classe floue $h$ est définie par :* $M(h) = \sum_{x \in X} m_h(x)$.

Pour un ensemble fini classique $H$, on a $M(H) = card(H)$. Le degré d'imbrication d'une classe $h$ dans une classe $h'$ est mesuré par la proportion, au sens de la magnitude, des éléments de $h$ appartenant à $h'$. Plus précisément, on a :

**Définition 2.3** *Etant donnés $h, h'$, deux éléments de $\mathcal{F}(X)$, l'imbrication de $h$ dans $h'$ est définie par :*

$$S(h, h') = \frac{M(h \bigcap h')}{M(h)}$$

Lorsque $H$ et $H'$ sont des ensembles finis classiques, l'imbrication de $H$ dans $H'$ s'écrit :

$$S(H,H') = \frac{M(H \bigcap H')}{M(H)} = \frac{card(H \bigcap H')}{card(H)}$$

On a les propriétés suivantes :

**Proposition 2.2** *Pour tout $h, h', h'' \in \mathcal{F}(X)$ et pour tout $x \in X$ on a :*

1. $S(\{x\}, h) = m_h(x)$ ;
2. $S(h, \emptyset) = 0$ ;
3. $h \subseteq h' \iff S(h,h') = 1$ ;
4. $h' \subset h \iff 0 < S(h,h') < 1$ ;
5. $h \bigcap h' = \emptyset \iff S(h,h') = S(h',h) = 0$ ;
6. $h' \subseteq h'' \Rightarrow S(h,h') \leq S(h,h'')$ .

### 2.5. *Imbrication d'un sous ensemble flou dans une hiérarchie semi-floue*

Dans cette section, on se donne une hiérarchie semi-floue $\mathcal{F}$. La stratégie adoptée pour mesurer l'imbrication d'un sous ensemble flou $Y$ dans une hiérarchie semi-floue $\mathcal{F}$ est la suivante : on commence par "situer" l'ensemble $Y$ dans la hiérarchie, c.-à-d. on recherche d'une part la plus petite classe de $\mathcal{F}$ contenant $Y$ (c.-à-d. $\overline{Y}$) et d'autre part les éléments de $\mathcal{F}$ maximaux parmi ceux contenus dans $Y$. Puis, on calcule aussi bien le degré d'imbrication de $Y$ dans ces éléments maximaux que le degré d'imbrication de $\overline{Y}$ dans $Y$.

Si $E$ est une famille de sous-ensembles flous, en considérant l'ensemble ordonné $(\mathcal{F}(X), \subset)$, on note $Max(E)$ (respectivement $Min(E)$) l'ensemble des éléments maximaux (respectivement minimaux) de $E$. Soit un ensemble flou non vide $Y \in \mathcal{F}(X)$. On désigne par $Y^u$ l'ensemble des classes de $\mathcal{F}$ qui majorent $Y$ et $Y^l$ l'ensemble des classes de $\mathcal{F}$ qui minorent $Y$. Soit $\overline{Y} = Min(Y^u)$. Pour tout $h, h' \in Y^u$, on a $Y \subseteq h$ et $Y \subseteq h'$. Donc $Y \subseteq h \cap h'$. Comme $\mathcal{F}$ est une hiérarchie semi-floue, on a $h \subseteq h'$ ou $h' \subseteq h$ de sorte que $Y^u$ est une chaîne qui admet un seul élément minimal qui en est donc son plus petit élément : $\overline{Y} = Min(Y^u)$. On assimilera $\overline{Y}$ à cet élément unique. $\overline{Y}$ existe toujours car $Y \subseteq X$ et $\mathbf{1} \in \mathcal{F}$. On note $\underline{Y}$ l'ensemble des minorants de $Y$ maximaux dans $Y^l$ : $\underline{Y} = Max(Y^l)$. Le degré d'imbrication d'un sous ensemble flou $Y$ dans une hiérarchie semi-floue $\mathcal{F}$ est mesuré par :

$$S(Y,\mathcal{F}) = \frac{1}{1+|\underline{Y}|}(S(\overline{Y},Y) + \sum_{y \in \underline{Y}} S(Y,y)).$$

On a la proposition :

**Proposition 2.3** *Soit $\mathcal{F}$ une hiérarchie semi-floue, on a :*

1. $0 \leq S(Y,\mathcal{F}) \leq 1$ ;
2. $Y \in \mathcal{F} \iff S(Y,\mathcal{F}) = 1$.

Cette proposition nous permet de considérer une hiérarchie semi-floue comme un ensemble flou de $\mathcal{F}(\mathcal{F}(X))$ muni de la fonction d'appartenance $m_{\mathcal{F}}(Y) = S(Y,\mathcal{F})$.

## 3. Comparaison de deux hiérarchies semi-floues

Soient $\mathcal{F}_1$ et $\mathcal{F}_2$ deux hiérarchies semi-floues définies sur un ensemble $X$. L'inclusion de $\mathcal{F}_1$ dans $\mathcal{F}_2$ est l'inclusion ensembliste classique : $\mathcal{F}_1 \subseteq \mathcal{F}2 \iff (\forall h \in \mathcal{F}_1 \Rightarrow h \in \mathcal{F}_2)$. Soient $Y \subset X$ et $\mathcal{F}_Y$ une hiérarchie semi-floue définie sur $Y$, et $\mathcal{F}$ une hiérarchie semi-floue sur $X$. On dira que $\mathcal{F}_Y \subset \mathcal{F}$ si $(\forall h \in \mathcal{F}_Y \Rightarrow h \in \mathcal{F})$. La magnitude d'une hiérarchie semi-floue $\mathcal{F}$ a pour expression $M(\mathcal{F}) = \sum_{Y \in \mathcal{F}(X)} m_{\mathcal{F}}(Y)$.

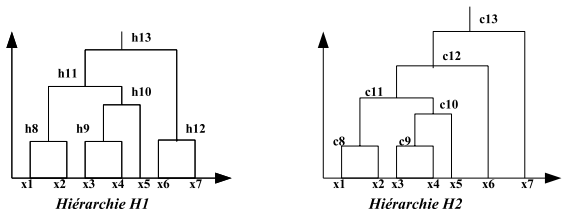



**FIG. 1.** *Comparaison de hiérarchies*

### 3.1. *Imbrication d'une hiérarchie semi-floue dans une autre hiérarchie semi-floue*

Soient deux hiérarchies floues $\mathcal{F}_1$ et $\mathcal{F}_2$ sur $X$. L'ensemble flou $\mathcal{F}_1 \bigcap \mathcal{F}_2$ a pour fonction d'appartenance $m_{\mathcal{F}_1 \bigcap \mathcal{F}_2} = m_{\mathcal{F}_1} \wedge m_{\mathcal{F}_2}$. L'imbrication de $\mathcal{F}_1$ dans $\mathcal{F}_2$ est mesurée par :

$$S(\mathcal{F}_1, \mathcal{F}_2) = \frac{M(\mathcal{F}_1 \bigcap \mathcal{F}_2)}{M(\mathcal{F}_1)}$$

où $M(\mathcal{F}_1 \bigcap \mathcal{F}_2) = \sum_{Y \in \mathcal{F}(X)} m_{\mathcal{F}_1} \wedge m_{\mathcal{F}_2}(Y)$. Les propriétés de cet indice d'imbrication sont résumées ci-dessous

**Proposition 3.1** *Soient $\mathcal{F}_1$ et $\mathcal{F}_2$ deux hiérarchies semi-floues sur $X$ alors :*

*1.* $S(\mathcal{F}_1, \mathcal{F}_2) = 1 \Leftrightarrow \mathcal{F}_1 \subseteq \mathcal{F}_2$;

*2.* $S(\mathcal{F}_1, \mathcal{F}_2) = S(\mathcal{F}_2, \mathcal{F}_1) = 1 \Leftrightarrow \mathcal{F}_1 = \mathcal{F}_2$;

*3.* $S(\mathcal{F}_1, \mathcal{F}_2) = 1$ *et* $S(\mathcal{F}_2, \mathcal{F}_3) = 1$ *impliquent* $S(\mathcal{F}_1, \mathcal{F}_3) = 1$;

*4.* $0 < S(\mathcal{F}_1, \mathcal{F}_2) \leq 1$;

*5. Si* $\mathcal{F}_1 \subseteq \mathcal{F}_2$ *et* $\forall h_2 \in \mathcal{F}_2$, $\exists h_1 \in \mathcal{F}_1$ *tel que* $h_1 \subseteq h_2$ *alors :* $\forall\, Y \in \mathcal{F}(X), \quad S(Y, \mathcal{F}_1) \leq S(Y, \mathcal{F}_2)$;

*6. Soient* $A, B \subset X$, $\mathcal{F}_A, \mathcal{F}_B$ *des hiérarchies semi-floues quelconques définies respectivement sur* $A$ *et* $B$. *Alors* $A \cap B = \emptyset$ *est équivalent à* $S(\mathcal{F}_A, \mathcal{F}_B) = S(\mathcal{F}_B, \mathcal{F}_A) = 0$.

### 3.2. *Imbrication de hiérarchies classiques*

Les indices d'imbrication définis précédemment s'appliquent aux hiérarchies classiques car ces dernières sont des cas particuliers des hiérarchies semi-floues. La Figure 1 présente deux hiérarchies classiques $H1$ et $H2$ ayant 13 classes et qui diffèrent par les classes $h_{12}$ et $c_{12}$. Les ensembles considérés pour le calcul des degrés d'imbrication mutuels de $H1$ et $H2$ sont les classes de $H1 \cup H2 = \{h_1, ..., h_{12}, c_{12}, c_{13}\}$. L'expression du degré d'imbrication de deux classes $h, h'$ utilisée est $S(h, h') = \frac{M(h \bigcap h')}{M(h)} = \frac{card(h \bigcap h')}{card(h)}$ (cf. définition 2.3). Le tableau 1 donne les éléments utiles permettant de calculer :

- $S(h_{12}, H2) = \frac{S(c_{13}, h_{12}) + S(h_{12}, x_6) + S(h_{12}, x_7)}{3} = \frac{0.286 + 0.5 + 0.5}{3} = 0.428$,
- $S(c_{12}, H1) = \frac{S(h_{13}, c_{12}) + S(c_{12}, x_6) + S(c_{12}, h_{11})}{3} = \frac{0.857 + 0.166 + 0.833}{3} = 0.618$,
- $M(H1) = 13 + S(c_{12}, H1) = 13,618$,
- $M(H2) = 13 + S(h_{12}, H2) = 13,428$,
- $M(H1 \cap H2) = \Sigma_{i=1}^{11} S(h_i, H1) + S(h_{12}, H2) + S(c_{12}, H1) + S(c_{13}, H1) = 13.046$.

On trouve $S(H1, H2) = \frac{M(H1 \cap H2)}{M(H1)} = 0,95$ et $S(H2, H1) = \frac{M(H1 \cap H2)}{M(H2)} = 0,97$.

**TAB. 1.** *Eléments pour le calcul des degrés d'imbrication mutuelle de H1 et de H2*

| $Y$ | $\underline{y}$ | $\overline{Y}$ | $S(\overline{Y},Y)$ | $S(Y,\underline{y})$ | $S(Y,H2)$ | $Y$ | $\underline{y}$ | $\overline{Y}$ | $S(\overline{Y},Y)$ | $S(Y,\underline{y})$ | $S(Y,H1)$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $h_8$ | $c_8$ | $c_8$ | 1 | 1 | 1 | $c_8$ | $h_8$ | $h_8$ | 1 | 1 | 1 |
| $h_9$ | $c_9$ | $c_9$ | 1 | 1 | 1 | $c_9$ | $h_9$ | $h_9$ | 1 | 1 | 1 |
| $h_{10}$ | $c_{10}$ | $c_{10}$ | 1 | 1 | 1 | $c_{10}$ | $h_{10}$ | $h_{10}$ | 1 | 1 | 1 |
| $h_{11}$ | $c_{11}$ | $c_{11}$ | 1 | 1 | 1 | $c_{11}$ | $h_{11}$ | $h_{11}$ | 1 | 1 | 1 |
| $h_{12}$ | $x_6$ | $c_{13}$ | 0,286 | 0,5 | | $c_{12}$ | $h_{11}$ | $h_{13}$ | 0,857 | 0,833 | |
| $h_{12}$ | $x_7$ | | | 0,5 | 0,428 | $c_{12}$ | $x_6$ | | | 0,166 | 0,618 |
| $h_{13}$ | $c_{13}$ | $c_{13}$ | 1 | 1 | 1 | $c_{13}$ | $h_{13}$ | $h_{13}$ | 1 | 1 | 1 |

## 4. Conclusion

Dans ce papier, nous avons introduit la notion de hiérarchie semi-floue en associant une fonction d'appartenance floue à chaque classe d'une hiérarchie usuelle. Par ailleurs, nous avons proposé un indice permettant de mesurer le degré d'imbrication entre deux hiérarchies classiques ou semi-floues.

## 5. Bibliographie


L. A. Zadeh (1965). Fuzzy sets, Basic notion in fuzzy set theory.Inform. and control 8, pages 338-358. MathScinet,1965.

J. P. Benzécri et Collaborateurs (1976) L'Analyse des données :1 La Taxinomie, Dunod,1976.

B. Kosko (1992) Neuronal networks and fuzzy systems. Prentice-hall International Editions,1992.

J. Bezdec(1981). Pattern Recognition with Fuzzy Objective Functions. New York : Plenum, 1981.

# Cartographie d'un corpus de domaine médical

**Thibault Roy (1), Aurélie Névéol (2, 3)**

*(1) Laboratoire GREYC UMR 6072 CNRS*
*Université de Caen / Basse-Normandie*
*Boulevard Maréchal Juin 14032 Caen Cedex*
*thibault.roy@info.unicaen.fr*

*(2) Equipe CISMeF*
*aneveol@insa-rouen.fr*

*(3) NLM, 3600 Rockville Pike, Bethesda, MD 20894*
*Etats-Unis*

*RÉSUMÉ. Cet article présente les premiers résultats d'une expérience menée dans le cadre de la cartographie d'un corpus de documents médicaux. Notre objectif est de proposer à des experts dans le domaine de la santé (médecins, documentalistes, etc.) des vues globales sur des corpus situés dans ce domaine. Afin de fournir de telles vues, nous utilisons la plate-forme ProxiDocs de cartographie et de catégorisation de corpus permettant de prendre en considération les particularités du domaine médical. Les cartes ainsi construites à partir du corpus d'étude permettent de visualiser des proximités et des regroupements entre documents du corpus.*

*MOTS-CLÉS : Cartographie de Corpus, Catégorisation de Textes, Terminologies Médicales, Indexation*

## 1 Introduction

Cet article présente les résultats d'une expérience récente menée dans le cadre de la cartographie d'un corpus de documents du domaine médical. Après avoir présenté la problématique générale de la représentation visuelle d'ensembles documentaires, nous précisons nos objectifs à travers la cartographie d'un corpus de documents médicaux. La deuxième partie détaille le corpus d'étude ainsi que la méthode de cartographie exploitée. La troisième partie présente la carte construite à partir du corpus d'étude et une analyse détaillée de cette carte. Enfin, nous concluons sur les résultats obtenus.

## 2 Cadre de travail

Le nombre de documents électroniques textuels produits et échangés chaque jour ne cesse de croître. Afin d'isoler les principales informations contenues dans des ensembles de documents, il peut être intéressant d'en proposer des représentations globales. Depuis quelques années, des outils d'analyse textuelle exploitent une technique de visualisation particulière appelée cartographie. À la manière d'une carte routière mettant en évidence des villes et des routes les reliant, une carte d'un ensemble de données textuelles met en évidence des proximités sémantiques et des liens entre entités textuelles, tels des mots, des textes, etc. Depuis 2001, les métamoteurs de recherche cartographiques KartOO (http://www.kartoo.com) et MapStan (http://www.mapstan.net) sont disponibles sur l'Internet. De nombreux logiciels dédiés à l'analyse de données textuelles proposent également des résultats d'analyses

sous forme de cartes. Parmi ces logiciels, nous pouvons citer Hyperbase d'Etienne Brunet, BI de Michel Kerbaol ou encore Lexico3 de l'équipe CLA2T de Paris III.
Les documents scientifiques dans le domaine de la santé ne sont pas épargnés par l'essor du numérique. Plusieurs projets se donnent alors pour objectif de guider les utilisateurs dans leur recherche d'information en santé. Ainsi, la fondation Suisse HON (Health On the Net – http://www.hon.ch) propose un portail vers une information de santé de qualité dans plusieurs langues européennes. La base documentaire MEDLINE® (http://www.pubmed.gov) recense l'ensemble des publications scientifiques dans le domaine de la santé depuis plusieurs décennies. Depuis 1995, le Catalogue et Index des Sites Médicaux Francophones (CISMeF – http://www.cismef.org) recense des ressources de santé institutionnelles à l'usage des professionnels de santé, des étudiants en médecine et du grand public. Afin de retrouver des informations pertinentes dans de tels ensembles documentaires, les méthodes traditionnelles consistent à interroger des bases documentaires à l'aide de mots-clés fournis aux moteurs de recherche dédiés. Notre objectif est de proposer à des experts de la santé (médecins, documentalistes, etc.) des vues globales sur des ensembles de documents médicaux. De telles vues doivent permettre de localiser les principales informations contenues dans l'ensemble documentaire, mais aussi des similarités et des différences entre documents de l'ensemble. Pour ce faire, nous proposons d'utiliser la plate-forme ProxiDocs de cartographie et de catégorisation de corpus [ROY 05] à laquelle des informations liées au domaine de la santé ont été intégrées.

## 3 Présentation du corpus et de la méthode de cartographie

### *3.1 Corpus d'étude*

Pour cette étude, nous avons travaillé avec un corpus de 70 ressources[1] extraites aléatoirement du catalogue CISMeF dans le cadre de différentes campagnes d'évaluation de systèmes d'indexation automatique. Chaque ressource du corpus de travail comporte une indexation à l'aide de descripteurs du thésaurus MeSH® (Medical Subject Headings). Cette indexation se présente sous la forme d'une liste pondérée de mots-clés ou de paires mot-clé/qualificatif issus du MeSH. La pondération « majeur » dénote les thèmes traités en profondeur dans la ressource, et la pondération « mineur » signale les thèmes traités plus succinctement.

### *3.2 Méthode de construction des cartes*

La catégorisation du corpus en spécialités médicales est effectuée grâce à un outil bibliométrique [DAR 05] utilisant récursivement l'algorithme de catégorisation décrit dans [NEV 04]. Cet algorithme est fondé sur l'indexation MeSH des ressources, et exploite les liens sémantiques existant entre les mots-clés MeSH et les spécialités médicales d'une part, les qualificatifs MeSH et les spécialités médicales d'autre part. Ainsi, chaque descripteur MeSH attribué à une ressource permet de catégoriser la ressource sous la (les) spécialité(s) médicale(s) auxquelles renvoie le descripteur. Par exemple, une ressource indexée avec le mot-clé <*diabète*> relève de la spécialité « endocrinologie ». Le score attribué à « endocrinologie » sera de 100 si <*diabète*> est un thème majeur pour la ressource et de 1 si c'est un thème mineur. À partir du classement des spécialités établi avec la méthode précédente sur le corpus d'étude, nous obtenons une information globale sur cet ensemble.
Les spécialités ainsi classées servent de point de départ à la cartographie de l'ensemble documentaire. Le premier traitement réalisé consiste à attribuer une structure vectorielle à chaque ressource : une ressource est représenté par un vecteur de nombres réels compris dans un espace de dimension égale au nombre de spécialités où chaque coordonnée du vecteur est le score de la spécialité correspondante, l'ordre des coordonnées dans le vecteur étant similaire au classement global des spécialités. Les scores des spécialités de chaque ressource sont déterminés à l'aide de la méthode précédente, mais en ne prenant cette fois-ci en

[1] Afin de rendre compte de la multiplicité des documents électroniques que cela soit du point de vue de leurs formats ou des usages auxquels ils sont destinés, nous utiliserons le terme de « ressource ».

considération que la ressource et non la globalité du corpus. De cette manière, une ressource est représentée de la façon suivante :

$$Vecteur_{Res} = (Score_{Virology}(Res), Score_{Infectiology}(Res), Score_{Virology}(Res), etc.)$$

Si des spécialités apparaissant dans l'ensemble ne sont pas présentes dans la ressource, des valeurs nulles sont placées aux coordonnées correspondantes dans le vecteur. Ce processus est alors répété pour chaque ressource de l'ensemble étudié. Ainsi, un espace de grande dimension où les ressources prennent place a pu être construit.
Dans notre étude, 78 spécialités (sur 126) ont été utilisées pour catégoriser l'ensemble des ressources à l'aide de la méthode précédente ; l'espace des ressources possède donc 78 dimensions. Afin de visualiser graphiquement les documents prenant place dans un tel espace, nous avons choisi d'en réaliser une projection vers un espace à deux dimensions. La plate-forme ProxiDocs permet de réaliser cette opération selon différentes méthodes statistiques (cf. [ROY 04] pour plus de détails sur ces méthodes). Dans cette étude nous avons choisi d'utiliser la méthode de projection de Sammon pour les résultats satisfaisants qu'elle donne dans la projection d'espaces de grande dimension [SAM 69]. Des Analyses en Composantes Principales [BOU 80] et des Analyses Factorielles des Correspondances [BEN 80] ont également réalisées à l'aide de la plate-forme dans cette étude, mais les résultats obtenus se sont révélés moins pertinents.

Entrée : un espace de ressources à n dimensions (n > 2)
Sortie : un espace à deux dimensions où les ressources prennent place

1. Placer chaque ressource aléatoirement dans l'espace d'arrivée à deux dimensions (le placement aléatoire se fait dans $[0,1]^2$).
2. Pour chacune des ressources de l'ensemble : tester si les distances euclidiennes dans l'espace de départ à n dimensions entre la ressource courante et les autres ressources sont respectées dans l'espace d'arrivée à 2 dimensions (une faible constante près fixée empiriquement).
3. Si ce n'est pas le cas, les autres ressources peuvent effectuer un déplacement minimal (valeur du déplacement minimal donnée en entrée de l'algorithme) dans l'espace à 2 dimensions afin de « tendre » vers une situation où les distances entre chacune des ressources sont respectées dans l'espace à 2 dimensions.
4. Reprendre à l'étape 2. jusqu'à ce que les distances entre chaque ressource soient respectées entre l'espace de départ à n dimensions et l'espace d'arrivée à 2 dimensions.

Afin de mettre en évidence des regroupements entre les ressources ainsi projetées, nous avons choisi d'appliquer une Catégorisation Hiérarchique Ascendante (CHA) [BOU 80]. Son fonctionnement dans le cadre de notre étude peut se résumer par les deux étapes suivantes :

Entrée : un espace de ressources prenant place dans un espace à 2 dimensions
Sortie : un ensemble de n groupes de ressources, le nombre n étant choisi empiriquement par l'utilisateur

1. Parmi les entités à catégoriser, chercher les deux entités les plus proches (c'est-à-dire, dont la distance euclidienne est la plus petite) dans l'espace à deux dimensions. Ces deux entités sont ensuite agrégées en un nouveau groupe.
2. Calculer les distances entre le nouveau groupe et les entités restantes. La configuration est alors identique à celle de l'étape 1, hormis que l'on a seulement n-1 entités à classer.

Et ainsi de suite, on cherche de nouveau les deux entités les plus proches, que l'on agrége et ceci jusqu'à obtenir le nombre de groupes choisis par l'utilisateur.
Une fois les étapes de projection et de catégorisation réalisées, nous retournons des représentations graphiques de l'ensemble documentaire que nous appelons des cartes. Ces cartes permettent alors à

l'utilisateur de naviguer sur l'ensemble et de visualiser de façon interactive différentes informations, telles des regroupements entre ressources de l'ensemble. La partie suivante de cet article présente en détails une carte obtenue à partir du corpus d'étude.

## 4 Cartographie du corpus d'étude

Dans cette partie, nous présentons une carte construite selon la méthode détaillée précédemment à partir du corpus catégorisé en spécialités médicales. Cette carte met visuellement en évidence 12 groupes de ressources obtenus par CHA (nombre de groupes choisi empiriquement).

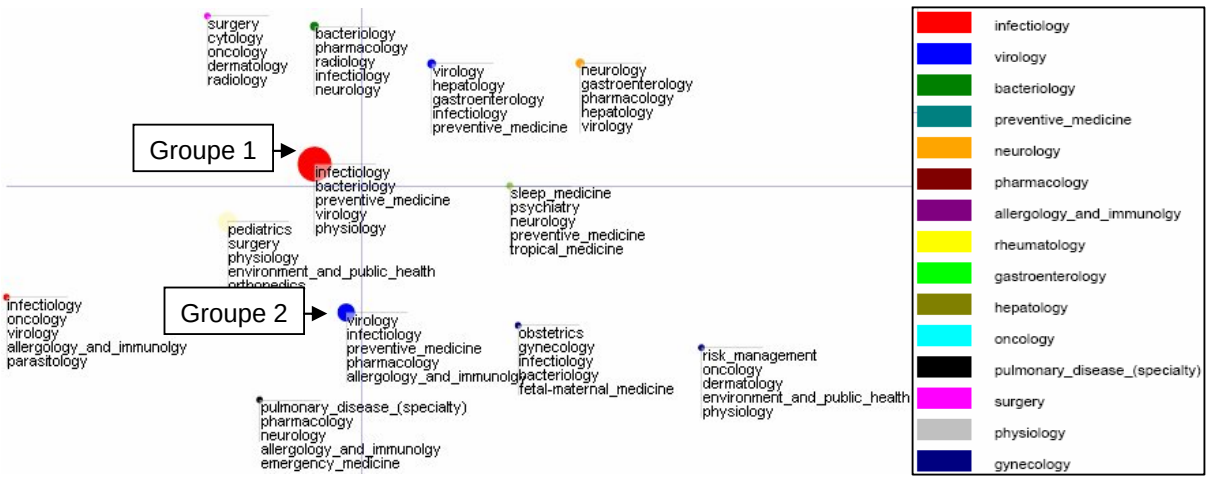



**Figure 1 : Carte du corpus construite avec la plate-forme ProxiDocs. Version interactive de cette carte au format SVG : http://www.info.unicaen.fr/~troy/cismef/**

Chaque groupe de ressources est représenté par un disque de taille proportionnelle à sa cardinalité. La couleur attribuée à chaque disque correspond à sa spécialité majoritaire, c'est-à-dire celle ayant le score le plus élevé dans les ressources du groupe. Une légende attribuant une couleur aux 15 spécialités majoritaires dans les groupes est disponible sur la partie droite de la figure. Chaque groupe est caractérisé par ses cinq spécialités de score le plus élevé. Chaque disque sur la carte est un lien hypertexte vers un rapport détaillé sur les propriétés du groupe. La carte révèle des regroupements entre ressources du corpus d'étude par rapport aux spécialités médicales.

Afin de construire cette carte, nous n'avons pris en considération que des spécialités médicales principales relevant d'un domaine médical (comme *infectiology*, *virology*, *neurology*, etc.). Les spécialités médicales dites transversales, c'est-à-dire des spécialités ne constituant pas un tout et étant applicables aux spécialités médicales principales, n'ont pas été prises en considération dans la cartographie (nous avons par exemple comme spécialités médicales transversales : *therapeutics*, *anatomy*, *diagnosis*, *economics*, *ethics*, *organization and administration*, etc.). La carte ne révèle donc que les domaines médicaux abordés dans le corpus et non les moyens de mettre en œuvre de tels domaines. Dans une étude préliminaire conservant l'ensemble des spécialités au même niveau, nous avions observe une influence significative des spécialités transversales sur la construction des cartes. La prise en considération de ces spécialités, globalement majoritaires dans l'ensemble des groupes, avait pour effet de masquer les thématiques dénotées par les spécialités principales. Il semblait donc plus pertinent de les étudier séparément.

La carte ainsi met en évidence la répartition des spécialités principales dans le corpus d'étude. Le groupe 1 sur la carte possède 36 ressources, les trois spécialités majoritaires dans ce groupe sont *infectiology*, *bacteriology* et *preventive medecine*. Un parcours rapide des ressources de ce groupe révèle qu'elles abordent des thématiques assez variées, certes liées aux spécialités principales, mais sans réelle lien entre les ressources. Au contraire le groupe 2 contenant 11 ressources et possédant comme spécialité majoritaire *virology*, *infectiology* et *preventive medecine* regroupe des ressources toutes étroitement liées au domaine de la virologie (par exemple, sur des ressources traitant du virus de la grippe et des différents vaccins

existants contre ce virus). Ce phénomène se retrouvent dans une très grande majorité des autres groupes de la carte : les ressources abordent des thématiques étroitement liées aux spécialités majoritaires du groupe les contenant.

## 5 Conclusion

Nous avons présenté dans cet article les premiers résultats d'une expérience dédiée à la cartographie d'un corpus du domaine médical. La carte ainsi construite a permis de visualiser des proximités et des différences entre documents. Cette carte a révélé une répartition des spécialités très différentes de celle obtenue lors de l'analyse globale du corpus. Ainsi, des groupes de documents de spécialités majoritaires particulièrement « enfouies » dans le classement global du corpus ont pu être mis en évidence. D'une certaine manière, des signaux faibles dans l'analyse globale du corpus sont ressortis à travers la carte, ceci à l'aide d'une prise en considération d'un niveau d'analyse de granularité différente : le groupe de documents.

## 6 Bibliographie


[BEN 80] Benzecri J.-P., *L'analyse des données - tome 2 : l'analyse des correspondances*, éditions Bordas, 1980.

[BOU 80] Bouroche J.-M., Saporta G., *L'Analyse des Données*, Paris : PUF, 1980.

[DAR 05] Darmoni S.J., Névéol A., Renard J.M., Gehanno J.F., Soualmia L.F., Dahamna B., Thirion B., "A MEDLINE Categorization Algorithm", *BMC*, sous presse, 2005.

[NEV 04] Névéol A., Soualmia L.F., Douyère M., Rogozan A., Thirion B. et Darmoni S.J., "Using CISMeF MeSH "Encapsulated" Terminology and a Categorization Algorithm for Health Resources", International *Journal of Medical Informatics* Vol. 73(1), 57-64, 2004.

[ROY 04] Roy T., Beust P., "ProxiDocs, un outil de cartographie et de catégorisation thématique de corpus", Actes des *Journées internationales de l'Analyse des Données Textuelles*, 978-987, 2004.

[ROY 05] Roy T., "Une plate-forme logicielle dédiée à la cartographie thématique de corpus", Actes de *TALN/RECITAL*, 545-554, 2005.

[SAM 69] Sammon J. W., "A Nonlinear Mapping for Data Structure Analysis", *IEEE Transactions on computers* C-18(5), 401-409, 1969.

# Un algorithme GEM pour le débruitage de signaux

**Allou Samé, Etienne Côme, Latifa Oukhellou, Patrice Aknin**

*Institut National de Recherche sur les Transports et leur Sécurité (INRETS)*
*2, avenue du général Malleret-Joinville*
*94114, Arcueil, France*

*{same,come,oukhellou,aknin}@inrets.fr*

RÉSUMÉ. *Dans le cadre du diagnostic de défauts dans le domaine ferroviaire, cet article propose une méthode pour le débruitage de signaux, mise en œuvre par un algorithme EM généralisé (GEM). La méthode proposée est basée sur un modèle de régression dans lequel le bruit, supposé additif, est distribué suivant un mélange de densités normales. Des simulations menées sur des signaux simulés mettent en évidence de bons résultats de l'algorithme proposé.*

MOTS-CLÉS : *Modèle de mélange, régression, algorithmes EM et GEM, débruitage de signaux, bruit non symétrique*

## 1. Introduction

Dans le cadre du diagnostic du système de transmission voie-machine dans le domaine ferroviaire [FES 01], nous avons été amenés à débruiter puis à paramétriser des signaux d'inspection ferroviaire. La méthode proposée dans cet article pour réaliser simultanément ces deux tâches est basée sur un modèle de régression ; les coefficients de régression étant directement utilisés comme paramètres. Les modèles de mélange [MCL 00] très utilisés dans le domaine de la classification automatique pour modéliser des populations hétérogènes, sont utilisés ici comme distribution du bruit dont la structure physique est complexe. Dans la seconde section, notre modèle de régression est détaillé ; la troisième section présente l'algorithme utilisé pour estimer les paramètres du modèle ; enfin une étude expérimentale menée sur des signaux simulés, dont les paramètres sont connus, est effectuée dans la quatrième section pour évaluer la méthode proposée.

## 2. Bruit additif et modèle de mélange

Nous représentons chaque signal par un échantillon indépendant $((x_1, y_1), \ldots, (x_n, y_n))$ où les variables $x$ et $y$, définies sur $\mathbb{R}$, représentent respectivement la variable indépendante (par exemple le temps) et la variable dépendante (le signal à un instant donné $x$). Les fonctions de régression polynômiales du second degré ont été choisies pour leur cohérence avec notre application. Toute autre classe de fonctions de régression linéaires est utilisable. Un signal observé est donc modélisé par

$$y = ax^2 + bx + c + \varepsilon,$$

où $\varepsilon$ est un bruit additif dont la distribution est indépendante de $x$. Habituellement, $\varepsilon$ est un bruit distribué suivant une densité normale de moyenne nulle (bruit électronique centré par exemple) et l'estimation des paramètres du modèle revient à résoudre un problème classique de moindres carrés. Dans notre application, la présence d'une source supplémentaire de bruit liée au processus de mesure nous a conduit à considérer que le bruit $\varepsilon$ était distribué suivant le mélange de deux densités normales

$$f(\varepsilon) = \pi_1 \mathcal{N}(\varepsilon; 0, \sigma_1^2) + \pi_2 \mathcal{N}(\varepsilon; \mu, \sigma_2^2),$$

où $\mathcal{N}(\cdot;\mu,\sigma^2)$ est la densité normale de moyenne $\mu$ et de variance $\sigma^2$, et $\pi_1, \pi_2$ sont les proportions du mélange vérifiant $\sum_{k=1}^{2}\pi_k = 1$. Les composantes $\mathcal{N}(\cdot;0,\sigma_1^2)$ et $\mathcal{N}(\cdot;\mu,\sigma_2^2)$ du mélange sont relatives aux deux sources de bruit évoquées. La flexibilité bien connue des modèles de mélange permet au modèle de régression considéré de gérer aussi bien des bruits additifs symétriques que non-symétriques [BAR 05] comme c'est le cas dans notre application. La section suivante montre comment les paramètres du modèle proposé peuvent être estimés.

## 3. Estimation des paramètres par un algorithme EM généralisé

Les paramètres du modèle sont estimés par maximisation de la log-vraisemblance qui peut s'écrire

$$L(\boldsymbol{\Psi}; x_1, \ldots, x_n, y_1, \ldots, y_n) = \sum_{i=1}^{n} \log\Big[\sum_{k=1}^{K} \pi_k \mathcal{N}\big(y_i\ ;\ ax_i^2 + bx_i + c_k, \sigma_k^2\big)\Big],$$

où $K = 2$ et $\boldsymbol{\Psi} = (a, b, c_1, c_2, \pi_1, \pi_2, \sigma_1^2, \sigma_2^2)$, avec $c_1 = c$ et $c_2 = c+\mu$. Cette écriture de la vraisemblance montre clairement que le modèle proposé est un mélange de régressions [DES 88] contraint, puisque les coefficients $a$ et $b$ sont les mêmes pour chaque composante du mélange. La log-vraisemblance étant difficile à maximiser directement, l'algorithme EM [DEM 77] est utilisé pour effectuer la maximisation. A partir d'un vecteur paramètre initial $\boldsymbol{\Psi}^{(0)}$, l'algorithme EM consiste alors à répéter les deux étapes suivantes jusqu'à la convergence.

**Etape E (Espérance)** Cette étape consiste à calculer l'espérance $Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(q)})$ de la vraisemblance complétée conditionnellement aux données observées et au vecteur paramètre courant $\boldsymbol{\Psi}^{(q)}$. Elle nécessite principalement de calculer les probabilités a posteriori $t_{ik}^{(q)} = \frac{\pi_k^{(q)}\mathcal{N}(y_i;\, a^{(q)}x_i^2 + b^{(q)}x_i + c_k^{(q)},\, \sigma_k^{2\,(q)})}{\sum_{\ell=1}^{K}\pi_\ell^{(q)}\mathcal{N}(y_i;\, a^{(q)}x_i^2 + b^{(q)}x_i + c_k^{(q)},\, \sigma_\ell^{2\,(q)})}$.

**Etape M (Maximisation)** Cette étape consiste à calculer le vecteur paramètre $\boldsymbol{\Psi}^{(q+1)}$ qui maximise, par rapport à $\boldsymbol{\Psi}$, la quantité $Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(q)})$ qui, dans notre situation, s'écrit

$$Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(q)}) = \sum_{k,i} t_{ik}^{(q)} \log \pi_k - \frac{1}{2}\sum_{k,i} t_{ik}^{(q)}\Big[\log(2\pi) + \log\sigma_k^2 + \frac{(y_i - (ax_i^2 + bx_i + c_k))^2}{\sigma_k^2}\Big].$$

Comme dans la situation classique des modèles de mélange gaussiens, on peut vérifier que les proportions sont données par $\pi_k^{(q+1)} = \sum_{i=1}^{n} t_{ik}^{(q)}/n$. La maximisation par rapport aux paramètres $a$, $b$, $c_k$ et $\sigma_k^2$ ne peut, quant à elle, pas se faire analytiquement. Par contre, il est possible de maximiser $Q$ par rapport à $a$, $b$, $c_k$, pour des paramètres $\sigma_k^2$ fixés, et inversement. Cela nous permet de calculer des paramètres $a^{(q+1)}$, $b^{(q+1)}$, $c_k^{(q+1)}$ et $\sigma_k^{2\,(q+1)}$ garantissant simplement $Q(\boldsymbol{\Psi}^{(q+1)}; \boldsymbol{\Psi}^{(q)}) \geq Q(\boldsymbol{\Psi}^{(q)}; \boldsymbol{\Psi}^{(q)})$. Cette alternative, souvent employée quand l'étape M est difficile à effectuer directement, nous fournit un algorithme EM généralisé (GEM) [DEM 77, MCL 97] dont les propriétés de convergence sont les mêmes que celles de l'algorithme EM.

## 4. Expérimentations sur des signaux simulés

Cette section évalue l'algorithme proposé sur des signaux simulés en termes de précision de débruitage. Chaque signal est simulé suivant une fonction de régression polynômiale de degré deux avec un bruit distribué suivant un mélange de densités gaussiennes. Pour choisir les paramètres de simulation, nous considérons la nouvelle paramétrisation $\alpha(x-\beta)^2 + \gamma$ du polynôme du second degré. Dans cette formulation, la concavité/convexité est réglée par le paramètre $\alpha$, le paramètre $\beta$ contrôle la position de l'axe de symétrie de la fonction de régression et le paramètre $\gamma$ permet de régler la valeur du polynôme pour $x = 0$. Notons que ce dernier paramètre n'a aucune influence sur la qualité des estimations des autres paramètres. Ici, seuls les polynômes de régression concaves ($\alpha < 0$) sont considérés pour leur adéquation avec les signaux réels. La taille de l'échantillon considéré dans toutes les simulations est de $n = 600$.

Trois différents ensembles de simulations sont considérés : le premier analyse l'effet de la concavité du polynôme de régression (ou du paramètre $\alpha$) sur la précision des paramètres ; le second observe les effets du rapport

des variances $\sigma_1^2/\sigma_2^2$ et du rapport des proportions $\pi_1/\pi_2$ sur la qualité des paramètres ; le troisième ensemble de simulations analyse l'effet de la position du centre $\mu$ de la seconde composante du mélange sur l'estimation. Nous avons choisi des moments $\mu$ toujours en liaison avec les connaissances a priori sur la nature de la seconde composante du bruit de notre application. La qualité de l'estimation des paramètres est mesurée par la valeur absolue de l'écart entre la vraisemblance calculée sur le paramètre estimé $\widehat{\mathbf{\Psi}}$ et celle calculée sur le paramètre de simulation $\mathbf{\Psi}_{vrai}$, divisée par la taille d'échantillon $n$, $|L(\mathbf{\Psi}_{vrai}) - L(\widehat{\mathbf{\Psi}})|/n$. Puisque la vraisemblance croît généralement avec la taille de l'échantillon, la normalisation par $n$ permet d'obtenir un critère de comparaison très peu sensible à cette taille. Pour chaque vecteur de paramètres, 20 réalisations de bruit sont générées et l'erreur d'estimation commise est moyennée sur ces 20 simulations (Monte-Carlo). La section suivante détaille les trois ensembles de simulations.

### 4.1. *Effet de la concavité du polynôme de régression sur la qualité des paramètres estimés*

L'effet de la concavité du polynôme de régression sur la qualité de l'estimation est observé en considérant les valeurs décroissantes de $\alpha$ suivantes : -0.0008, -0.0016, -0.002, -0.01, -0.1. Deux situations différentes de symétrie du polynôme de régression correspondant à $\beta = 0.5 \times 600 = 300$ (signal symétrique sur $[1; 600]$) et $\beta = 0.75 \times 600 = 450$ (signal non-symétrique sur $[1; 600]$), sont considérées. Les autres paramètres de simulation, choisis en fonction des situations réelles, sont : $\gamma = 150$, $\pi_1 = 0.75$, $\pi_2 = 0.25$, $\sigma_1^2 = 20$, $\sigma_2^2 = 100$ and $\mu = -10$. La figure 1 montre un exemple de signal simulé symétrique avec la densité de bruit correspondante (la densité mélange est représentée en trait continu et les composantes du mélange en pointillés). La figure 2 montre la courbe de régression estimée et la densité mélange estimée du bruit, correspondant au signal de la figure 1.

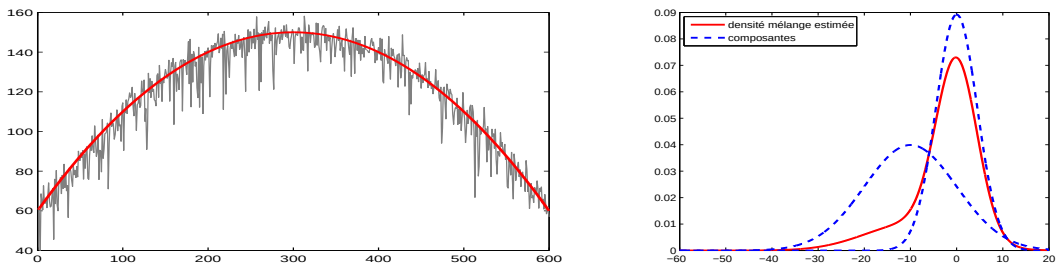



**FIG. 1.** *A gauche : exemple de signal symétrique généré et courbe de régression correspondante ($\alpha = -0.001$, $\beta = 300$, $\gamma = 150$) ; à droite : densité mélange du bruit et composantes associées.*

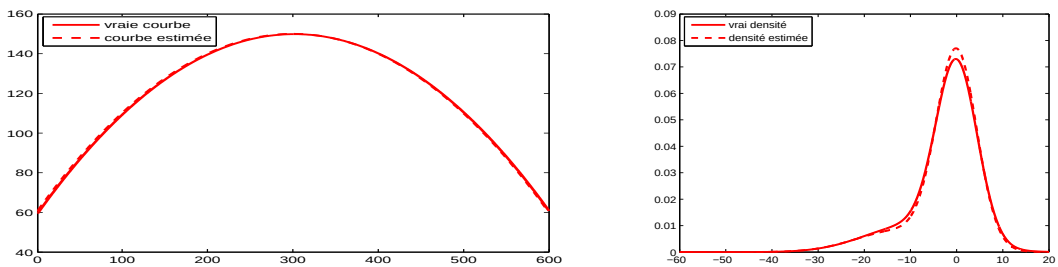



**FIG. 2.** *A gauche : courbe de régression estimée (en pointillés) et vrai courbe de régression (en trait continu) pour le signal représenté dans la figure 1 ; à droite : densité mélange estimée (en pointillés) et vraie densité mélange (en trait continu).*

Le tableau 1 reporte, pour les deux situations de symétrie, l'erreur d'estimation en fonction du paramètre $\alpha$. On remarque que les erreurs sont quasi-identiques pour les deux situations de symétrie ; ce qui prouve que

les coefficients $\alpha$ et $\beta$ n'ont aucune influence sur la qualité de l'estimation. Les bonnes performances de notre algorithme GEM peuvent être attribuées au peu de paramètres du modèle à estimer qui est dû à la dimension peu élevée des données (dimension 1) et au faible nombre de composantes du mélange ($K = 2$).

| $\alpha$ | -0.0008 | -0.0016 | -0.002 | -0.01 | -0.1 |
|---|---|---|---|---|---|
| erreur (signal symétrique) | 0.005 | 0.006 | 0.005 | 0.005 | 0.005 |
| erreur (signal non symétrique) | 0.007 | 0.005 | 0.004 | 0.007 | 0.006 |

**TAB. 1.** *Erreur d'estimation obtenue avec l'algorithme proposé en fonction du degré de concavité du polynôme de régression pour des signaux symétrique et non-symétrique.*

Dans toute la suite, pour simplifier, nous nous restreignons à des polynômes de régression symétriques sur l'intervalle temporel $[1; 600]$, c'est-à-dire à $\beta = 300$.

### 4.2. *Effet du rapport des variances $\sigma_1^2/\sigma_2^2$ et du rapport des proportions $\pi_1/\pi_2$ sur la précision d'estimation*

L'influence du rapport des variances est étudié en observant la qualité des paramètres pour différentes valeurs de $\sigma_1^2/\sigma_2^2$. Les valeurs 1/3, 1/2, 1, 2 et 3 de ce rapport ont été retenues tandis que les autres paramètres de simulation ont été fixés, conformément aux signaux réels, à $\alpha = -0.001$, $\beta = 300$, $\gamma = 150$, $\sigma_1^2 = 20$, $\pi_1 = 0.75$, $\pi_2 = 0.25$, $\mu = -10$. De la même manière, nous considérons les valeurs 1/3, 1/2, 1, 2 et 3 du rapport de proportions pour observer son effet sur la qualité des estimations, les autres paramètres étant fixés à $\alpha = -0.001$, $\beta = 300$, $\gamma = 150$, $\sigma_1^2 = 20$, $\sigma_2^2 = 100$, $\mu = -10$. Le tableau 2 montre clairement que la qualité d'estimation ne varie presque pas avec le rapport des variances ou le rapport des proportions. Les bonnes performances de notre algorithme GEM peuvent être attribuées aux mêmes raisons que celles évoquées dans la section 4.1.

| $\sigma_1^2/\sigma_2^2$ | 1/3 | 1/2 | 1 | 2 | 3 |
|---|---|---|---|---|---|
| erreur | 0.006 | 0.006 | 0.006 | 0.006 | 0.005 |

| $\pi_1/\pi_2$ | 1/3 | 1/2 | 1 | 2 | 3 |
|---|---|---|---|---|---|
| erreur | 0.006 | 0.006 | 0.005 | 0.006 | 0.005 |

**TAB. 2.** *Erreurs d'estimation obtenues en fonction du rapport des variances et en fonction du rapport des proportions.*

### 4.3. *Effet de la position du centre $\mu$ de la densité de la seconde composante du mélange, sur la qualité de la solution*

De façon similaire aux sections précédentes, le tableau 3 fournit les résultats d'estimation obtenus en fonction de la position du centre $\mu$, les autres paramètres étant fixés à $\alpha = -0.001$, $\beta = 300$, $\gamma = 150$, $\pi_1 = 0.75$, $\pi_2 = 0.25$, $\sigma_1^2 = 20$ et $\sigma_2^2 = 100$. On observe qu'il n'y a presque pas d'influence sur la qualité des estimations. Ces bonnes performances peuvent être aussi attribuées au faible nombre de paramètres du modèle à estimer.

| $\mu$ | -20 | -15 | -10 | -5 | 0 |
|---|---|---|---|---|---|
| erreur | 0.006 | 0.006 | 0.006 | 0.007 | 0.006 |

**TAB. 3.** *Erreur d'estimation en fonction de la position du centre $\mu$ de la seconde composante du mélange.*

## 5. Conclusion

Une méthode originale pour le débruitage de signaux est proposée dans cet article. Celle-ci est basée sur un modèle de régression où le bruit, supposé additif, suit un mélange de distributions gaussiennes unidimensionnelles.

Pour estimer les paramètres du modèle, un algorithme EM généralisé (GEM) est proposé. L'étude expérimentale menée sur des signaux simulés montre que l'algorithme fournit d'assez bons résultats d'estimation et par conséquent un débruitage assez fin des signaux. Une étude expérimentale permettant d'étendre, sur des signaux réels, la méthode proposée dans cet article, à une version intégrant un modèle de mélange non nécessairement gaussien est en cours de développement.

## 6. Bibliographie


[BAR 05] BARTOLUCCI L., SCACCIA L., The use of mixtures for dealing with non-normal regression errors, *Computational statistics and data analysis*, vol. 48, 2005, p. 821-834.

[DEM 77] DEMPSTER A. P., LAIRD N. M., RUBIN D. B., Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society Series B*, vol. 39, n$^o$ 1, 1977, p. 1-38.

[DES 88] DESARBO W. S., CRON W. L., A maximum likelihood methodology for clusterwise linear regression, *Journal of classification*, vol. 5(1), 1988, p. 249-282.

[FES 01] FESSANT F., AKNIN P., VILETTE F., Modélisation électrique du circuit de voie, élément du système de transmission voie-machine des TGV, *Revue 3EI*, vol. 27, 2001, p. 46-52.

[MCL 97] MCLACHLAN G., KRISHNAN T., *The EM algorithm and extentions*, Wiley, 1997.

[MCL 00] MCLACHLAN G., PEEL D., *Finite mixture models*, Wiley, 2000.

# Classification des images ISAR pour la reconnaissance des cibles

**Abdelmalek TOUMI, Brigitte HOELTZENER et Ali KHENCHAF**

*Laboratoire E3I2 – EA 3876*
*Ecole Nationale Supérieure d'Ingénieurs des Etudes et Techniques d'Armement (ENSIETA)*
*2 rue François Verny 29806 Brest Cedex 9, France*
*{toumiab, hoeltzbr, Ali.khenchaf@ensieta.fr}*

*RÉSUMÉ. Le domaine de la reconnaissance de formes connaît aujourd'hui une activité importante en raison de la grande panoplie d'applications qu'il permet d'aborder, ceci face à la croissance du nombre et à la complexité des demandes exprimées dans les secteurs porteurs comme les systèmes de défense et de la surveillance pour la sécurité. La problématique générale présentée dans ce papier concerne les systèmes intelligents, dédiés à l'aide à la décision dans le domaine radar. En cela, on se retrouve à un carrefour d'approches aussi variées que spécifiques dans le contexte du processus d'extraction de connaissances à partir de données.*

*MOTS-CLÉS : Classification, ISAR, Reconnaissance des cibles radar.*

## 1 Introduction

La reconnaissance des formes (RF) est la première étape du long processus de compréhension de notre environnement. Par conséquent, on se retrouve dans l'optique des systèmes de l'intelligence artificielle ou encore de la reconnaissance artificielle de forme qui traite de la prise de décision automatique. C'est dans ce contexte que des travaux de recherche importants sont déployés pour permettre aux systèmes radars de réaliser des tâches liées à l'intelligence artificielle (ex. : capture d'expertise sur les données et inférence de règles pour qualifier les données – [TOU 06]), de percevoir l'environnement au-delà du système sensoriel puis de réaliser des étapes de perception de plus en plus fines. Cette présentation, concerne davantage la description des traitements adoptés pour l'extraction des informations à partir des données, avec une focalisation particulière faite sur le problème de l'extraction des primitives. Ces primitives doivent être les mieux adaptées aux images ISAR (*Inverse synthetic Aperture Radar*) reconstruites à partir d'un signal radar brut. La dernière partie présente l'architecture globale de la reconnaissance ainsi que les résultats obtenus au titre de ces travaux. Dans [TOU 06] on trouvera la présentation de la chaîne de traitements adaptée au domaine radar et inspirée du processus d'extraction de connaissances à partir de données (ECD) [FRA 91].

## 2 Données radar expérimentales

Pour la phase expérimentale, nous avons eu recours à des données acquises dans la chambre anéchoïque du laboratoire E3I2 de dimension finie simulant un espace libre. Nous avons utilisé 11 maquettes (A10, F104, F14, F4, Mig29, Tornado, Harrier, F15, F16, F117, F18) à l'échelle 1/48$^{ième}$, Chaque cible est illuminée par une rafale d'émission d'une bande de fréquence de *[11,65 : 18] GHz* avec un pas fréquentiel

de *50 MHz* pour un angle de rotation $\theta$ donné. La simulation est répétée pour chaque angle du domaine angulaire *[-5° : 95°]* avec un pas angulaire de *0.5°*. La polarisation d'émission est horizontale et celle de la réception est verticale. On obtient alors 201 réponses, chacune correspond à un angle donné. La première transformation consiste à reconstruire l'image ISAR à partir des signaux réfléchis par la cible.

### *2.1 Images ISAR*

Une des propriétés intéressante qui est liée à la forme intime de la cible, est sa réflectivité. Elle est définie comme la distribution spatiale de toute la capacité de réflexion concentrée dans quelques régions restreintes, appelées *points brillants* ou *centres de réflexion*. Pour plus de précision, en projetant les points brillants sur l'axe de visée du radar, on obtient les *profils de distance*. Une deuxième projection des points brillants sur l'axe perpendiculaire est effectuée pour conserver l'information transverse et obtenir une image (bidimensionnelle) dite image ISAR (l'information transverse est obtenue par l'analyse spectrale - Doppler- du signal reçu) [MUS 96]. Nous obtenons 162 images ISAR pour chaque cible en utilisant la Transformée de Fourier (TF-2D), d'une taille de 256x256. Chaque image correspond à l'image de la cible observée sous un angle β.

Une première approche pour l'extraction des primitives basée sur la forme de la cible est présentée dans [TOU 06]. Elle utilise la segmentation des images ISAR par la méthode de partage des eaux mosaïque (LPE). Cette approche a montré ses limites en travaillant sur des données volumineuses.

### *2.2 Signature de l'image ISAR*

Des travaux récents sur la classification des images ISAR peuvent être trouvés dans [MUS 96][KIM 05]. Il est à noter que dans le cas des images ISAR, les paramètres de reconstruction sont liés au mouvement de la cible imprévisible, entraînant une variation en rotation et changement d'échelle. Pour proposer des descripteurs invariants à la translation et au changement d'échelle, nous avons procédé par projection de l'image ISAR dans un plan polaire pour obtenir une nouvelle image appelée *Image Polaire* (voir figure 1). Ceci a été réalisé à partir de l'algorithme *«Polaire»* suivi de la projection sur l'axe-r noté $I_r(r)$ et de la projection sur l'axe-$\theta$ noté $I_\theta(\theta)$ :

*Algorithme-Polaire*

$$I_P(r_m,\theta_n) = I(x_k,y_k)$$
$$(x_k,y_k) = (x_0,y_0) + (r_m\cos\theta_n + r_m\sin\theta_n)$$
$$m = 1,...,N_r \; et \quad n = 1,...,N_\theta . et \quad k = 1,...,N_r N_\theta$$
$$où \qquad r_m = R_{\min} + (m-1)\Delta r \qquad et \quad \theta_n = -\pi + (n-1)\Delta\theta$$
$$\Delta r = \frac{R_{\max} - R_{\min}}{N_r - 1}; et \quad \Delta\theta = \frac{2\pi}{N_\theta - 1}$$

Et

$$I_r(r) = \int_{-\pi}^{\pi} I_p(r,\theta)d\theta \approx \sum_{n=1}^{N_\theta} I_P(r_m,\theta_n).$$
$$I_\theta(\theta) = \int_{R_{\min}}^{R_{\max}} I_p(r,\theta)dr \approx \sum_{m=1}^{N_r} I_P(r_m,\theta_n).$$

Avec

$I(x_i,y_i)$ : image ISAR

$I_P\ (r_m,\ \theta_n)$ : image polaire reconstruite

Pour vérifier l'invariance de l'image polaire au changement d'échelle et à la rotation, nous avons effectué une rotation de $\pi/4$ et un changement d'échelle de $1/\sqrt{2}$ de l'image ISAR originale présentée dans la figure 1.a (on a choisi le changement d'échelle qui correspond à la rotation pour garder une même taille initiale de l'image après rotation). Rappelons ici, que la technique d'interpolation choisie est l'interpolation linéaire pour réduire l'échelle de l'image.

Nous pouvons constater, le changement entre les deux vecteurs de projection $I_r(r)$ entre l'image originale (cf. figure 1.a) et l'image résultat (cf. figure 1.b). Ceci s'explique par l'effet produit par la méthode d'interpolation utilisée pour effectuer le changement d'échelle de l'image ISAR originale. Par contre, la rotation est représentée par un décalage bien apparent spécialement sur $I_\theta(\theta)$ et l'image polaire $I_P\ (r,\theta)$ ( voir figure 1.b) qui correspond à une rotation de π/4 qu'on a effectuée. Pour assurer une invariance à cette rotation, les descripteurs de Fourier notés $df_\theta$ sont calculés à partir de $I_\theta(\theta)$. Seulement la moitié des descripteurs est prise en compte. Finalement, dans la base d'apprentissage, une image ISAR

*k* d'une cible *P* sera représentée par la *signature polaire* composée par quatre vecteurs *V1, V2, V3 et V4* où *V1* est le vecteur $I_r$ ( correspond à la projection de l'image polaire sur l'axe des rayons-*r)*, *V2* est le vecteur $df_\theta$ ( vecteur des descripteurs de Fourier), *V3* est le vecteur $I_\theta$, et *V4* est l'image polaire compressée par l'ACP- $ACP(I_p(r,\theta))$.

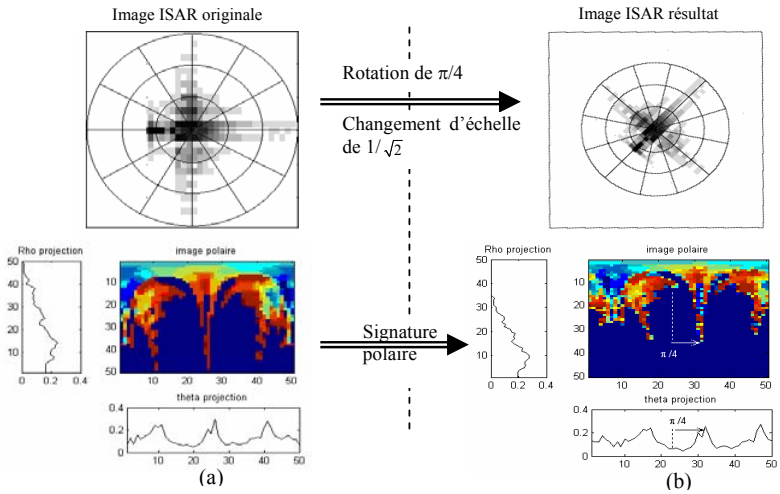



**Figure 1.** *a- Image ISAR d'une cible aérienne et sa signature polaire. b- Image ISAR avec un changement d'échelle et de rotation et sa signature polaire correspondante,* $(N_r = N_\theta = 50, R_{\min} = 0, R_{\max} = 50)$.

## *2.3 Architecture de système de reconnaissance*

Dans l'objectif de construire une base d'apprentissage la plus représentative de chaque cible, l'idée est d'effectuer une classification non supervisée pour chaque classe (cible). Les images qui représentent les centres des sous-classes, sont sélectionnées pour constituer la base d'apprentissage. Nous avons choisi pour cela, d'utiliser la méthode de *Ward* qui opère par classification hiérarchique ascendante (CAH) [WAR 63]. D'autres techniques de classification non supervisée [ASS 05] peuvent êtres appliquées comme K-means [HAR 79] ou les cartes auto-organisatrices de Kohonen [BAL 03]. L'architecture de notre système de reconnaissance se base quant à lui, sur trois classifieurs (voir figure 2), chacun d'entre eux se focalise sur une partie de la *signature* de chaque image. Le premier classifieur intègre le calcul des coefficients de corrélation normalisés $C1(p,k)$ entre $I_r$ de l'image requête (image inconnue) et le vecteurs *V1* de chaque image *k* de la cible *p* de la base d'apprentissage avec $k = 1,..,K_p$ et $p = 1,..,P$ où $P$ et $K_p$ sont respectivement, le nombre total des cibles et le nombre total des

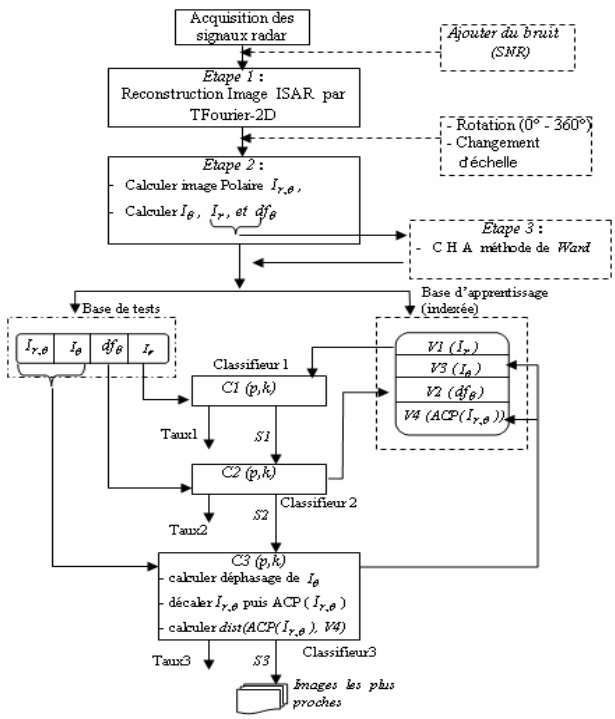



**Figure 2**- *Architecture de la reconnaissance.*

images de la cible *p* dans la base d'apprentissage. Les $\sum_{i=1}^{P} K_i$ coefficients calculés par le classifieur *C1* sont ensuite ordonnés par ordre *décroissant* et l'index des images correspondant à cet ordre est récupéré. Enfin, seulement un pourcentage $\lambda\%$ de cet index est envoyé au classifieur *C2*. Nous notons *S1* le vecteur des indexes envoyé par le premier classifieur vers le second classifieur *C2*. L'idée ici est fondée sur l'hypothèse que l'image la plus proche de l'image requête se trouve parmi les $\lambda\%$ des images de la base d'apprentissage dont les coefficients de corrélation sont les plus grands. Ensuite, le classifieur *C2* calcule les coefficients de corrélation normalisés $C2(p,k)$ sur le vecteur des descripteurs de Fourier $df_\theta$ entre une image requête et les images de l'index *S1* (i.e. $(p,k) \in S1$). Les coefficients $C2(p,k)$ sont ensuite ordonnés par ordre *décroissant* et de la même manière, seulement les premières $\eta\%$ des images de nouvel index sont retenues. Le vecteur des indexes, noté *S2,* est envoyé au classifieur *C3*. Dans le dernier classifieur, on commence par aligner l'image polaire (image requête) avant de la compresser. En effet, on calcule la valeur du décalage entre les deux vecteurs $I_\theta$ et *V3* pour décaler l'image polaire requête. Une approche systématique qui estime la valeur du décalage de translation (*shifts translation*) entre deux vecteurs $I_\theta$ et *V3* est le « *matching score* ». Cette approche est définie comme la valeur maximale des coefficients de corrélation normalisés de tous les décalages linéaires entre deux vecteurs [LI 93]. Par la suite, une mesure de similarité (*C3(p,k),* avec $(p,k) \in S2$) via la *distance Euclidienne* est calculée entre une image requête alignée puis compressée et une images référencée dans S2. Le vecteur *de distance C3(p,k)* est ordonné par ordre *croissant*. Seules les premières images qui correspondent aux plus petites distances référencées dans *C3* sont affichées (l'utilisateur peut intervenir pour changer le nombre d'images à afficher).

## 3 Résultats

Pour les tests réalisés, nous avons utilisé 11 maquettes (chaque maquette correspond à une cible). La base de données pour les expérimentations est constituée de 1782 images ISAR. La base de données est divisée en deux bases, une de test et une d'apprentissage (base de test est complémentaire de la base d'apprentissage). Dans l'ensemble des résultats, $\lambda = 30\%$ et $\eta = 10\%$. Nous avons obtenu un taux de reconnaissance proche de 100%. Pour tester la robustesse de notre architecture, nous avons bruité le signal brut avant la reconstruction des images ISAR. Par la suite, nous avons effectué une rotation aléatoire (entre 0° et 360°) et un changement d'échelle entre $[1:\sqrt{2}]$ pour toute la base de données comme il est illustré par la figure 2. Les résultats obtenus sont présentés dans le tableau 1 en fonction du rapport signal/bruit (SNR) et des plages (a-b) de nombre de classes. Selon les plages (a-b) de nombre de partition, une meilleure partition dans cet intervalle a été obtenue en utilisant la méthode de Ward. Ce qui assure une plus petite inertie intraclasse et une plus grande inertie interclasse. Les taux de bonne classification sont calculés à partir de la matrice de confusion de l'ensemble des classes.

| *Taux* | | ***Taux 1 (%)*** | ***Taux 2 (%)*** | ***Taux 3 (%)*** | ***Taille de la base d'apprentissage*** |
|---|---|---|---|---|---|
| **(a-b)** | **SNR** | | | | |
| 17-20 | + 5 | 72,555 | 80,811 | 86,294 | 202 images |
| | -1 | 66,510 | 73,337 | 80,883 | 200 images |
| | -5 | 59,952 | 60,024 | 71,398 | 203 images |
| 17-23 | + 5 | 74,158 | 81,446 | 87,245 | 210 images |
| | -1 | 67,329 | 74,152 | 82,007 | 209 images |
| | -5 | 60,710 | 60,290 | 72,041 | 214 images |

**Tableau 1**- *Résultats de la classification.*

## 4 Conclusion

Tout système de reconnaissance de forme est fortement corrélé aux primitives caractérisant les formes. Dans ce sens, nous avons proposé la projection de l'image ISAR dans le plan polaire afin de reconstruire une nouvelle image appelée image polaire. A partir de cette dernière et afin de réduire la dimensionnalité de l'espace représentatif de l'image polaire, la projection sur l'axe-$\theta$ et l'axe-$r$ ainsi qu'une compression via l'ACP ont été réalisées. L'architecture hiérarchique a permis une classification raffinée sur trois niveaux avec le souci de réduire les temps calcul du système.
Pour tenir compte des conditions réelles d'application, nos futurs travaux s'orientent vers l'intégration d'indicateurs voire de modèles environnementaux (ex. : fouillis terrestre, de mer) afin d'étudier la robustesse et prévoir des modes d'adaptation du système à l'environnement.

## 5 Bibliographie


[ASS 05] ASSELIN DE BEAVILLE J. P., KETTAF F. Z., " *Bases théoriques pour l'apprentissage et la décision en reconnaissance des forme* ", Cépaduès, 2005.

[BAL 03] BALMAT J.F., LAFONT F., " Multi-model architecture supervised by Kohonen map ", *in: Sciences of Electronic, Technology of Information and Telecommunication SETIT'03*, 2003, p. 98-104, Sousse, Tunisia.

[FRA 91] FRAWLEY W. J., PIATETSKY-SHAPIRO G., MATHEUS C. J., " Knowledge Discovery in Databases: An Overview. Knowledge Discovery in Databases " , 1991, p. 1-30.

[HAR 79] HARTIGAN J. A., WONG M. A., '' A -Means Clustering Algorithm ", *Journal of Applied Statistics*, vol. 28, 1979, p. 100-108.

[KIM 05] KIM K-T., SEO D-K., KIM H-T., " Efficient Classification of ISAR Images ", *IEEE trans on Antennas and Propagation*, vol. 53, no. 5, 2005, p.1611-1621.

[LI 93] Li H. J., YANG S. H., " Using range profiles as features vectors in identify aerospace objects ", *IEEE tans. Antennas propagation.*, vol 41, no°3, 1993, p. 261-268.

[MUS 96] MUSMAN S., KERR D., BACHMANN C., " Automatic recognition of ISAR ship images ", *IEEE Trans.Aerospace Electron*, Syst., vol 32, n° 4, 1996, p. 1392-1404.

[TOU 06] TOUMI A., HOELTZENER B., KHENCHAF A., " Préparation des données Radar pour la reconnaissance/identification de cibles aériennes ", *EGC06, RNTI,* Vol 2, 2006, p. 675-680, Lille.

[WAR 63] WARD J. H., " Hierarchical Grouping to Optimize an Objective Function ", *Journal of American Statistical Association*, no. 64, 1963, p. 236-244.

# Models for Clustering Data: Least-Squares Fitting Approach


## Vichi Maurizio

*University of Rome "La Sapienza"*
*Dept. of Statistics, Probability and Applied Statistics.*
*P.le A. Moro, 5*
*I-00185 ROME, Italy*



*RÉSUMÉ. Dans ce papier, à partir d' un tableau de dissimilarités, on étudie deux modèles pour la recherche d'une partition des objets en classes homogènes. On utilise l'approche des moindres carrés pour estimer les paramètres des modèles proposés.*



*MOTS-CLÉS : classification, moindres carrés.*


## 1 Introduction

Let $O = \{o_1,\ldots,o_I\}$ be a finite set on $I \geq 1$ units (objects) and suppose that a $(I \times I)$ dissimilarity matrix $\mathbf{D}=[d_{il}]$ on $O$ has been observed. $\mathbf{D}$ satisfies the usual properties: (*i*) $d_{il} \geq 0$, $(i,l=1,\ldots,I)$; (*ii*) $o_i=o_l \Rightarrow d_{il} = 0$ for $(i,j=1,\ldots,I)$; (*iii*) $d_{il} = d_{li}$ for $(i,l=1,\ldots,I)$.
The cluster analysis or unsupervised classification of objects through the matrix $\mathbf{D}$ is, in this paper, seen as the statistical modeling problem of fitting an expected *classification model* (e.g., partition, hierarchy, pyramid, etc.) - which is one-to-one associated to a specific dissimilarity *classification matrix* $\mathbf{D}_c$.- to the dissimilarity data $\mathbf{D}$. Formally the cluster analysis problem can be statistically stated as follows

$$\mathbf{D} = \mathbf{D}_c + \mathbf{E}, \tag{1}$$

where $\mathbf{E}$ is the error matrix describing the part of the observed dissimilarity matrix $\mathbf{D}$ which is not explained by the classification matrix $\mathbf{D}_c$. It represents the extent to which an observed dissimilarity matrix $\mathbf{D}$ differs from its expected classification model represented by $\mathbf{D}_c$. In order to complete the model description the characteristics of the matrix $\mathbf{E}$ must be specified. Customary specification is that the expected value of $e_{il}$ is supposed zero. For this paper, the classification matrix corresponding to a partition will be discussed and the model free least-squares estimation method will be adopted.

## 2 Notation

For the convenience of the reader, the terminology used in this paper is listed here:

| | |
|---|---|
| $I, P$ | number of objects; number of classes of the partition of objects; |
| $O = \{o_1,\ldots,o_I\}$ | set of $I$ objects to be classified or simply its set of indices; |
| $C$ | partition $C=\{C_1\ldots,C_p,\ldots,C_P\}$ of $O$, where $C_p$ is the $p^{th}$ subset of objects of $O$ or simply the subset of indices pertaining to these objects; |

| | |
|---|---|
| $\mathbf{D}=[d_{il}]$ | $(I \times I)$ matrix specifying dissimilarities between objects $(o_i, o_j)$, $i, j=1,\ldots,I$; |
| $\mathbf{M}=[m_{ip}]$ | $(I \times P)$ membership matrix specifying a *partition* $C$ of objects in $P$ classes, where $m_{ip}=1$ if the $i^{th}$ object $o_i$ belongs to $p^{th}$ class $C_p$, $m_{ip}=0$ otherwise. Therefore, matrix **M** is constrained to be binary and row-stochastic, i.e., with one nonzero element per row; |
| $\mathbf{P}=[p_{il}]$ | $(I \times I)$ matrix specifying an 2-ultrametric distance between objects $(o_i, o_l)$, $i, l=1,\ldots,I$, i.e., an ultrametric matrix with at most 2 different off-diagonal values; |
| $\mathbf{E}=[e_{il}]$ | $(I \times I)$ matrix of error terms. |

## 3 The partition model and the associated dissimilarity classification matrix

### *3.1 Partitioning Model with Equal Heterogeneity and Isolation*

A classification matrix specifying a partition with equal heterogeneity and isolation in now discussed. To any *partition C*, defined as a set of disjoint subsets of *O* such that their union is *O* itself, a (2-*Ultrametric*) *classification matrix* $\mathbf{D}_c=\mathbf{P}$ can be associated. In fact, **P** is an ultrametric matrix with off-diagonal elements that can assume one of at most 2 different values $0 < \alpha_1 \leq \alpha_2$. Formally: $\mathbf{P}=[p_{il}]$, $p_{il} \geq 0$, $p_{il} = p_{li}$, $p_{ij} \leq \max(p_{ik}, p_{lk}) \; \forall \; (i,l,k)$ and $p_{il} \in \{0, \alpha_1, \alpha_2\} \; \forall \; (i,l)$. Triplets of elements of **P** are of one of the three types $(\alpha_1, \alpha_1, \alpha_1)$, $(\alpha_1, \alpha_2, \alpha_2)$, $(\alpha_2, \alpha_2, \alpha_2)$, respectively, corresponding to three, two, none objects in a same class. These triplets in turn satisfy the ultrametric inequality.

Matrix **P** can be written as a function of $\alpha_1$, $\alpha_2$ and of a $(I \times I)$ binary matrix $\mathbf{S} \equiv [s_{il}]$ specifying a particular similarity matrix, where entries $s_{ii} = 1$, $(i=1,\ldots,I)$; $s_{il} = 1$ (resp., 0), if objects $i^{th}$ and $l^{th}$ belong (resp., do not belong) to the same class of the partition $C$ $(i, l = 1, ..., I)$. Thus,

$$\mathbf{P} = \alpha_2(\mathbf{11'} - \mathbf{S}) + \alpha_1(\mathbf{S}-\mathbf{I}), \tag{2}$$

where **1** is a vector of ones and **I** is the identity matrix of order *I* and $0 < \alpha_1 \leq \alpha_2$. Now, since $\mathbf{S} = \mathbf{MM'}$, where **M** is a binary membership matrix specifying a partition *C*, matrix **P** can be rewritten,

$$\mathbf{P} = \alpha_2(\mathbf{11'} - \mathbf{MM'}) + \alpha_1(\mathbf{MM'}-\mathbf{I}). \tag{2'}$$

This classification matrix identifies the most parsimonious partitioning model where the within clusters dissimilarities of the classes of the partition *C* are supposed all equal to $\alpha_1$ while the between clusters dissimilarities are hypothesized all equal to $\alpha_2$.

The value $\alpha_1$ measures the *heterogeneity* or lack of cohesion of objects in each class of *C*; while $\alpha_2$ evaluates the *isolation* between classes of *C*. Therefore this model supposes equal heterogeneity in each class of *C* and equal isolation between classes. Of course, it is suitable that $\alpha_1 \leq \alpha_2$ as required in a 2-ultrametric matrix.

Let $C$ be the set of partitions of *O* and let $D_{\mathbf{P}}$ be the set of 2-ultrametric matrices **P** with fixed $\alpha_1$ and $\alpha_2$, thus, Vicari & Vichi (2000), show that there exists a *bijection* between $C$ and $D_{\mathbf{P}}$. Furthermore, any 2-Ultrametric matrix can be represented by a dendrogram with at most two levels (height of the tree), briefly named 2-dendrogram.

### *3.2 Partitioning Model with Different Heterogeneity and Isolation*

In the previous partitioning model equal heterogeneity or lack of cohesion in the classes and equal isolation between classes has been hypothesized. In this section different heterogeneity within clusters and isolation between clusters is hypothesized in the partition and the corresponding classification matrix is discussed. To any *partition C* of *O* a dissimilarity classification matrix $\mathbf{D}_c=\mathbf{Q}$ can be associated, where objects $o_i \in C_p$ have associated values ${}_{\mathrm{W}}d_{pp}>0$ of the diagonal matrix $\mathbf{D}_{\mathrm{W}}$; while objects $o_i \in C_p$ and $o_l \in C_h$ $(p \neq h)$, have associated values ${}_{\mathrm{B}}d_{ph}>0$ of $\mathbf{D}_{\mathrm{B}}$. Each value ${}_{\mathrm{W}}d_{pp}$ of $\mathbf{D}_{\mathrm{W}}$ is a measure of *heterogeneity* or lack of

cohesion of objects in the class $C_p$ of $C$; while each values ${}_{\mathrm{B}}d_{ph}$ is a measure of isolation between classes $C_p$ and $C_h$. Therefore, a dissimilarity classification matrix is defined as follows

$$\mathbf{Q} = \mathbf{M}\mathbf{D}_{\mathrm{B}}\mathbf{M}' + \mathbf{M}\mathbf{D}_{\mathrm{W}}\mathbf{M}' - \mathrm{diag}(\mathbf{M}\mathbf{D}_{\mathrm{W}}\mathbf{M}'), \tag{3}$$

where $\mathbf{D}_{\mathrm{B}}$ is a $(P \times P)$ "core" dissimilarity matrix specifying the expected dissimilarity between pairs of clusters of the partition $C$, $\mathbf{D}_{\mathrm{W}}$ is a diagonal matrix of order $P$ specifying the expected dissimilarity within each cluster. Matrices $\mathbf{D}_{\mathrm{B}}$ and $\mathbf{D}_{\mathrm{W}}$ control the heterogeneity and isolation of the partition $C$.
Therefore, the dissimilarity partitioning matrix $\mathbf{Q}$ identifies a more flexible partition, where both the expected within clusters dissimilarities and the expected between clusters dissimilarities may differ between clusters. If matrix $\mathbf{D}_{\mathrm{B}} = \alpha_2 (\mathbf{1}_P\mathbf{1}'_P - \mathbf{I}_P)$ and matrix $\mathbf{D}_{\mathrm{W}} = \alpha_1\mathbf{I}_P$, with $\alpha_1 \leq \alpha_2$, then matrix (3) coincides with matrix (2′), i.e., $\mathbf{P}=\mathbf{Q}$, (where $\mathbf{1}_P$ and $\mathbf{I}_P$ is a vector and a square matric of order $p$).
Matrix $\mathbf{Q}$ specifies a partition $C$ in "*well-structured clusters*" if the largest heterogeneity of a class of $C$ is smaller of equal to the smallest isolation between two classes of $C$, i.e., formally

$$max\{{}_{\mathrm{W}}d_{pp} \in \mathbf{D}_{\mathrm{W}}\} \leq min\{{}_{\mathrm{B}}d_{ph} \in \mathbf{D}_{\mathrm{B}} : (p \neq h)\}. \tag{4}$$

Given two matrices $\mathbf{D}_{\mathrm{W}}$ and $\mathbf{D}_{\mathrm{B}}$ specifying a partition $C$ in "*well-structured clusters*", if $\mathbf{D}_{\mathrm{B}}$ is an ultrametric matrix with at most $P$-1 different values, then matrix $\mathbf{Q}$ is an ultrametric matrix with at most $2P$-1 levels (internal nodes of the dendrogram).

## 4 LS estimation of the partitioning models

The semi-parametric least-squares estimation of model (1) when the classification matrix is a 2-ultrametric matrix corresponds to find $\mathbf{P}$ that minimizes the following quadratic constrained problem

$$F(\mathbf{P}) = \| \mathbf{D} - \mathbf{P} \|^2$$
subject to $\mathbf{P}$ to be 2-Ultrametric matrix. [P1]

If in equation (2), $\alpha_2 = 1$ and $\alpha_1=0$, then matrix $\mathbf{P}$ becomes: $\mathbf{P} = \mathbf{1}\mathbf{1}' - \mathbf{S}$. In this case, it is interesting to see that problem [P1] is equivalent to the linear 0/1-integer programming *clique-partitioning* problem (Reigner 1965). If integer variables $s_{il}$ are replaced by continuous variables: $0 \leq s_{il} \leq 1$, $(1 \leq i < l \leq I)$ the relaxation of [P1], (with $\alpha_2 = 1$ and $\alpha_1=0$), is a linear programming problem which solution, when all $s_{il} \in \{0, 1\}$, is also the solution of [P1]. In practice, it appears that the relaxation of [P1] often, but not invariably, has a 0/1 solution (Grötschel and Wakabayashi 1989). There can be more than one optimal solution, and different solutions can be identified by solving a series of linear programming problems in which different small random quantities are added to the right hand sides of inequalities in [P1]. If the solution that is obtained is not integer, more elaborate algorithms are required to provide heuristic solutions to the clique-partitioning problem; several algorithms of this type are reviewed by Hansen *et al*. (1994).
Including expression (2′) in problem [P1] and rewriting constraints on the binary matrix $\mathbf{M}$, problem [P1] needs to minimize, with respect to $\alpha_1$, $\alpha_2$ and $\mathbf{M}$,

$$F(\alpha_1, \alpha_2, \mathbf{M}) = \|\mathbf{D} - \alpha_2(\mathbf{1}\mathbf{1}' - \mathbf{M}\mathbf{M}') - \alpha_1(\mathbf{M}\mathbf{M}' - \mathbf{I})\|^2$$
subject to [P2]
$0 < \alpha_1 \leq \alpha_2$;
$m_{ip} \geq 0$ $(i=1,\ldots,I;\ p=1,\ldots,P)$, $\sum_{p=1}^{P} m_{ip} = 1$ $(i=1,\ldots,I)$.

Problem [P2] can be solved considering a coordinate descent algorithm, which alternates between the update of $\alpha_1$, $\alpha_2$ and $\mathbf{M}$. The update of $\alpha_1$ and $\alpha_2$ are given by

$$\alpha_1 = tr\mathbf{M}'\mathbf{D}\mathbf{M}/tr(\mathbf{M}\mathbf{M}'-\mathbf{I})^2 = \frac{2\sum_{p=1}^{P}\sum_{\substack{i,j\in C_p \\ i<l}} d_{il}}{\sum_{p=1}^{P} n_p^2 - I}, \quad \alpha_2 = (tr\mathbf{1}'\mathbf{D}\mathbf{1} - tr\mathbf{M}'\mathbf{D}\mathbf{M})/tr(\mathbf{1}'\mathbf{1}-\mathbf{M}\mathbf{M}')^2 = \frac{2\sum_{p=1}^{P-1}\sum_{h=p+1}^{P}\sum_{\substack{i\in C_p \\ l\in C_h \\ i<l}} d_{il}}{I^2 - \sum_{p=1}^{P} n_p^2}$$

where $n_p$ is the size of cluster $C_p$.
The update of the matrix $\mathbf{M}$ is given by solving for each row of $\mathbf{M}$ an assignment problem. At each step of the coordinate descent algorithm $F(\alpha_1, \alpha_2, \mathbf{M})$ does not increase; thus since this function is bounded below the algorithm stops to a stationary point that turns to be at least a local minimum of the problem.
The semi-parametric least-squares estimation of model (1) when the partitioning matrix (3) is considered, corresponds to find $\mathbf{Q}$ that minimizes the following quadratic constrained problem

$$\begin{aligned} & F_2(\mathbf{M}, \mathbf{D}_B, \mathbf{D}_w) = \| \mathbf{D} - \mathbf{M}\mathbf{D}_B\mathbf{M}' - \mathbf{M}\mathbf{D}_W\mathbf{M}' + \mathrm{diag}(\mathbf{M}\mathbf{D}_W\mathbf{M}') \|^2 \\ & \text{subject to} \\ & \max\{{}_W d_{pp} \in \mathbf{D}_W\} \le \min\{{}_B d_{ph} \in \mathbf{D}_B : (h \ne p)\} \\ & m_{ip} \ge 0 \qquad (i=1,\ldots,I;\ p=1,\ldots,P), \\ & \sum_{p=1}^{P} m_{ip} = 1 \qquad (i=1,\ldots,I), \end{aligned} \qquad \text{[P3]}$$

Problem [P2] can be solved by considering a coordinate descent algorithm, which alternates between the update of $\alpha_1$, $\alpha_2$ and $\mathbf{M}$. The update of $\alpha_1$ and $\alpha_2$ are given by

$${}_W d_{pp} = \frac{\sum_{i=1}^{I}\sum_{\substack{l=1 \\ i\ne j}}^{I} d_{il} m_{ip} m_{jp}}{\sum_{i=1}^{I}\sum_{\substack{l=1 \\ i\ne j}}^{I} m_{ip} m_{lp}} \qquad {}_B d_{ph} = \frac{\sum_{i=1}^{I}\sum_{\substack{l=1 \\ i\ne j}}^{I} d_{il} m_{ip} m_{lh}}{\sum_{i=1}^{I}\sum_{\substack{l=1 \\ i\ne j}}^{I} m_{ip} m_{lh}}.$$

The update of the matrix $\mathbf{M}$ is given by solving for each row of $\mathbf{M}$ an assignment problem as in the previous algorithm.

## 5 References


[GRÖ 89] GRÖTSHEL, M. & WAKABAYASHI, Y. (1989), A Cutting Plane Algorithm for a Clustering Problem, *Mathematical Programming*, 45, 59-96.

[HAN 94] HANSEN P., JAUMARD, B. & SANLAVILLE E. (1994), Partitioning Problems in Clustering: A Review of Mathematical Programming Approaches, 228-240.

[RÉG 65] RÉGNIER S. (1965), Sur quelques aspects mathématiques des problèmes de classification automatique, *I.C.C. Bulletin*, Rome.

[VIC 00] VICARI D. & VICHI M. (2000), Non-Hierarchical Classification Structures. In: *Data Analysis,* Studies in Classification Data Analysis and Knowledge Organization, Springer, Eds W. Gaul, O. Opitz, M. Schader, 51-66

# Index