Xèmes rencontres de la Société francophone de classification

- 2003 -

Neufchâtel - Suisse

Sur des Modèles Probabilistes de Classification pour le Data Mining

Helena Bacelar-Nicolau
Laboratório de Estatística e Análise de Dados (LEAD)
Faculdade de Psicologia e Ciências da Educação, Universidade de Lisboa
Alameda da Universidade
1694-013 Lisboa, Portugal
hbacelar@fpce.ul.pt

Fernando Nicolau
Departamento de Matemática
Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa
Campus Quinta da Torre
2929-516 Caparica, Portugal
cladlead@fpce.ul.pt

RÉSUMÉ. Nous référons ici des méthodes et des techniques concernant les modèles probabilistes VAL de classification basées sur l'Affinité des «vecteurs profil» et sur la Validité de l'Affinité. Des propriétés sur la stabilité, la robustesse et la validation des modèles probabilistes VAL et des modèles empiristes qui leur sont associées seront traitées aussi bien que leur extension à l'analyse classificatoire des données complexes ou heterogènes. L' étude a été partiellement développée dans le cadre de nos Programmes de Coopération Franco-Portugaise ou Européenne sur les modèles statistiques pour le data mining ou sur le traitement des «données complexes» .

MOTS-CLÉS : classification, validation, similarité, affinité, simulation, fonction cumulative de répartition.

1. Introduction

La classification ou analyse classificatoire concerne généralement un ensemble de méthodes et techniques multivariées pour la recherche de groupements d'unités de données (individus, échantillons, sous-ensembles d'une population, …) et/ou de groupements de variables, à partir d'un ensemble d'unités de données décrites par des variables. On cherche à obtenir des classes homogènes et bien séparées, selon différents critères. Dans les problèmes pratiques qui nous sont usuellement apportés nous sommes très souvent concernés avec la classification de variables lorsque nous avons à traiter des données issues des sciences humaines – questionnaires de psychologie et des sciences de l'éducation, par exemple – avec la classification des unités de données plutôt en biomédecine et biologie, et avec la classification simultanée des variables et des unités de données dans les domaines de l'économie et des enquêtes de marché.

D'autre part les ensembles d'éléments à classifier, soit dans le cas de la classification d'unités de données, soit lorsque nous cherchons des typologies de variables (ou les deux) proviennent de plus en plus des grandes bases de données comprenant des données hétérogènes et complexes, d'où nous voulons extraire l'information utile et relevante sous la forme de structures de classification.

Le *data mining* concerne à son tour un ensemble diversifié de méthodes exploratoires pour analyser et découvrir des relations entre les objets ou les variables dans les grandes bases de données. Dans ce contexte l'analyse classificatoire peut donc être un outil puissant pour la représentation synthétisée des données et la découverte d'information nouvelle.

Nous avons ainsi eu besoin dans nos projets de recherche statistique appliquée, de développer et utiliser des modèles et techniques de classification exploratoires qui soient simultanément robustes et flexibles. En plus, nous voulons évaluer la qualité et la validité des résultats obtenus. Plus précisément nous cherchons à utiliser des modèles de haute qualité, en ce qui concerne non seulement la validité externe et la validité interne, mais aussi leur «validité intrinsèque». Les *modèles probabilistes* VAL, basées sur des mesures de ressemblance fonctions des *vecteurs profil* répondent, par leurs propriétés, à cette exigence de qualité et flexibilité.

2. Sur les vecteurs profil et les modèles de classification empiriques de base

Dans notre présentation nous nous référerons surtout aux modèles de classification hiérarchique pour des données complexes, où les données observées peuvent être représentées dans une «matrice généralisée», les lignes décrivant les unités de données, les colonnes décrivant les variables, et chaque cellule, croisement d'une ligne avec une colonne, comprenant une ou plusieurs valeurs [BAC 02].

Un modèle de classification sera ici basé sur un coefficient de similarité fonction des vecteurs profil (vecteurs de probabilité décrivant une distribution de probabilité conditionnelle, s'il s'agit d'une population) associés aux paires d'éléments - des unités de données ou des variables – de l'ensemble à classifier. La fonction utilisée n'est autre que le produit interne entre les racines carrées des vecteurs profil, soit le coefficient d'affinité [MAT 51], [BAC 81], [BAC 02].

Dans le cas d'un modèle de classification empirique nous travaillerons donc sur le quadrant positif de la sphère de rayon unité et centrée à l'origine, lorsque nos données sont décrites par des fréquences ou des valeurs réelles positives, ou sur toute la sphère, si nos données sont décrites par des fréquences entières ou des valeurs réelles quelconques, le coefficient d'Ochiai étant le cas particulier pour des données binaires. L'extension à la classification des données complexes ou des mélanges de données suit alors un schéma simple conduisant à la définition d'un coefficient d'affinité pondéré généralisé, dont tous les cas précédents apparaîtrent comme des cas particuliers.

Nous montrons que le choix des profils et du coefficient d'affinité (simple ou généralisé ou pondéré) comme mesure de la similarité, apporte déjà plusieurs propriétés importantes aux modèles de classification empiriques y basées, par rapport à d'autres coefficients (le

coefficient de corrélation de Pearson ou la distance du qui-deux, par exemple). Rappelons aussi que la distance d'Hellinger, associée à l'affinité, avait été utilisée par M. Volle en 1979, dans l'analyse factorielle sphérique, présenté comme alternative à l'analyse factorielle des correspondances et son extension au traitement des données hétérogènes a été programmée et utilisée par L. Bacelar-Nicolau en 2001 [BAC 02], avec de très bons résultats.

2. Sur les modèles de classification probabilistes VAL

Dans le cas des modèles de classification hiérarchique, la première étape d'un modèle étant le choix du coefficient de ressemblance entre pairs d'éléments à classifier, la deuxième consistera à définir un critère d'agrégation entre classes et la troisième à évaluer la validité des résultats obtenus. Cependant, dans beaucoup de situations issues dans la pratique, l'étude des grandes bases de données comprise, on dispose d'information a priori sur la structure sous jacente aux données non négligeable, dont on devrait tenir compte dans la procédure d'extraire de nouvelle information sur leur structure classificatoire. Cette connaissance a priori pourra alors prendre le rôle des hypothèses de référence, dans un contexte probabiliste plus approprié à la nature du problème à résoudre.

Dans un modèle VAL de classification probabiliste nous comptons de même trois étapes: dans la première on calcule les valeurs du coefficient d'affinité normalisé ou standardisé (exact ou asymptotiquement) sous les hypothèses de référence considérées, dans la deuxième on peut déterminer les coefficients de similarité probabiliste associés, à savoir les valeurs correspondantes des fonctions de répartition – coefficients mesurés dans une échelle de probabilité - et dans la troisième on sélectionne des critères d'agrégation (empiriques ou probabilistes), souvent extraits d'une famille paramétrique de méthodes, ce qui permettra des études plus aisées sur la stabilité ou sur la recherche de consensus parmi les membres de la famille. Dans les modèles probabilistes le coefficient probabiliste est en fait une mesure de la Validité de l'Affinité (le Lien de base) et l'étape concernant l'étude de la validation est donc déjà (au moins partiellement) incluse dans la définition du modèle. Il s'agit là d'une «validité intrinsèque» au modèle, qui, remarquons le bien, est toujours dans le cadre des modèles exploratoires de classification (puisque nous ne faisons pas ici d'inférence statistique).

Nous référerons en plus des modèles classificatoires probabilistes basés sur le coefficient d'affinité pondéré généralisé et leurs applications à la classification des données hétérogènes (e.g. [BAC 02], [NIC 99]), d'autres travaux et résultats complémentaires, concernant l'application de l'affinité et ses extensions à l'analyse discriminante [SOU 02], à la représentation géométrique des données [PIN 00], à l'effet des données manquantes sur la classification hiérarchique des variables [SIL 02], aux problèmes de validation et à la vitesse de convergence concernant les résultats asymptotiques sur des données simulés [SOU 02].

3. Bibliographie

[BAC 81] BACELAR-NICOLAU, H. "Contributions to the Study of Comparison Coefficients in Cluster Analysis", 1981, Univ. Lisbon

[BAC 02] BACELAR-NICOLAU, L. & BACELAR-NICOLAU, H. "Hierarchical Classification with a Probabilistic Model using SAS: Applications to Human and Social Sciences". EMPG- 2001, Lisbon (to be published).

[BAC 02] BACELAR-NICOLAU, H. "On the Generalised Affinity Coefficient for Complex Data", *Byocybernetics and Biomedical Engineering*, vol.22, nº 1, 2002, p. 31-42

[MAT 51] MATUSITA, K., "On the Theory of Statistical Decision Functions", *Ann. Instit. Stat. Math*., vol.III, 1951, p.1-30.

[NIC 99] NICOLAU, F.C. & BACELAR-NICOLAU, H. "Clustering Symbolic Objects Associated to Frequency or Probability Laws by the Weighted Affinity Coefficient", *Applied Stochastic Models and Data Analysis. Quantitative Methods in Business and Industry Society*, H. Bacelar-Nicolau, F. C. Nicolau & Jacques Janssen (Eds.), INE, Lisboa, Portugal, 1999, p.155-158

[PINC 00] PINTO DORIA, I., LE CALVÉ, G. & BACELAR-NICOLAU, H. "Comparison of Ultrametrics Obtained with Real Data, Using the PL and the VALAW Coefficients", *Data Analysis, Classification and Related Methods; Series: Studies in Classification, Data Analysis, and Knowledge Organization*, Kiers, Rasson, Groenen, Schader (Eds.), Springer, 2000, p. 160-165

[SIL 02] SILVA, A.L, BACELAR-NICOLAU, H. & SAPORTA, G. "Missing Data in Hierarchical Classification of Variables - a Simulation Study" in *Classification Clustering and Data Analysis*, 2002, p.121-128, Springer.

[SOU 02] SOUSA FERREIRA, A., CELEUX, G. & BACELAR-NICOLAU, H. "New developments on combining models in discrete discriminant analysis by a hierarchical coupling approach". *Applied Stochastic Models and Data Analysis*; G. Govaert, J. Janssen, N. Limnios (Eds.), UTC, 2001, p. 430-435

[SOU 02] SOUSA, A., SILVA, O., BACELAR-NICOLAU, H. & NICOLAU, F. "Validação em Classificação Hierárquica". *JOCLAD-2002* (to be published).

* Ce travail a été partiellement supporté par le Programme de Coopération Scientifique et Technique Franco-Portugaise MSPLDM-542-B2 (Ambassade de France au Portugal et Ministère de la Science et de l'Enseignement Supérieur - ICCTI) co-dirigé par H. Bacelar-Nicolau (LEAD-FPCEUL) et G. Saporta (Chaire de Stat. Appliquée-CNAM), par le Project d'Analyse des Données Multivariées (CEAUL-FCT) dirigé par H. Bacelar Nicolau, et par le CMA-UNL.

Réalisation de dissimilarités

François Brucker

ENST Bretagne
Département d'intelligence artificielle et sciences cognitives
BP 832
29285 Brest CEDEX (Breizh)
francois.brucker@enst-bretagne.fr

RÉSUMÉ. *Nous étudions dans cette communication une nouvelle approche pour l'étude des classes induites par une dissimilarité. En considérant un sous-ensemble particulier de l'ensemble des classes d'une dissimilarité, nous montrons qu'il est possible de décrire exhaustivement la dissimilarité d'origine. De plus, les classes considérées sont assez peu nombreuses, représentables graphiquement et polynomialement calculables.*

MOTS-CLÉS : *dissimilarité, classes d'une dissimilarité*

1. Introduction

Il y a deux principales approches en classification, l'une c'est de voir les classes comme des intersections de cliques maximales des graphes seuils d'une dissimilarité (dans la lignée de Jardine et Sibson 1971, Janowitz 1978 et Bertrand 2000), l'autre est de voir les classes comme des parties connexes d'un graphe (par exemple les classes d'une pyramide sont des parties connexes d'un chemin). Nous montrerons dans cette communication que l'on peut, grâce aux réalisations de dissimilarités, combiner ces deux approches.

Nous travaillerons sur un ensemble X dont les éléments sont appelés *objets* et qui seront au nombre de n. Une *dissimilarité* sur X est une fonction d de $X \times X$ dans l'ensemble des réels positifs telle que $d(x, y) = d(y, x)$ pour $x, y \in X$ et $d(x, x) = 0$ pour $x \in X$. La dissimilarité d est dite *propre* quand $d(x, y) = 0 \Rightarrow x = y$.

On supposera que X est décrit par une dissimilarité propre d. Les dissimilarités étudiées dans ce travail seront toutes supposées propres. Dissimilarité et dissimilarité propre devront donc être considérées comme synonyme.

On appellera graphe seuil $G_h = (X, E_h)$ d'une dissimilarité d pour le seuil h, le graphe ayant X pour ensemble de sommets et admettant la paire xy comme arête si et seulement si $d(x, y) \leq h$.

On associe ainsi à toute dissimilarité d l'ensemble de ses *classes*, ces dernières étant toutes les intersections non vides des cliques maximales des graphes seuils G_h de d pour $h \in \mathbb{R}^+$. On peut évidemment restreindre les valeurs prises par h aux valeurs prises par d. Le diamètre d'une classe C de d est alors : $\text{diam}(C) = \max\{d(x, y) | x, y \in C\}$ et on a $d(x, y) = \min\{\text{diam}(C) | x, y \in C, c \in \mathcal{C}\}$ (*cf.* Batbedat 1988, Bertrand 2000 pour une bijection générale entre les dissimilarités et leurs classes).

Nous montrons dans cette communication qu'il n'est pas nécessaire d'étudier toutes les classes d'une dissimilarité, mais que l'on peut ne considérer qu'un nombre restreint d'entres elles. Cet ensemble de classes, appelé *réalisation* de la dissimilarité possède, entre autres, la propriété d'être calculable en temps polynomial et "représentable" par un graphe que nous exhiberons. Nous finirons par un exemple montrant une application pratique de ces résultats.

2. Réalisation d'une dissimilarité

Soit d une dissimilarité et $\mathcal{C}$ l'ensemble de ses classes. On note $\delta[d](x,y)$ ($x,y \in X$) le sous ensemble de X tel que :

$$\delta[d](x,y) = \cap\{C | x,y \in C, C \in \mathcal{C}\}$$

On peut montrer facilement que la fonction $\delta[d]$ de $X \times X$ dans 2^X vérifie les propriétés suivantes (Barthélemy, 2003) :

[BD$_1$] pour tout $x \in X$, $\delta[d](x,x) = \{x\}$,
[BD$_2$] pour tous $x,y \in X$, $\delta[d](x,y) = \delta(y,x)$,
[BD$_3$] pour tous $x,y \in X$, $\{x,y\} \subseteq \delta[d](x,y)$.
[BD$_4$] il existe $u,v \in X$ tels que $X = \delta[d](u,v)$.
[BD$_5$] pour tous $x,y \in X$ et tous $z,t \in \delta[d](x,y)$, $\delta[d](z,t) \subseteq \delta[d](x,y)$.

De plus, pour tous $x,y \in X$, on a $\delta[d](x,y) \in \mathcal{C}$ et diam$(\delta[d](x,y)) = d(x,y)$.

On appellera *réalisation de* d l'ensemble $\Delta \subseteq \mathcal{C}$ de tous les $\delta[d](x,y)$, $x,y \in X$. Cet ensemble suffit, si l'on associe à chacun de ses élément son diamètre, pour retrouver la dissimilarité d. Contrairement aux nombre exponentiel de classes que peut posséder d, Δ possède au plus $\frac{n(n-1)}{2}$ éléments.

Si l'on définit pour une dissimilarité d une boule de centre $x \in X$ et de rayon α comme étant l'ensemble $B(x,\alpha) = \{y | d(x,y) \leq \alpha\}$, la proposition suivante montre, de plus, que Δ peut être construit en temps polynomial, rendant possible l'utilisation "pratique" de ce modèle.

Proposition 1 *Soit d une dissimilarité. On a :*

$$\delta[d](x,y) = \cap_{z \in X}\{B(z, \max\{d(x,z), d(y,z), d(x,y)\})\}$$

L'exemple ci-dessous montre la réalisation de la dissimilarité d_1 de la table 1.
– $\delta[d_1](x,y) = \{x,y\}$
– $\delta[d_1](y,z) = \{y,z\}$
– $\delta[d_1](x,z) = \{x,z\}$
– $\delta[d_1](t,x) = \delta[d_1](t,z) = \{x,z,t\}$
– $\delta[d_1](t,y) = \{x,y,z,t\}$

TAB. 1. *La dissimilarité d_1.*

d :	x	y	z	t
x	0			
y	1	0		
z	2	1	0	
t	2	3	2	0

L'intérêt de réaliser une dissimilarité est triple. Tout d'abord, il n'est ni nécessaire de fixer le nombre de classes que l'on veut obtenir (comme pour les "k-means" par exemple) ni nécessaire d'approximer la dissimilarité par une autre dissimilarité satisfaisant un modèle particulier (comme le modèle hiérarchique pour une classification ascendante hiérarchique, par exemple).

De plus, les éléments de la réalisation d'une dissimilarité sont des classes au sens classique du terme (*ie.* des intersections de cliques maximales des graphes seuils assiociés à la dissimilarité).

Enfin, les classes sont formées à partir de couples d'éléments. On peut donc plus facilement choisir de conserver telle ou telle classe, plus pertinante qu'une autre (il n'est pas nécessaire d'étudier la classe de deux éléments

de X que l'on sait dissemblable), ou comprendre l'origine de ladite classe (c'est la plus petite classe de la dissimilarité contenant les deux éléments générateurs. Ou encore, c'est l'intersection de toutes les cliques maximales des graphes seuils assiociés à la dissimilarité contenant ces deux éléments).

3. Représentation d'une réalisation

Soit d une dissimilarité et Δ sa réalisation. L'algorithme ci-dessous construit un graphe $G_\Delta = (X, E_\Delta)$.

Initialisation :
- Soit $\Delta_0 \leftarrow \Delta$,
- Soit $G_0 = (X, E_0)$ un graphe tel que $E_0 = \{\phi\}$,
- Soit $A_0 \leftarrow \{\{x, y\} | x \neq y \in X\}$.

On suppose que les i premières étapes ont été effectuées. L'étape $i + 1$ consiste alors à :
- choisir $\{x_i, y_i\} \in A_i$ tel que $\delta[d](x_i, y_i)$ soit un élément minimal de Δ_i pour l'inclusion.
- Si $\delta[d](x_i, y_i)$ est une partie connexe de G_i, alors $E_{i+1} \leftarrow E_i$. $E_{i+1} \leftarrow E_i \cup x_i y_i$, sinon.
- On pose $A_{i+1} \leftarrow A_i \backslash \{x_i, y_i\}$.
- On pose $\Delta_{i+1} \leftarrow \{\delta[d](x, y) | \{x, y\} \in A_{i+1}$.

L'algorithme se termine lorsque A_i est vide (c'est à dire en $\frac{n(n-1)}{2}$ itérations). Chaque itération pouvant être effectuée en $\mathcal{O}(n^3)$ opérations, la complexité générale de l'algorithme est en $\mathcal{O}(n^5)$.

La proposition suivante montre le lien entre G_Δ, les classes et la réalisation de d.

Proposition 2 *Soit d une dissimilarité. Le graphe G_Δ qui lui est associé vérifie que :*
- *toute les classes de d sont des parties connexe de G_Δ,*
- *tous les $\delta[d](x, y)$ $(x, y \in X)$ sont des parties connexe de G_Δ.*

De plus, de tous les graphes qui vérifient l'une ou l'autre de ses propriétés (le graphe complet en est un), G_Δ est minimum en nombres d'arêtes.

De façon générale, trouver, pour un sous ensemble C de 2^X, un graphe G minimum en nombre d'arêtes tel que chaque élément de C soit une partie connexe de G est NP-difficile (Osswald, 2003). On a donc obtenu une instance polynomiale de ce problème pour le cas particulier des classes d'une dissimilarité (on peut monter que le résultat s'étend à tout sous ensemble fermé de 2^X).

De plus, l'algorithme ci-dessus permet de se doter d'un moyen simple de représentation de la binarisation d'une dissimilarité d : le graphe G_Δ. Il donne un moyen de connaître les liaisons existantes entre données sans regarder exhaustivement tous les éléments de Δ. Un graphe G_Δ associé à la table 1 est présenté en figure 1.

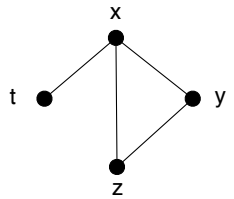


FIG. 1. *G_Δ associé à la table 1*

4. Exemple d'application

Nous reprenons ici un des exemples développés dans Barthélemy et Luong 1998. La matrice de dissimilarité utilisée est celle des proximité lexicales entre 23 œuvres de Jean Giraudoux. L'*indice de connexion lexicale* (Muller,

1977 ; Brunet, 1988), adaptation de l'indice de Jaccart (1908), permet de comparer le lexique entre deux textes en faisant intervenir la *fréquence* d'apparition des mots.

Les 23 textes de Jean Giraudoux dont la connexion lexicale sont, dans l'ordre chronologique de leur publication : *Provinciales* (Pro), *L'école des indifférents* (Ind), *Simon le Pathétique* (Sim), *Suzanne et le Pacifique* (Suz), *Siegfried et le Limousin* (S&L), *Juliette au pays des hommes* (Jul), *Bella* (Bel), *Églantine* (Egl), *Siegfried* (Sie, *Amphitrion 38* (Amp), *Aventures de Jérôme Bardini* (Bar), *Judith* (Jud), *Intermezzo* (Int), *Combat avec l'Ange* (Com), *La guerre de Troie n'aura pas lieu* (Gue), *Electre* (Ele), *Cantique des Cantiques* (Can), *Choix des Élues* (Elu), *Ondine* (Ond), *Apollon de Bellac* (Apo), *Sodome et Gomorrhe* (Sod), *La Folle de Chaillot* (Fol), *Pour Lucrèce* (Luc).

Par manque de place, nous ne montrerons ici que le graphe G_Δ (figure 2). On peut tout d'abord remarquer la grande proximité des différents types d'œuvres entre elles. En effet, les pièces antiques par exemple sont toutes des feuilles du *cantique des cantiques* et les différentes romans sont groupés par périodes (à part peut être la deuxième période, à cheval entre la première et la troisème période). On peut également remarquer la place centrale qu'occupe le *cantique des cantiques*, joignant les pièces antiques aux pièces modernes, et lien d'icelles vers les romans.

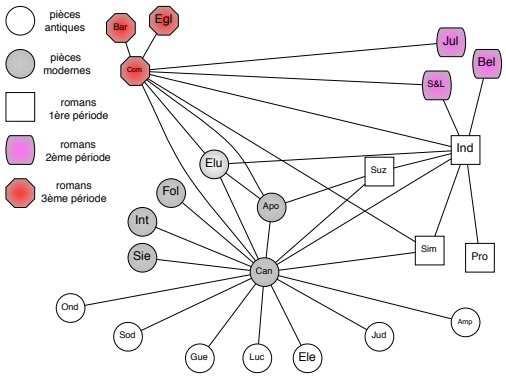


FIG. 2. *Binarisation des œuvres de Jean Giraudoux*

5. Bibliographie

BARTHÉLEMY, J.-P. (2003), communication personnelle.

BARTHÉLEMY, J.-P., LUONG, X. (1998), "Représenter les données textuelles par des arbres…," dans les actes de la quatrième journées internationales d'analyse de données textuelles (JADT'98), Université de Nice, 49–71.

BATBEDAT, A. (1988), "Les isomorphismes HTS et HTE (après la bijection de Benzécri-Johnson)," *Metron*, *46*, 47–59.

BERGE, C. (1987), *Hypergraphes*, Paris : Gauthier-Villars.

BERTRAND, P. (2000), *Set Systems and Dissimilarities European Journal of Combinatorics*, *21*, 727 – 743.

BRUNET, E. (1978), *Le vocabulaire de Jean Giraudoux structure et évolution*, Genève : Slatkine.

BRUNET, E. (1988), "Une mesure de la distance intertextuelle : la connexion lexicale," dans *Le nombre et le texte, Revue Informatique et Statistique dans les Sciences Humaines*, Université de Liège, 81–116.

JACCART, P. (1908), "Nouvelles recherches sur la distribution florale," *Bulletin de la Société Vanderbilt de Sciences Naturelles*, *44*.

JANOWITZ, M. F. (1978), "An order theoretic model for cluster analysis," *SIAM Journal on Applied Mathematics*, *34*, 55–72.

JARDINE, N., et SIBSON, R. (1971), *Mathematical Taxonomy*, London : Wiley.

MULLER, C. (1977), *Principes et méthodes de statistiques lexicales*, Paris : Hachette.

OSSWALD, C. (2003), communication personnelle.

Validation du consensus en classification hiérarchique

Guy Cucumel[1], François-Joseph Lapointe[2]

[1]*École des sciences de la gestion*
Université du Québec à Montréal
CP 8888, Succursale Centre-ville
Montréal (Québec)
H2T 1M5
Canada
cucumel.guy@uqam.ca

[2]*Département de sciences biologiques*
Université de Montréal
CP 6128, Succursale Centre-ville
Montréal (Québec)
H3C 3J7
Canada
françois-joseph.lapointe@umontreal.ca

RÉSUMÉ. *Les méthodes de recherche de consensus sont largement employées pour combiner des hiérarchies obtenues à partir de plusieurs jeux de données portant sur les mêmes observations. De nombreux algorithmes ont été proposés au cours des dernières décennies pour construire des hiérarchies consensus, mais la validation du résultat est rarement étudiée. Nous passons en revue quelques approches pour en établir la fiabilité et la stabilité. Lorsque celles-ci ne sont pas vérifiées, nous proposons une alternative : le consensus multiple.*

MOTS-CLÉS : *classification hiérarchique, consensus, validation, consensus multiple*

1. Introduction

Une méthode de recherche de consensus peut être décrite comme une fonction qui associe à un profil de k hiérarchies $P=\{H_1, H_2, \ldots, H_k\}$ une hiérarchie H_c représentative (en un certain sens) des hiérarchies initiales ([LEC 87] et [LEC 98]). Depuis le premier algorithme proposé par Adams [ADA 72], l'utilisation de hiérarchies consensus a largement augmenté et à donné naissance à de nombreux algorithmes, certains s'appliquant à des profils de hiérarchies indicées ([NEU 83], [STI 84], [FIN 85], [BAR 86] et [LAP 97]). La recherche d'un consensus

produit toujours une hiérarchie, indépendamment de l'accord ou du désaccord entre les hiérarchies initiales, ce qui soulève plusieurs questions. Quelle est la pertinence de cette solution ? Quel est l'effet d'une hiérarchie particulière du profil P sur le consensus obtenu ? Certains paliers du consensus sont-ils plus stables que d'autres ? Dans cette communication, nous abordons ces questions et nous passons en revue quelques approches permettant d'évaluer la fiabilité et la stabilité de hiérarchies consensus. Nous proposons également une solution alternative au consensus lorsqu'il y a un manque de fiabilité ou de stabilité : le consensus multiple.

2. Évaluation statistique de la congruence de hiérarchies

Étant donné un indice de congruence entre hiérarchies ([ROB 79], [DAY 83], [DAY 86] et [EST 85]), deux hiérarchies seront dites congruentes si elles sont plus semblables entre elles que ne le sont la plupart des paires de hiérarchies aléatoires construites sur les mêmes objets. Cette évaluation se fait à l'aide d'un test par permutations en trois étapes : 1) la ressemblance entre les deux hiérarchies est mesurée par l'indice de congruence, 2) une distribution statistique de l'indice de congruence est produite en mesurant la congruence entre des hiérarchies aléatoires et 3) une valeur critique est déterminée à partir de la distribution obtenue en tenant compte du risque d'erreur. Si la congruence entre les deux hiérarchies initiales est plus forte que la valeur critique, on rejette l'hypothèse nulle et on conclut à la congruence des deux hiérarchies.

3. Évaluation de la pertinence d'un consensus

Une hiérarchie consensus H_c peut être obtenue à partir de n'importe quel profil de hiérarchies P, mais elle ne sera utile que si elle résume bien P. Nous définissons un consensus comme étant pertinent s'il est plus proche de P que ne le sont la plupart des consensus, construits avec la même méthode que H_c, du profil de hiérarchies aléatoires dont chacun d'eux dérive. Le test se fait selon trois étapes analogues à celle du paragraphe précédent : 1) la distance entre H_c et P est mesurée par la somme des indices de congruence entre H_c et chacune des hiérarchies de P, 2) une distribution statistique de distances est produite en mesurant la distance entre les hiérarchies consensus dérivées de profils de hiérarchies aléatoires et ces profils et 3) une valeur critique est déterminée à partir de la distribution obtenue en tenant compte du risque d'erreur. Si la distance entre H_c et P est plus faible que la valeur critique, on rejette l'hypothèse nulle et on conclut à la pertinence du consensus.

4. Évaluation de l'effet des hiérarchies initiales sur un consensus

Une façon d'évaluer l'effet des hiérarchies initiales est de construire une série de consensus à partir de sous-profils de hiérarchies en retirant tour à tour une hiérarchie du profil. Cette procédure de type jacknife peut être utilisée pour générer k consensus, un pour

chaque sous-profil, comprenant k-1 hiérarchies. La méthode peut être généralisée en retirant un nombre plus important de hiérarchies du profil P lors de la création des sous-profils. La congruence entre les consensus obtenus peut être extrêmement informative sur l'effet de hiérarchies particulières lors de l'obtention du consensus.

5. Évaluation de la stabilité des paliers d'un consensus

Dans le prolongement de l'évaluation de l'effet des hiérarchies initiales sur un consensus, il est intéressant de vérifier la stabilité des paliers du consensus. Un palier sera d'autant plus stable qu'il est présent dans des consensus issus de sous-profils de P. La procédure de type jacknife décrite au paragraphe précédent peut être appliquée afin de déterminer pour chacun des paliers de la hiérarchie consensus issue de P le nombre de hiérarchies consensus issues des sous-profils qui le contiennent. Alternativement, on peut appliquer une procédure de type bootstrap en rééchantillonnant avec remplacement les hiérarchies du profil P de manière à obtenir des profils P' de même cardinalité que P. L'évaluation de la stabilité se fait alors comme décrit précédemment en substituant les profil P' aux sous-profils.

6. Consensus multiple

Lorsque les hiérarchies du profil initial sont très différentes, un consensus unique n'est pas très informatif. Lorsqu'une ou plusieurs méthodes de validation mettent en évidence un manque de fiabilité ou de stabilité, un consensus unique n'est pas nécessairement adapté à l'analyse des données. Il est par contre possible que plusieurs consensus permettent une meilleure représentation de la structure des données. Nous abordons dans ce paragraphe la question de la recherche d'un partitionnement du profil P en sous-profils conduisant chacun à une hiérarchie consensus.

La recherche d'un consensus multiple consiste à partitionner le profil P en m sous-profils (P_1, P_2, …, P_m) (m<k) dans lesquels les hiérarchies sont « relativement » semblables et à partir desquelles on construit m consensus (Hc_1, Hc_2, …, Hc_m) selon le même algorithme.

La manière la plus simple de partitionner ce profil est de calculer les k(k-1)/2 distances entre toutes les paires de hiérarchies. Un algorithme de classification hiérarchique est ensuite appliqué à la matrice de distances entre hiérarchies. La coupure de cette hiérarchie permet de partitionner P.

7. Bibliographie

[ADA 72] ADAMS E. N. III., " Consensus Techniques and the Comparison of Taxonomic Trees ", *Systematic Zoology*, vol. 21, 1972, p. 390-397.

[BAR 86] BARTHÉLEMY J. P., MCMORRIS F. R., " The Median Procedure for n-Trees ", *Journal of Classification*, vol. 3, 1985, p. 229-334.

[DAY 83] DAY W. H. E., " Properties of the Nearest Neighbor Interchange Metric for Trees ", *Journal of Theoretical Biology*, vol. 101, 1983, p. 275-288.

[DAY 86] DAY W. H. E., " Analysis of Quartet Dissimilarity Measures between Undirected Phylogenetic Trees ", *Systematic Zoology*, vol. 35, 1986, p. 325-333.

[EST 85] ESTABROOK G. F., MCMORRIS F. R., MEACHAM C. A., " Comparison of Undirected Phylogenetic Trees Based on Subtrees of Four Evolutionary Units", *Systematic Zoology*, vol. 34, 1985, p. 193-200.

[FIN 85] FINDEN C. R., GORDON A. D., " Obtaining Common Pruned Trees ", *Journal of Classification*, vol. 2, 1985, p. 225-276.

[HAR 67] HARTIGAN J. A., " Representation of Similarity Matrices by Trees ", *Journal of the American Statistical Association*, vol. 62, 1985, p. 1140-1158.

[LAP 97] LAPOINTE F.-J., CUCUMEL G., " The Average Consensus Procedure : combination of Weighted Trees Containing Identical or Overlapping Sets of Objects ", *Systematic Zoology*, vol. 46, 1984, p. 306-312.

[LEC 87] LECLERC B., CUCUMEL G., " Consensus en classification : une revue bibliographique ", *Mathématiques et sciences humaines*, n° 100, 1987, p. 109-128.

[LEC 98] LECLERC B., " Consensus of Classification : the Case of Trees ", *Data Analysis, Classification and Related Methods*, H. A. L. Kiers et al. (eds.), 1998, p. 81-90, Springer-Verlag, Berlin.

[NEU 83] NEUMANN D. A., " Faithful Consensus Methods for n-Trees ", *Mathematical Biosciences*, vol. 63, 1983, p. 271-287.

[ROB 79] ROBINSON D. F., FOULDS L. R., " Comparison of Weighted Labelled Trees ", *Lectures Notes in Mathematics*, vol. 748, 1979, p. 119-126, Springer-Verlag, Berlin.

[STI 84] STINEBRICKNER R., " An Extension of Intersection Methods from Trees to Dendrograms ", *Systematic Zoology*, vol. 33, 1984, p. 381-386.

[SPR 99] SPRINGER, M. S., AMRINE H. M., BURK A., STANHOPE M. J., " Additional Support for Afrotheria and Paenungalata, the Performance of Mitochondrial versus Nuclear Genes, and the Impact of Data Partitions with Heterogeneous Base Composition ", *Systematic Zoology*, vol. 48, 1999, p. 65-75.

Régression Inverse : de la réduction de dimension à la discrimination fonctionnelle

Louis Ferré

Equipe GRIMM
Université Toulouse Le Mirail
UFR SES, Département de Math-Info
31058 Toulouse Cedex
loferre@univ-tlse2.fr

RÉSUMÉ. *Les méthodes de régression inverse ont été intialement introduites pour contourner le fléau de la dimension dans les problèmes de régression non-paramétrique multivarié. Nous présentons ici une extension de la méthode de régression inverse par tranche (SIR) à des données de type fonctionnel. Nous proposons plusieurs estimateurs de l'espace exhaustif ainsi que plusieurs approches pour résoudre des problémes de prédiction dans le cas d'une variable réponse réelle. Les méthodes de régression inverse apparaissent alors autant comme des techniques de filtrage que comme des techniques de réduction de dimension. Nous étendons notre propos au cas où la variable réponse est qualitative. Dans ce cadre, notre approche fournit une réponse originale et pertinente pour traiter de l'affectation d'individus à des groupes, i.e., pour des problèmes de classifications supervisées. Nous présentons plusieurs exemples d'application.*

MOTS-CLÉS : *Régression Inverse, Perceptron multi-couches, Courbes, Discrimination, Reconnaissance de formes, Données Fonctionnelles*

1. Introduction

Pour contourner le problème du fléau de la dimension en régression non-paramétrique multivariée, de nombreuses méthodes ont été proposées. Parmi elles, la régression inverse par tranche, introduite par Li (1991), s'appuie sur un modèle particulièrement intéressant d'un point de vue statistique et présente de surcroît l'avantage de la simplicité. En effet, le modèle utilisé stipule l'existence d'un sous-espace "exhaustif" dans le sens où la régression de la variable réponse, Y, sur la variable explicative X, ne dépend que de la projection de X sur ce sous-espace. De nombreux travaux ont été menés d'une part pour éviter certains écueils de la régression par tranche conduisant entre autre aux méthodes SIR II (Li, 1991), pHd (Li, 1992), Save (Cook, 1991) et CME (Cook et Li, 2002), MAVE (Xia et al., 2002), CKMS (Yin et Cook, 2002) et d'autre part pour estimer la dimension de l'espace exhaustif voir, par exemple, Schott (1994), Li(1991), Ferré (1998), Villela (1998). Nous présenterons, dans un premier temps, une extension de la méthode de régression inverse par tranche au cas où la variable X est à valeurs dans un espace fonctionnel $\mathcal{F}$. Dans cette situation de plus en plus fréquente, les observations sont des courbes et l'importance de prendre en compte la spécificité de telles données est maintenant bien établie comme l'indiquent, par exemple, Ramsay et Silverman (1997).

Nous intéresserons ici à des problèmes de prédiction à partir d'une variable fonctionnelle. Ainsi, notre approche s'apparentera-t-elle à une méthode de filtrage dans la mesure où les données fonctionnelles sont projetées sur une base obtenue par régression inverse et les coordonnées sur cette base utilisées pour obtenir la prédiction. Usuellement, sont souvent utilisées soit des bases fixées a priori à partir de la nature des données (ondelettes,

0. Ce travail a été réalisé grâce aux contributions de A.F. Yao du Centre Océanique de Marseille et N. Villa de l'équipe GRIMM de Toulouse

séries de Fourier,...) mais indépendantes de celles-ci, soit des bases directement liées à la variable explicative et obtenues par Analyse en Composantes Principales. L'intérêt de la régression inverse est que, contrairement à l'ACP qui ne prend en compte qu'une information marginale, la régression inverse va permettre de construire des bases qui reflètent une information conditionnelle, information idoine dans les problèmes de régression. Néanmoins, le caractère infini dimensionel des données nécessite une étape préalable pour l'estimation de l'espace exhaustif, étape réalisée par filtrage (utilisation des projections de la variable explicative sur les premiers vecteurs de son ACP) ou par régularisation (pénalisation d'un opérateur de covariance).

Dans le cas de la prédiction d'une variable réelle, nous montrons comment notre approche peut-être utilisée. Après estimation de l'espace exhaustif, il est nécessaire d'estimer la fonction de lien entre Y et la projection de X sur l'espace exhaustif. On s'appuiera donc sur une méthode de régression non-paramétrique après projection des données. Deux approches seront privilégiées ici : l'une par méthode du noyau qui peut être remplacée par une méthode de lissage quelconque (splines, ondelettes, polynômes locaux,...), l'autre en utilisant un réseau de neurones supervisé de type perceptron multi-couches. Le couplage régression inverse et réseau de neurones apparaît alors comme un moyen de réaliser des réseaux de neurones dont les entrées sont fonctionnelles, fournissant une alternative à l'approche de Conan et Rossi (2002).

Nous nous intéressons ensuite à une variable réponse qualitative. Nous abordons ainsi des problèmes de discrimination et nous montrons que le modèle de régression inverse conduit naturellement à une modélisation des probabilités d'appartenance aux groupes et que l' estimation de l'espace exhaustif est identique à celle des espace discriminants de l'Analyse linéaire discriminante de Fisher dans le cas multivarié et à leurs variantes dans le cas fonctionnel. Alors que des règles d'affectation géométriques sont le plus souvent retenues, notre approche légitime l'utilisation des estimations des probabilités d'appartenance aux groupes comme règle de décision.

2. Modèle

Soit $(\Omega, \mathcal{B}, P)$ un espace probabilisé. On note Y la variable réponse qui est à valeurs dans $(\mathcal{D}_{\mathcal{Y}}, \mathcal{B}_{\mathcal{Y}})$ et X la variable explicative qui est à valeurs dans $(\mathcal{F}, \mathcal{B}_{\mathcal{F}})$. On note $Var(X) = \Gamma_X$ et $Var(E(X|Y)) = \Gamma_{E(X|Y)}$ et on suppose que le lien entre Y et X s'écrit selon le modèle :

$$Y = f(\pi_E X, \varepsilon) \qquad [1]$$

où E est un espace de dimension K, f est une fonction de $\mathbb{R}^{K+1}$ dans $\mathbb{R}$ et ε est une variable aléatoire réelle indépendante de X. Ce modèle traduit bien que l'information apportée par X sur Y est entièrement portée par $\pi_E X$, la projection de X sur E qui est l'espace "exhaustif." Sous la condition :

Condition 1 *Pour tout b de $\mathcal{F}$, il existe un vecteur C de $\mathbb{R}^K$ tel que $E(< b, X > |B) = C'B$ avec $B' = (< \theta_1, X >, \ldots, < \theta_K, X) >)$, oú $(\theta_k)_{k=1,\ldots,K}$ est une base de E,*

Dauxois et al. (2001) montrent que le sous-espace E contient le sous-espace propre associé aux K valeurs propres non-nulles de l'opérateur $\Gamma_X^{-1}\Gamma_{E(X|Y)}$ qui, sous des conditions convenables, existe et est compact.

3. Estimation

3.1. *Estimation de l'espace exhaustif*

La méthode de régression inverse par tranches consiste à estimer $\Gamma_{E(X|Y)}$ par $\hat{\Gamma}_{E(X|Y)}$ la matrice de covariance empirique des moyennes de X dans des "classes" obtenues par tranchage de $\mathcal{D}_{\mathcal{Y}}$. La matrice de covariance empirique de X usuelle, $\hat{\Gamma}$ est, elle, utilisée pour estimer Γ. Il est alors tentant d'estimer E par les vecteurs propres associés aux K plus grandes valeurs propres de $\hat{\Gamma}_X^{-1}\hat{\Gamma}_{E(X|Y)}$. Cependant, $\hat{\Gamma}_X$ ne possédant pas d' inverse borné, il n'est pas possible de procéder ainsi.

Pour contourner le problème, on peut utiliser :

– une approche de type filtrage en projetant X sur une base par exemple, celle correspondant à l'ACP (ou à la décomposition de Karunen-Loeve pour parler en termes de processus) de X (Ferré et Yao, 2003) ;
– une approche de type pénalisation (Ferré et Villa, 2003).

Pour cette dernière, nous utilisons le fait que les vecteurs propres de $\Gamma_X^{-1}\Gamma_{E(X|Y)}$ sont aussi les solutions emboîtées et sous contrainte de Γ^{-1}-orthogonalité du problème de minimisation en $\beta \in \mathcal{F}$ de :

$$\frac{<\Gamma_{E(X|Y)}\beta, \beta>}{<\Gamma_X\beta, \beta>}. \qquad [2]$$

Ainsi, la régularisation s'obtient en considérant le critère pénalisé :

$$\frac{<\Gamma_{E(X|Y)}\beta, \beta>}{<\Gamma_X\beta, \beta> + \delta < D^2\beta, D^2\beta>},$$

où D est l'opérateur de différentiation et δ est le paramètre de régularisation. Cette démarche est à rapprocher de celle de Leurgan et al. (1992) pour l'Analyse Canonique de courbes.

3.2. *Estimation de la fonction* f

L'estimation de la fonction f peut dépendre de la valeur de la dimension K de l'espace exhaustif. Ainsi, si K est raisonnablement petit, il sera possible d'utiliser une méthode non-paramétrique classique. Cependant, si la réduction de dimension n'est pas suffisante (n'oublions pas que nous sommes a priori en dimension infinie), ces méthodes peuvent s'avérer désastreuses. Il est alors pertinent de faire appel à des techniques qui ne sont pas sensibles au fléau de la dimension et dont la dimension de Vapnick-Chesnovski (Vapnick, 2000) est faible. C'est pourquoi, nous utilisons ici un réseau de neurones de type perceptron multi-couches dont les propriétés d'approximation universelle sont bien connues.

Notre démarche peut alors se présenter également comme un moyen de réaliser un réseau de neurones dont les entrées sont fonctionnelles et les cibles réelles :

– tout d'abord, on projette les entrées dans l'espace E,
– puis, on construit un perceptron (réel) dont les entrées sont les coordonnées des projections.

Nous démontrons dans Ferré et Villa (2003) la convergence des poids du réseau vers les poids optimaux ce qui justifie la méthode globale pour mettre en oeuvre des réseaux à entrées fonctionnelles. Notre approche se résume par le graphique de la Figure 1.

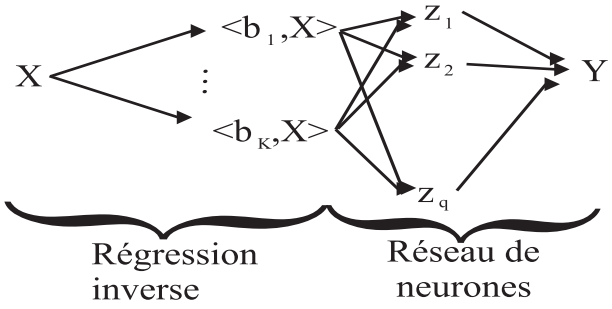


FIG. 1. *Diagramme récapitulatif*

Nous illustrons notre propos par un exemple sur des données de spectrométrie qui indique que notre démarche s'avère supérieure aux méthodes "classiques" et compétitive par rapport à des modèles sophistiqués et au faible pouvoir de généralisation comme celles présentées dans Borggaard et Thodberg (1992).

3.3. *Cas de la classification supervisée*

On considère maintenant que $\mathcal{D}_{\mathcal{Y}} = \{0,1\}^g$ et que $Y = (\mathbb{1}_{G_1}, ..., \mathbb{1}_{G_g})$ où $(G_j)_{j=1,...,g}$ définit une partition de Ω. Soit Z la variable aléatoire correspondant à un codage linéaire et définie par $Z = \sum_{j=1}^{g} j\mathbb{1}_{G_j}$. On a alors :

$$E(Y|X) = (P(Z = 1|X), ..., P(Z = g|X))$$

et on peut écrire le modèle (1) pour une version multivariée de Y en considérant f comme une fonction de $\mathbb{R}^K$ dans $\mathbb{R}^g$ et telle que :

$$E(Y|X) = f(\pi X) = (f_1(\pi X), ..., f_g(\pi X)).$$

Ce modèle vaut également pour le cas où $\mathcal{F} = \mathbb{R}^p$ et l'espace E est alors un CMS (Central Mean Subspace) au sens de Cook et Li (2002). En utilisant l'expression (2), on voit que cet espace n'est autre que l'espace discriminant de l'analyse discriminante linéaire de Fisher. Le problème de l'affectation au groupe s'identifie totalement dans notre modèle à celui de la prédiction de Y. En suivant la démarche ci-dessus, elle conduit à l'estimation de f, soit par une méthode à noyau, soit par un réseau de neurones et à affecter l'individu au groupe correspondant au f_g maximal. Ainsi, notre démarche revient à une estimation non-paramétrique des probabilités d'appartenance aux groupes après projection dans l'espace exhaustif et à la maximisation de la règle de Bayes.

Nous présentons plusieurs exemples pour expliquer la mise en oeuvre pratique de la méthode. La comparaison sur ces exemples avec d'autres méthodes démontre, là encore, la pertinence de notre approche surtout si on tient compte du rapport simplicité-performance.

4. Bibliographie

Borggaard, C. and Thodberg, H.H. (1992) Optimal minimal neural interpretation of spectra. *Analytic Chemistry*, **64**, 545-551.

Conan-Guez, B. and Rossi, F. (2002) Approche régularisée du traitement de données fonctionnelles par un perceptron multi-couches. *Actes des neuvièmes journées de la SFC, Toulouse*, 169-172.

Cook, R.D. (1991) Discussion of Li (1991) *J. Am. Statis. Ass.*, **86**,328-332.

Cook, R.D. and Li, B. (2002) Dimension reduction for the conditional mean in regression. *Ann. Statist.*, **30**, 455-474.

Dauxois, J., Ferré, L. and Yao, A.F. (2001) Un modèle semi-paramétrique pour variable aléatoire hilbertienne. *C.R. Acad. Sci. Paris*, **t.327**, série I, 947-952.

Ferré, L. (1998) Determining the dimension in Sliced Inverse Regression and Related Methods. *J. Am. Statist. Ass.*, **93**, 132-140.

Ferré L. and Villa N. (2003) Multi-layer Neural Network with Functional Inputs. Preprint.

Ferré, L. and Yao, A. F. (2003) Functional Sliced Inverse Regression analysis. *Statistics*, to appear.

Leurgan, S., Moyeed, R. and Silverman, B. W. (1993) Canonical correlation analysis when data are curves. *J. R. Statist. Soc.* B, **63**, 393-410.

Li, K. C. (1991) Sliced Inverse Regression for dimension reduction. *J. Amer. Statist. Ass.*, **86,** 316-342.

Li, K. C. (1992) On principal Hessian directions for data visualisation and dimension reduction : another application of Stein's lemma. *Ann. Statist.*, **87**, 1025-1039.

Ramsay, J. O. and Silverman, B. W. (1997) *Functional Data Analysis*, New-York : Springer Verlag.

Schott, J.R. (1994) Determining the dimensionality in sliced inverse regression. *J. Amer. Statist. Ass.*, **89,** 141-148.

Vapnick, V.N. (2000) *The nature of statistical learning*, 2nd ed. New-York : Springer-Verglag.

Velilla, S. (1998) Assessing the number of linear component in a general regression problem. *J. Amer. Statist. Ass.*, **93,** 1088-1098.

Yin X. and Cook R. D. (2002) Dimension reduction for the conditional kth moment *J. R. Statist. Soc.* B, **64**, 159-175.

Xia Y., Tong H., LI W.K. and Zhu L.X. (2002) An adaptative estimation of dimension reduction space *J. R. Statist. Soc.* B, **64**, 1-28.

La classification numérique au service de la géographie linguistique: brève présentation de la dialectométrie

Hans Goebl

Institut für Romanistik
Universität Salzburg
Akademiestraße 24
5020 Salzburg,
Autriche,
hans.goebl@sbg.ac.at

Que l'on ne s'étonne pas que les sciences humaines, elles aussi, se servent de méthodes quantitatives en matière de classification et de typologie! Une des applications les plus éloquentes en la matière est fournie par les différentes philologies modernes qui toutes, depuis de longues années, disposent d'instruments de recherche appelés "atlas linguistiques". Or, un atlas linguistique n'est rien d'autre un ouvrage cartographique en format in-folio qui renseigne, pour un réseau plus ou moins équidistant de N localités (souvent de petite taille), comment l'on y prononce un certain nombre (p) de concepts préalablement choisis. La manière de collecter les données géolinguistiques en question est très simple: il suffit de se rendre sur les lieux et d'y poser toujours les mêmes questions ("Comment dites-vous pour l'écureuil?") aux dialectophones aussi patients que chevronnés dans l'idiome local.

Or, la structure d'un atlas linguistique ressemble de très près à celle d'une matrice bidimensionnelle (N fois p) et permet, de ce fait, tous les calculs respectifs. Malheureusement les géolinguistes – en particulier ceux qui pratiquent les études romanes – ont mis près d'un siècle à découvrir cette possibilité prometteuse. Que ce soit par anti-mathématisme ou simple inadvertance, peu importe.

Ce n'est qu'en 1973 qu'a été mis sur pied, de la part du dialectologue français Jean Séguy, le terme (et aussi l'idée) d'une "dialectométrie" (DM). Depuis, la DM a pris un essor considérable, surtout à cause de la prise en compte parallèle des apports de la classification numérique, de la cartographie statistique et des multiples possibilités visualisatrices qu'offre l'informatique moderne.

Le propos central de notre conférence est de présenter les problèmes et méthodes de la DM actuelle à l'aide d'un jeu de méthodes-DM standards et d'exemples applicatifs tirés de l'"Atlas linguistique de la France" (ALF) qui, lui, constitue le plus grand atlas linguistique du domaine roman (N: 638 points d'enquête; p: 1421 cartes d'atlas). Les méthodes-DM illustrées vont de la visualisation de vecteurs isolés de la matrice de similarité (N fois N) jusqu'au calcul de dendrogrammes et à leur spatialisation concomitante.

Dans notre conférence nous nous servirons de transparents multicolores et aussi, par le biais d'un vidéoprojecteur, d'un logiciel-DM très puissant (VDM: "Visual DialectoMetry") qui permet non seulement de gérer les données atlantographiques de base et d'effectuer

rapidement les différents calculs-DM, mais aussi de visualiser non moins rapidement les résultats numériques des calculs-DM respectifs.

Un accent particulier sera mis sur les filiations interdisciplinaires de la DM qui vont de la géographie quantitative jusqu'à la génétiques des populations.

Barycentric Representation of Profiles in Correspondence Analysis: Some New Results

Willem J. Heiser

Department of Psychology
Leiden University
P.O. Box 9555
2300 RB Leiden, The Netherlands
Heiser@Fsw.LeidenUniv.nl

RESUME: *The geometry of correspondence analysis is discussed without considering dimension reduction, but nevertheless profiting from a weighted least squares framework. An interpretation is given to the profile-to-vertex distances in the barycentric representation of profiles that is characteristic for the classic correspondence analysis formulation. A new type of supplementary point, called the shadow point, is defined and its properties are discussed.*

MOTS-CLES: *multinomial distributions, center of gravity, chi-squared distance, inertia, concentration index, unfolding, biplot, affine geometry*

1. Introduction

There has been some debate about the correct interpretation of distances between row elements and column elements in a joint display of a correspondence table. The conventional view is that we can scale this joint display in such a way that either the distances between rows can be interpreted, or the distances between columns, but never directly the distances between rows and columns ([HEI 83]; [GRE 87]). [CAR 86] proposed an alternative scaling of the coordinates for which they claimed that both between-set and within-set squared distances could be interpreted, but [GRE 89] has shown that this claim is not warranted. This paper tries to clarify the issue, and proposes a new data reconstruction method on the basis of distances.

2. Four roads to the same formulas

As is well-known, there are several ways to reach the formulas that are characteristic for correspondence analysis. Each approach considers a specific type of data, and a specific goal of analysis. First, we have the classic French approach, where the data are a set of profiles, which can be relative frequencies (multinomial distributions), chemical compositions, time budgets or the like, and where the goal is to approximate the within-set distances between these profiles. Second, there is the less familiar French approach also known as contiguity

analysis ([LEB 69], [LEB 01]), which was independently studied in [HEI 81], where the data are a similarity relation or a bipartite graph, and the goal is to embed the graph in Euclidean space with minimal between-set distances. Third, we have the classic Anglo-Saxon optimal or dual scaling approach, where the data are two (or more) categorical variables, and the goal is to find quantifications (optimal scalings) as to maximize their homogeneity in terms of the canonical correlation. Fourth, and finally, there is a less familiar approach where one studies how bivariate distributions can be approximated with sets of orthogonal functions.

The first two approaches are asymmetric in their treatment of rows and columns, and will be studied here; the last two approaches treat rows and columns symmetrically, and for them, the issue of how to deal with interest distances has been studied recently by [NIS 03].

3. Asymmetric treatment of rows and columns: The barycentric representation

Before any dimension reduction, the original asymmetric representation of the data in correspondence analysis is a barycentric configuration of profile points with respect to the unit profiles, which are hypothetical profiles for which all mass is concentrated in one cell. It is shown that a between-set distance interpretation is possible in any barycentric configuration or plot. The data can be reconstructed perfectly with the between-set distance in comparison with the distance towards some specific supplementary point, called the *shadow point*. The shadow point of profile *i* with respect to vertex *j* is the location of the center of gravity for *i* if column *j* is not taken into account. The distance involved is not of the chi-squared type, but simply Euclidean. The result is equally valid in the full-dimensional space as in a reduced space obtained by projection, because affine transformations preserve ratios of distances.

4. Bibliography

[CAR 86] CARROLL, J.D., GREEN, P.E., & SCHAFFER, C.M., Interpoint distance comparisons in correspondence analysis. *Journal of Marketing Research*, 23, 1986, p. 271-280.

[GRE 89] GREENACRE, M.J., The Carroll-Green-Schaffer scaling in correspondence analysis: A theoretical and empirical appraisal. *Journal of Marketing Research*, 26, 1989, p. 358-365.

[GRE 87] GREENACRE, M.J. & HASTIE, T., The geometric interpretation of correspondence analysis. *Journal of the American Statistical Association*, 82, 1987, 437-447.

[HEI 81] HEISER, W.J., *Unfolding Analysis of Proximity Data*. Unpublished doctoral dissertation, Leiden University, The Netherlands, 1981.

[HEI 83] HEISER, W.J. & MEULMAN, J., Analyzing rectangular tables with joint and constrained multidimensional scaling. *Journal of Econometrics*, 22, 1983, p. 139-167.

[LEB 69] LEBART. L., Analyse statistique de la contiguite. *Publication de l'Institut de Statistiques de l'Universite de Paris*, 28, 1969, p. 81-112.

[LEB 01] LEBART L., Representing words and texts through contiguity analysis, In: G. Govaert, J. Janssen, N. Limnios (eds), *ASMDA2001, Applied Stochastic Models and Data Analysis.* UTC, Compiègne, 2001, pp. 654-659.

[NIS 03] NISHISATO, S. & CLAVEL, J.G., A note on between-set distances in dual scaling and correspondence analysis. *Behaviormetrika*, 30, 2003, p. 87-98.

Visualisation de graphes et de partitions

Ludovic Lebart

CNRS - ENST
46 rue Barrault
75013, Paris, France
lebart@enst.fr

RÉSUMÉ. Etant donnés n objets décrits par p variables, la série des graphes de leurs k plus proches voisins (0 < k < n), puis les valeurs spectrales des matrices associées à ces graphes, et enfin les analyses de contiguïté correspondantes sont autant d'outils permettant d'explorer la configuration des n point dans l'espace à p dimensions. Ces approches, intermédiaires entre la classification et l'analyse en axes principaux, fournit également des outils de visualisation de partitions ou de familles de partitions.

MOTS-CLÉS : Visualisation de partitions. Spectre Laplacien d'un graphe. Analyse de contiguïté. Visualisation de partitions. Graphes des plus proches voisins.

1. Introduction

Pour construire une représentation visuelle d'une partition de *n* objets décrits par *p* variables, il existe deux approches courantes :

(A) Construire la partition en s'efforçant d'optimiser un critère, puis, dans un second temps, représenter les classes dans un graphique plan.

(B) Construire simultanément la partition et la représentation, ce qui induit des contraintes sur la partition, mais peut conduire à une meilleure représentation.

La seconde approche est justifiée ou soutenue par les deux arguments suivants :

Dans beaucoup d'applications, la partition n'est en fait qu'une dissection, c'est-à-dire un découpage qui peut isoler des vraies classes, mais qui n'est pas dépourvu d'arbitraire. Si plusieurs partitions sont également valides, pourquoi ne pas choisir, parmi celles-ci, celle qui donne la meilleure représentation visuelle ?

Le second argument est voisin du premier, mais plus technique : puisque dans l'état actuel des algorithmes de partitionnement, on n'obtient que des optima locaux, est-on sûr que les contraintes supplémentaires (imposées par une représentation plane des classes, par exemple) nuisent vraiment à la qualité de la partition ?

Un exemple caractéristique de la démarche (A) consiste à faire une partition classique (k-means, nuées dynamiques, ou partitionnement mixte : classification hiérarchique, coupure du dendrogramme, optimisation de la coupure par réaffectation du type k-means) puis à représenter les classes (par leurs centres, et/ou leurs enveloppes convexes, et/ou des ellipses de densité) dans le plan (1, 2) d'une analyse en axes principaux du tableau (n, p), ou dans le plan (1, 2) d'une analyse discriminante de la partition. Un exemple caractéristique de la démarche (B) est donné par les cartes auto-organisées de Kohonen [KOH 89], [COT 97], [THI 97]. La série des graphes des « k plus proches voisins » (k variant de 1 à n-1) (de même que la série des graphes définis par un seuil de distance croissant) va permettre de définir des sous-espaces de représentations permettant de visualiser plusieurs partitions.

2 Variance locale, graphes de contiguïté

Soient *n* objets décrits par *p* variables, conduisant à une *(n, p)* matrice **X**. Les *n* objets sont aussi les sommets d'un graphe symétrique ***G*** dont la matrice *(n, n)* associée est **M** ($m_{ii'} = 1$ si les sommets *i* et *i'* sont joints par une arête, $m_{ii'} = 0$ sinon). ***G*** peut être externe (données géographiques), interne (graphe des k plus proches voisins [*ppv*], ou graphe des couples d'éléments situés à une distance $d < d_k$), comme ce sera le cas ici.

y étant une variable aléatoire prenant ses valeurs sur chaque sommet *i* de ***G*** , ayant *m/2* arêtes, une première définition de la variance locale $v^c(y)$ est:

$$v^c(y) = (1/2m)\sum m_{ii'}(y_i - y_{i'})^2$$

Notons que si ***G*** est un graphe complet, $v^c(y)$ n'est rien d'autre que *v(y)*, la variance empirique classique. Quand les observations sont distribuées aléatoirement sur le graphe, $v^c(y)$ et *v(y)* estiment tous deux la variance de *y*.

Le coefficient de contiguïté *c(y)*, [GEA 54] , s'écrit : $c(y) = v^c(y) / v(y)$. Une valeur du coefficient $c(y) << 1$ indique une autocorrélation spatiale positive pour la variable *y*. Une modification de la définition du coefficient *c(y)* ([MOM 88], [ESC 89]) va rendre la variance locale compatible avec la variance "intra" (*within*) quand le graphe décrit une partition des observations (i.e. une série de cliques [sous-graphes complets]).

La variance locale sera redéfinie comme:

$$v^*(y) = (1/n)\sum (y_i - m_i^*)^2$$

Dans cette dernière formule, la *moyenne locale* est définie comme :

$$m_i^* = (1/n_i)\sum_k m_{ik} y_k$$

C'est la moyenne des valeurs adjacentes au sommet *i*. Si ***G*** est régulier : $v^*(y) = v^c(y)$.

2. Principaux résultats

On note par **N** la *(n,n)* matrice diagonale ayant le degré de chaque sommet *i* comme élément diagonal n_i (n_i dénote ici n_{ii}). **y** est le vecteur dont la $i^{ème}$ composante est y_i.

Le nouveau *c(y)* s'écrit alors : $c(y) = \mathbf{y}'(\mathbf{I} - \mathbf{N}^{-1}\mathbf{M})'(\mathbf{I} - \mathbf{N}^{-1}\mathbf{M})\mathbf{y} / \mathbf{y}'\mathbf{y}$.

1 – Pour un graphe régulier ***G*** donné [LEB 73] : *Min [c(y)]* = $(1- \sqrt{\lambda_{max}})^2$, où λ_{max} est la plus grande valeur propre de l'analyse des correspondances (AC) de la matrice **M** associée au graphe, y étant le vecteur propre associé. Cette propriété qui rend compte du bon pouvoir descriptif de l'AC, a été depuis étudiée et utilisée ans un cadre plus général [KOR 02].

2 – Plus généralement, que le graphe soit régulier ou non, le spectre de la matrice : **N-M** (*matrice Laplacienne* du graphe, [MOH 91]) a d'importantes propriétés relatives à la structure du graphe, l'ordre de multiplicité de la valeur propre nulle (valeur propre 1 en AC de **M**) étant le nombre de composantes connexes du graphe. Le rapprochement avec l'opérateur de Laplace est déjà dans [BEN 73]. Ces propriétés concernent en particulier les nombres chromatiques du graphe et de nombreuse inégalités [CHU 97], [MOH 97].

3 – Généralisation à des observations multivariées sur un graphe [LEB 69].
Si **X** désigne la *(n,p)* matrice donnant les valeurs de *p* variables pour chacun des *n* sommets du graphe, décrit par sa matrice associée **M**, la matrice des covariances locales s'écrit :

$$\mathbf{V^*} = (1/n)\ \mathbf{X'(I - N^{-1}M)' (I - N^{-1}M) X}$$

Soit **u** un vecteur définissant une combinaison linéaire *u(i)* des *p* variables pour le sommet *i*:
Avec les notations précédentes, la variance locale de la variable *u(i)* vaut : *v*(u)* = **u' V* u**.
Le coefficient de contiguïté de cette combinaison linéaire s'écrit : *c(u)* = **u' V* u / u' V u**, où **V** est la matrice des covariances classique. La recherche de **u** qui minimise *c(u)* donne des fonctions de contiguïté minimale, dont les fonctions discriminantes de Fisher constituent un cas particulier lorsque le graphe est formé de plusieurs graphes complets. C'est l'Analyse de Contiguïté, qui est aussi un "*projection pursuit algorithm*" [BUR 91], ou une recherche de projection privilégiée dans l'esprit des travaux de [ART 82].

4 – Chacun des graphes $\mathbf{G}_k$ associé aux k *ppv* (ou à un seuil de distance d_k) pourra donc, d'une part être décrit par ses valeurs spectrales, d'autre part donner lieu à une analyse de contiguïté, conduisant à un second spectre qui implique directement la matrice **X**. L'évolution de ces spectres en fonction de k permet de choisir les nombres de *ppv* ou le seuil qui peuvent correspondre à des partitions ou à des espaces de représentation intéressants [LEB 00].

5 – Les plans principaux correspondants fournissent simultanément une visualisation des observations, avec d'éventuels *dépliages* (*unfoldings*) permis par le caractère non-linéaire de l'opération.

6 – On peut ainsi représenter des *classes obtenues sans contrainte* sur le ou les plans principaux de l'analyse de contiguïté (sur de tels plans, chaque point peut être assorti d'une

zone de confiance *bootstrap* par exemple). Plusieurs partitions peuvent être représentées sur le même fond.

3. Bibliographie

[ART 82] ART D., GNANADESIKAN R., KETTENRING J.R., Data Based Metrics for Cluster Analysis, *Utilitas Mathematica*, 1982, 21 A, p 75-99.

[BEN 73] BENZECRI, J.P., *Analyse des Données: Correspondances*. 1973, Dunod, Paris.

[BUR 91] BURTSCHY B., LEBART L., Contiguity analysis and projection pursuit. In*: Appl. Stoch. Mod. and Data Anal*. R. Gutierrez et al., 1991, Eds, World Scientific, Singapore, p 117-128.

[CHU 97] CHUNG F.R.K., *Spectral Graph Theory*. CBMS Reg. Conf. Ser. Math. 92, American Mathematical Society, 1997.

[COT 97] COTTRELL M., ROUSSET P., The Kohonen Algorithm: a powerful tool for analysing and representing multidimensional qualitative and quantitative data. In: *Biological and Artificial Computation : From Neuroscience to Technology*. J. Mira, R. Moreno-Diaz, J. Cabestany, (eds), 1997, Springer, p 861-871.

[ESC 89] ESCOFIER B., Multiple correspondence analysis and neighbouring relation. In: *Data Analysis, Learning Symbolic and Numeric Knowledge*. Diday E. (ed.), 1989, Nova Science Publishers, New York, p 55-62.

[GEA 54] GEARY R.C., The Contiguity Ratio and Statistical Mapping, *The Incorporated Statistician*, 1954, 5, p 115-145.

[KOH 89] KOHONEN T., *Self-Organization and Associative Memory*. 1989, Springer-Verlag, Berlin.

[KOR 02] KOREN Y., CARMEL L., HAREL D., ACE: a Fast Multiscale Eigenvectors Computation for Drawing Huge Graphs, *Proceedings of IEEE Information Visualization, 2002*, p 137-144.

[LEB 69] LEBART L., Analyse Statistique de la Contiguité, *Publ. de l'ISUP*. 1969, XVIII, p 81-112.

[LEB 13] LEBART L., TABARD N. *Recherches sur la description automatique des données socio-économiques*, Rapport CREDOC-CORDES n° 4172, convention. 13-1971, 1973.

[LEB 00] LEBART, L., Contiguity Analysis and Classification, In: W. Gaul, O. Opitz and M. Schader (Eds): *Data Analysis*. 2000, Springer,Berlin, p 233--244.

[MOH 91] MOHAR B., The Laplacian Spectrum of Graphs, *Graph Theory, Combinatorics and Application*, 2, 1991, p 871-898.

[MOH 97] MOHAR B., Some Applications of Laplace Eigenvalues of Graphs, *Graph Symmetry, Algebraic Methods and Application*, Hahn G., Sabidussi G., NATO Ser. C., 497, Kluwer, 1997, p 225-275.

[MOM 88] MOM A., *Méthodologie Statistique de la Classification des réseaux de transport*. Thèse, Université des Sciences et Techniques du Languedoc, 1988, Montpellier.

[THI 97] THIRIA S., LECHEVALLIER Y., GASCUEL O., CANU S., *Statistique et Méthodes Neuronales*, 1997, Dunod, Paris.

Mesures pour la construction de graphes d'induction

Djamel A. Zighed

Laboratoire ERIC
Université Lumière Lyon 2
5, av. Pierre Mendès-France
69676 Bron Cedex
France
zighed@univ-lyon2.fr

RÉSUMÉ. *Les graphes d'induction sont d'un usage fort répandu dans le domaine de la fouille de données (Data Mining). Les modèles les plus utilisés sont les arbres, car ils présentent de multiples avantages : facilité de mise en œuvre, simplicité dans l'interprétation,... Ces avantages ne se perdent pas quand nous faisons appel aux graphes latticiels qui sont plus généraux. Dans cette présentation, je donnerai une formalisation générale de la construction d'un graphe d'induction, que le résultat final, sur un jeu de données, soit un graphe arborescent ou non. J'insisterai davantage sur les critères mis en œuvre dans ce contexte*

MOTS-CLÉS : *Graphes d'induction, arbres de décision, apprentissage, mesures de qualité d'une partition*

1. Introduction

Les arbres d'induction et, plus généralement, les graphes d'induction [ZIG 00] occupent une place privilégiée dans le domaine de la fouille de données (*Data Mining*). Ils sont appréciés pour la simplicité des algorithmes qu'ils utilisent, pour la facilité d'interprétation des résultats qu'ils produisent et pour leur temps de réponse rapide. Ces méthodes permettent la construction d'arbres n-aires avec par exemple ID3 [QUI 86], C4.5 [QUI 93], CHAID [KAS 80], [MOR 63], Arbogodaï [ZIG 03], d'arbres binaires avec les méthodes telles que CART [BRE 84] ou des graphes latticiels comme avec SIPINA [ZIG 92].

Toutes les méthodes de construction fonctionnent selon le même principe. Elles cherchent, au moyen des variables dites explicatives, $X_1, \ldots, X_j, \ldots, X_p$, à engendrer une succession de partitions $P_k(\Omega)$, $k = 0, \ldots, K$, sur l'ensemble d'apprentissage Ω visant à optimiser un critère $\mathcal{I}$. Ce critère mesure généralement le degré de séparabilité des modalités de la variable à prédire. Si cette séparabité est bonne cela signifierait qu'il est possible de prédire Y connaissant seulement les $X_1, \ldots, X_j, \ldots, X_p$). Dans les graphes d'induction, qu'ils soient arborescents ou latticiels, les heuristiques sont généralement descendantes et partent de la partition grossière Ω. Dans les méthodes arborescentes, les partitions engendrées sont de plus en plus fines car elles sont construites par segmentation. Dans les graphes latticiels, le passage d'une partition P_k à la suivante P_{k+1} peut résulter de l'éclatement d'un élément de P_k ou du regroupement de deux ou plus de ses éléments. Ainsi, la partition P_{k+1} peut être plus ou moins fine que P_k. Le processus s'arrête dès qu'il n'y a plus d'amélioration du critère : $\mathcal{I}(P_{k+1}) \leq \mathcal{I}(P_k)$. De nombreux algorithmes proposent des conditions d'arrêt pas toujours liées au critère à optimiser. Par exemple, fixer a priori la taille minimale d'une classe qui devient une condition d'arrêt supplémentaire.

Dans cet article, nous allons examiner deux aspects fondamentaux :

– Le critère d'évaluation d'une partition dans le contexte de l'apprentissage supervisé. Nous ne considérons pas un catalogue des mesures de qualité d'une partition, mais tentons de définir un ensemble de propriétés qui nous paraissent importantes pour mesurer la qualité d'une partition. Nous proposons ensuite une famille

de critères baptisés critères d'incertitude que nous mettrons en perspectives par rapport aux mesures classiquement proposées dans la littérature. Ces mesures sont issues soit de la statistique comme le Khi-Deux ou bien issues de la théorie de l'information comme les mesures d'entropie.

– Les stratégies de passage d'une partition à une autre. En somme, pourquoi serions-nous contraints de procéder par raffinements successifs en partant de la partition grossière pour construire un arbre ? Nous pourrions tout aussi bien choisir un autre point de départ qui peut être la partition la plus fine engendrée par l'arbre maximal issu du croisement de toutes les variables explicatives. Ainsi, nous faisons appel à des stratégies de recherche de partitions ascendantes et nous n'excluons pas les structures latticielles.

Dans la section suivante, nous allons introduire quelques définitions et notations préliminaires. Dans la section 3, nous proposons une liste de propriétés pour caractériser une bonne mesure de qualité. Dans la section 4, nous abordons les stratégies de passage d'une partition à une autre. Nous tentons ainsi de définir une stratégie générique qui engloberait l'ensemble des méthodes à base de graphes d'induction. Nous examinerons en particulier les stratégies ascendantes qui sont encore peu exploitées dans ce contexte. Dans la section 5, nous donnons quelques indications sur des expérimentations encore en cours qui seront exposées ultérieurement.

2. Définitions et notations

Nous nous plaçons dans un contexte d'apprentissage supervisé. Nous souhaitons, au moyen d'un échantillon d'apprentissage Ω dont les individus ω_i, $i = 1, \ldots, n$ sont caractérisés par un ensemble de variables explicatives parfois dites prédictives notées $X_1, \ldots, X_j, \ldots, X_p$, construire un modèle $\theta(X_1, \ldots, X_j, \ldots, X_p; \Omega)$ qui prédit (ou ajuste) au mieux un ensemble de variables $Y_1, \ldots, Y_t, \ldots, Y_m$ dites à prédire. Le cas le plus fréquemment rencontré et celui d'une seule variable à prédire, c'est-à-dire $m = 1$. Pour simplifier les notations, on notera $\mathbf{X} = (X_1, \ldots, X_j, \ldots, X_p)$ et $\mathbf{Y} = (Y_1, \ldots, Y_t, \ldots, Y_m)$. La qualité de la prédiction (ou de l'ajustement) est le plus souvent évaluée au moyen d'un échantillon test Ω' par une mesure d'écart entre ce qui est prédit par le modèle induit et ce qui est attendu 1. Le processus de construction des partition est itératif. On note T le nombre de classes de partition sur Ω obtenue à l'itération k.

$$\Delta((\theta(\mathbf{X}; \Omega')), (\mathbf{Y}; \Omega')) \quad [1]$$

$P_k^T(\Omega)$ désigne la partition en T classes obtenue à l'itération k sur l'ensemble d'apprentissage Ω. $T = 1$ désigne la partition grossière Ω.

$f(\mathbf{Y}/t)$ désigne la distribution de probabilité de $\mathbf{Y}$ sur la classe t.

3. Mesure de qualité sur une partition

Dans ce papier, nous allons considérer le cas classique de la mise en œuvre des méthodes à base de graphes d'induction. Nous supposons que nous avons une seule variable à prédire, c'est-à-dire $\mathbf{Y} = Y$. De plus, Y prend ses valeurs sur un ensemble fini discret $y_1, \ldots, y_i, \ldots, y_m$ sans structure mathématique particulière. Nous considérons également que les variables $\mathbf{X} = (X_1, \ldots, X_j, \ldots, X_p)$ sont toutes discrètes prenant leurs valeurs dans des ensembles finis.

$$X_j : \Pi \mapsto \{x_j^1, \ldots\ x_j^t, \ldots\ x_j^{\alpha_j}\} \quad [2]$$

où Π désigne la population concernée, ainsi $\Omega \subset \Pi$ et α_j désigne le cardinal de l'ensemble des valeurs possibles pour X_j. Précisons d'ores et déjà que quels que soient les types des variables X_j, les domaines des valeurs que nous traitons effectivement en apprentissage supervisé sont toujours finis :

$$|X_j(\Omega)| = \alpha_j \quad [3]$$

et

$$|Y(\Omega)| = m \quad [4]$$

On considère une partition sur Ω en T classes $P_k^T(\Omega)$ notées c_t avec $t = 1, \ldots, T$. Pour alléger la notation on omettra l'indice k s'il n'y a pas d'ambiguïté. On notera $n_{.t}$ l'effectif de la classe c_t de $P^T(\Omega)$ et n_{it} l'effectif de la modalité y_i de la variable à prédire Y dans la classe c_t. On désignera par $n_{i.}$ l'effectif de la modalité y_i. Ainsi, à toute partition $P^T(\Omega)$ sur l'échantillon d'apprentissage, nous pouvons associer un tableau (voir tableau 1) qui s'apparente à une table de contingence $\Theta(m, T)$ dont les T colonnes représentent les éléments de la partition induite par le graphe d'induction sur l'ensemble Ω, et les m lignes, la partition induite par Y sur Ω.

TAB. 1. *Table de contingence variable à prédire $\times$ partition*

Θ	c_1	$\ldots$	c_t	$\ldots$	c_T	
y_1	n_{11}	$\ldots$	n_{1t}	$\ldots$	n_{1T}	$n_{1.}$
$\vdots$						$\vdots$
y_i	n_{i1}	$\ldots$	n_{it}	$\ldots$	n_{iT}	$n_{i.}$
$\vdots$						$\vdots$
y_m	n_{m1}	$\ldots$	n_{mt}	$\ldots$	n_{mT}	$n_{m.}$
	$n_{.1}$	$\ldots$	$n_{.t}$	$\ldots$	$n_{.T}$	n

La partition grossière correspondra à un tableau ayant seulement une seule colonne correspondant à la distribution a priori sur les modalités de Y.

A tout tableau $\Theta(m, T)$; $m \geq 2$; $T \geq 1$ nous souhaitons associer une mesure de qualité positive ou nulle que nous appelons mesure d'incertitude. Il convient peut être de préciser qu'il ne s'agit pas d'une mesure d'entropie car elle ont souvent été baptisées mesure d'incertitude :

$$\begin{aligned} \mathcal{I} : \mathbb{N}^{m \times T} &\longmapsto \mathbb{R}^+ \\ \Theta(m, T) \in \mathbb{N}^{m \times T} &\longmapsto \mathcal{I}(\Theta(m, T)) \geq 0 \end{aligned}$$

Nous souhaitons que cette mesure nous renseigne sur le degré de d'indétermination de Y à partir des classes de la partition. Par exemple, si chaque classe de la partition est associée à une seule modalité de Y alors l'indétermination est nulle. Cela signifie que $P(Y = y_i / P_k^T(\Omega)) = 1$ pour tout $y_i \in \{y_1, \ldots, y_m\}$. En revanche, dans le cas où la connaissance de la classe ne nous apporte aucune connaissance sur Y, c'est-à-dire $P(Y = y_i / P_k^T(\Omega)) = \frac{1}{m}$ pour tout $y_i \in \{y_1, \ldots, y_m\}$ alors l'indétermination est maximale. C'est pour ces raisons que nous retenu le terme d'incertitude. Dans ce qui suit nous allons énumérer quelques propriétés qui nous paraissent importantes. Signalons, que de nombreux auteurs ont proposé des caractérisations analogues mais pas nécessairement avec la même approche. Parmi eux on peut citer les travaux de [RAF 95] qui ont conduit au critère BIC, [WEH 90] qui a défini un critère proche de celui que nous proposons ou [AKA 83].

Minimalité L'incertitude $\mathcal{I}$ devra être minimale si chaque classe de la partition ne comporte que des individus appartenant à la même modalité de Y. Autrement dit, $\mathcal{I}$ est minimal, si $\forall t$; $t = 1, \ldots, T$, $\exists i$; $i \in \{1, \ldots, m\}$ tels que $n_{it} = n_{.t}$. On notera que si $\mathcal{I}$ est minimale, cela ne signifie pas que $\mathcal{I}$=0. Cela veut dire que parmi toutes les distributions possibles sur le tableau, seules celles qui vérifient $n_{it} = n.t$ pour tout t sont minimales.

Maximalité L'incertitude $\mathcal{I}$ devra être maximale si dans chaque classe de la partition, les effectifs sont équirépartis sur l'ensemble des modalités. Autrement dit, $\mathcal{I}$ est maximale si $\forall t$; $t = 1, \ldots, T$ et $\forall (i, j)$; $(i, j) \in \{1, \ldots, m\}^2$ nous avons $n_{it} = n_{jt}$.

Sensibilité à l'effectif Soit $\Theta(m, T)$ un tableau de contingence de m lignes et T colonnes associé à une partition. Si les effectifs du tableau sont multipliés par un facteur $\alpha > 1$ alors la mesure d'incertitude $\mathcal{I}$ devra diminuer. Autrement dit, pour $\alpha > 1$, $\mathcal{I}(\alpha \times \Theta(m, T)) < \mathcal{I}(\Theta(m, T))$. Cette propriété permet de mieux contrôler le processus de généralisation et rend la comparaison inter-partitions possible sur des populations différentes.

Sensibilité à la complexité Nous dirons qu'une partition est plus complexe qu'une autre si le nombre de classes de la première est plus grand que celui de la seconde. Si une partition possède deux ou plusieurs classes ayant la même distributions sur Y, le regroupement de ces classes identiquement distribuées conduit à une nouvelle partition identique à la précédente du point de vue des distributions mais de complexité plus faible. De ce fait, la propriété de sensibilité à la complexité permet de gérer la réduction de la complexité. Si nous réduisons la complexité d'une partition par regroupement de classes identiquement distribuées, la valeur de l'incertitude devra diminuer. Formellement, nous pouvons réécrire $\mathcal{I}(\Theta(m,T))$ en fonction des colonnes c_t : $\mathcal{I}(\Theta(m,T)) = \mathcal{I}(c_1,\ldots,c_t,\ldots,c_T)$. Cette propriété de sensibilité à la complexité signifie que si nous considérons c_k et c_t deux classes de la partition $P^T(\Omega)$ telles que $|c_k| = \alpha|c_t|$ où $\alpha > 0$ alors $\mathcal{I}(\ldots,c_t,\ldots,c_k,\ldots) > \mathcal{I}(\ldots,c_t+c_k,\ldots)$. Nous pouvons encore dire qu'à distributions équivalentes, la partition de complexité plus faible devra être meilleure au sens du critère retenu.

Insensibilité à toutes permutations des classes ou des modalités Lorsqu'il n'existe pas d'ordre naturel sur les classes, c'est-à-dire que toute permutation sur les colonnes est possible on pourra alors énoncer la propriété qui suit : si nous permutons les lignes et/ou les colonnes en bloc, la valeur du critère devra rester identique. Si nous reprenons l'écriture du critère en fonction des colonnes, nous aurons $\mathcal{I}(c_1,\ldots,c_t,\ldots,c_T) = \mathcal{I}(c_{\sigma_1},\ldots,c_{\sigma_t},\ldots,c_{\sigma_T})$ où σ est une permutation sur les colonnes. Identiquement, nous pouvons écrire la même expression pour les lignes, $\mathcal{I}(y_1,\ldots,y_i,\ldots,y_m) = \mathcal{I}(y_{\sigma_1},\ldots,y_{\sigma_i},\ldots,y_{\sigma_m})$ ou σ est une permutation sur les lignes.

Indépendance Lors du passage d'une partition $P_k^T(\Omega)$ à la partition suivante $P_{k+1}^U(\Omega)$, la variation du critère ne devra dépendre que des classes éclatées ou regroupées. $\mathcal{I}(\ldots,c_t,\ldots,) - \mathcal{I}(\ldots,c_{t_0},c_{t_1}\ldots) = f(c_t,c_{t_0},c_{t_1})$, où $c_{t_0}; c_{t_1}$ sont les classes issues de l'éclatement ou bien les classes regroupées. Cette propriété est particulièrement utile sur le plan informatique car elle permet de travailler localement de manière indépendante d'une classe à une autre.

Toutes ces propriétés s'inspirent en partie des propriétés des mesures d'entropie [ACZ 75].

Nous proposons une famille de mesures vérifiant les propriétés énoncées. Nous donnons quelques unes de ces propriétés sans démonstration. Considérons un paramètre $\lambda > 0$.

$$\mathcal{I}(P^T(\Omega)) = -\sum_{t=1}^{T} \frac{n_{.t}}{n} \sum_{i=1}^{m} \frac{n_{it}+\lambda}{n_{.t}+m\lambda} \log \frac{n_{it}+\lambda}{n_{.t}+m\lambda} \quad [5]$$

$$\mathcal{I}(P^T(\Omega)) = \sum_{t=1}^{T} \frac{n_{.t}}{n} \sum_{i=1}^{m} \frac{n_{it}+\lambda}{n_{.t}+m\lambda} (1 - \frac{n_{it}+\lambda}{n_{.t}+m\lambda}) \quad [6]$$

On peut noter que si $\lambda = 0$ nous retrouvons les mesures d'entropies classiques. Généralement nous fixons $\lambda = 1$. Nous retrouvons ainsi l'estimateur des probabilités de Laplace $\frac{n_{it}+\lambda}{n_{.t}+m\lambda}$. En fait, ce résultat est très général et dit que pour toutes les mesures d'entropie, utilisant les estimateurs des probabilités de Laplace au lieu de celui du maximum de vraisemblance, vérifient les propriétés requises.

4. Stratégies de recherche de partitions

Le recherche de la meilleure partition sur un ensemble est écartée d'office pour les problèmes de complexité en temps de calcul. Nous ne pouvons considérer que des heuristiques et c'est ce que nous faisons dans ce papier.

Dans un but prédictif, nous allons chercher la partition qui minimise la mesure d'incertitude. Considérons une partition sur l'ensemble d'apprentissage $P^T(\Omega)$ en T classes. Si $T = 1$ cela signifie que nous sommes en présence de la partition grossière qui contient Ω. Si $T = \alpha = \prod_{j=1}^{p} \alpha_j$ cela signifie que nous sommes en présence de la partition la plus fine que l'on peut engendrer par segmentation en utilisant toutes les variables prédictives $(X_1,\ldots,X_p)$. On rappelle que α_j désigne le nombre de valeurs distinctes de X_j observées sur Ω. A l'étape k, la partition $P_k^T(\Omega)$ a une incertitude $\mathcal{I}(P_k^T(\Omega))$. Pour passer à l'itération $(k+1)$ nous avons le choix entre raffiner la partition par segmentation d'une de ses classes c_t au moyen de l'une des variables prédictives X_j soit

de réduire sa complexité par regroupement de deux éléments c_{t_1} et c_{t_2}. Pour passer de $P_k^T(\Omega)$ à $P_{k+1}^U(\Omega)$ il est nécessaire que $\mathcal{I}(P_{k+1}^U(\Omega)) < \mathcal{I}(P_k^T(\Omega))$. L'algorithme va consister à rechercher parmi les partitions possibles par segmentation ou par regroupement, celle qui minimise le critère. Le processus s'arrête dès qu'aucune amélioration n'est possible. Cette stratégie a été adoptée dans l'algorithme SIPINA [ZIG 92] mais de manière exclusivement descendante. Nous proposons d'explorer également les partitions selon un autre parcours qui partirait par exemple de la partition la plus fine et qui effectuerait des agrégations visant à minimiser l'incertitude. Cette approche que nous avons commencé à expérimenter [RIT 03], [MUH 01] donne des résultats intéressants qui sont en cours de validation et que nous présenterons ultérieurement.

5. Conclusion et perspectives

Dans ce papier, nous avons voulu résumer nos travaux sur les graphes d'induction en proposons à la fois des critères répondant à certaines exigences pratiques et en offrant de nouvelles stratégies exploratoires pour construire ces partitions. Certes, l'approche ascendante risque d'altérer un peu la lisibilité que nous avons avec les arbres. Cependant, les règles induites resteront toujours aisées à lire puisqu'elles s'exprimerons toujours sous la forme de règles logiques de la forme **SI** *condition* **alors** *conclusion*. Où *condition* est une disjonction de conjonction de prédicats logiques et *conclusion* est une distribution de probabilités sur les modalités de la variable Y à prédire. Nos premiers résultats montrent une certaine stabilité et surtout une meilleure généralisation en utilisant des approches non arborescentes.

6. Bibliographie

[ACZ 75] ACZÈL J., DOROCZY Z., *On measures of Information and Their Characterizations*, vol. 115 de *Mathematics in Science and Engineering*, Academic Press, 1975.

[AKA 83] AKAIKE H., Information Measures and Model Selection, *Bulletin of the International Statistical Institue*, vol. 50, 1983, p. 277–290.

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

[KAS 80] KASS G. V., An exploratory technique for investigating large quantities of categorical data, *Applied Statistics*, vol. 29, n^o 2, 1980, p. 119–127.

[MOR 63] MORGAN J. N., SONQUIST J. A., Problems in the Analysis of Survey Data, and a Proposal, *Journal of the American Statistical association*, vol. 58, 1963, p. 415–434.

[MUH 01] MUHLENBACH F., ZIGHED D. A., D'HONDT S., Génération de règles par compression, *Extraction des connaissances et apprentissage*, vol. 1, n^o 1-2, 2001, p. 93–104.

[QUI 86] QUINLAN J. R., Induction of decision trees, *Machine learning*, , n^o 1, 1986, p. 81-106.

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

[RAF 95] RAFTERY A. E., Bayesian Model Selection in Social Research, MARSDEN P., Ed., *Sociological Methodology*, p. 111-163, The American Sociological Association, Washington, DC, 1995.

[RIT 03] RITSCHARD G., Partition BIC optimale de l'espace des prédicteurs, *RNTI*, vol. , n^o 01, 2003, p. 99-110, SFDS 2003.

[WEH 90] WEHENKEL L., Une approche de l'intelligence artificielle appliquee a l'evaluation de la stabilite transitoire des reseaux electriques, PhD thesis, Faculte des Sciences Appliquees - Universite de Liege, 1990.

[ZIG 92] ZIGHED A., AURAY J., DURU G., *SIPINA : Methode et logiciel*, Lacassagne, 1992.

[ZIG 00] ZIGHED D. A., RAKOTOMALALA R., *Graphes d'induction : apprentissage et data mining*, Hermes Science Publications, Paris, 2000.

[ZIG 03] ZIGHED D., RITSCHARD G., ERRAY W., SCUTURICI V.-M., Abogodaï, a New Approach for Decision Trees, , Ed., *Principles of Data mining and Knowledge Discovery*, vol. de *LNCS*, , 2003, , Springer Verlag, page , To appear.

Classification connexe

Catherine Aaron

SAMOS-MATISSE
Université Paris 1
75013 Paris
catherine_aaron@hotmail .com

RÉSUMÉ. On s'intéresse ici à la construction d'une méthode de classification non supervisée sous la seule hypothèse de connexité des classes. Cette hypothèse est l'une des plus générales qu'on puisse faire sur la forme des classes. Après avoir défini une notion de connexité adaptée a des espaces discrets on montrera que la classification hiérarchique par la distance minimum mène à l'obtention de classes connexes. On définira alors une distance intra classe rendant compte de la connexité afin de mettre au point une méthode de choix du nombre de classes.

MOTS-CLÉS : classification, non-supervisée, connexité

1. Algorithme de classification

On dispose d'un ensemble de N points de $\mathbb{R}^P$: $E = \{x_1, ..., x_N\}$ que l'on souhaite segmenter en classes connexes du point de vue d'une distance d .Etant donnée la nature discrète du problème la notion de connexité, capacité à lier deux points d'un ensemble par un chemin continu de point de l'ensemble, doit être adaptée aux espaces discrets.

1.1. N*otion de connexité et espaces discrets*

On définit ici une notion de connexité par seuil comme il suit : E est $\delta - connexe$ si et seulement si on peut lier tous ses couples de points par un chemin constitué de points de E deux à deux distants d'au plus δ .

On montre alors que pour tout δ il existe une unique partition $\delta - connexe$ minimale (au sens ou le nombre de classe est minimal, soit encore que tout regroupement de classes n'est pas $\delta - connexe$). On notera alors $p(\delta)$ le nombre de classe associé à une telle segmentation. La fonction p qui, à δ , associé $p(\delta)$ est alors une fonction en escalier, décroissante, à valeur dans $\{1, ..., N\}$ On notera :

- $\delta_{\min}(k) = borne \inf\{\delta / p(\delta) = k\}$: plus petit seuil pour un nombre de classe donné.
- $\delta_{\max}(k) = borne \sup\{\delta / p(\delta) = k\}$: plus grand seuil pour un nombre de classe donné.

Avec $\delta_{\min}(k) = \delta_{\max}(k+1)$

1.2. C*lassification hiérarchique associée*

On montre que la classification hiérarchique par la distance minimum[1] entre deux ensembles va mener à l'obtention de toutes les classifications $\delta - connexes$ minimales possibles. En effet pour passer d'une classification minimale en $k+1$ classe à une classification minimale en k classes, on montre que la seule possibilité est le regroupement des deux classes les plus proches au sens de la distance minimum. Ce qui correspond à l'algorithme de classification de classification hiérarchique par la distance min.

[1] Les classifications hiérarchiques ascendantes et descendantes sont équivalentes dans le cas du choix de la distance minimum

2. Distance intra-classe

Une fois l'ensemble des segmentations connexes maximales effectuées on souhaite obtenir un indicateur permettant de choisir les segmentations les plus significatives. Pour cela on se base sur les méthodes classiques de minimisations de la distance intra classe, mais ici les notions de distances intra-classes moyennes de type euclidienne ou, de manière similaire les minimisations de variances intra classes, ne sont pas pertinentes.

2.1. *Distance entre deux points*

Pour tenir compte de la notion de connexité, la distance entre deux points doit faire intervenir tous les autres points de l'ensemble. L'exemple ci-contre illustre ce propos : dans les deux cas les points A et B sont à la même distance (euclidienne) mais, du point de vue de la connexité, les situations sont très différentes.

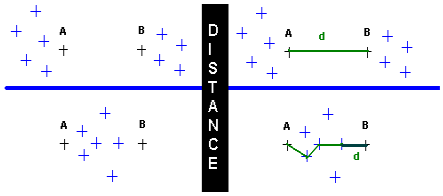


Figure 1. *Distance entre deux points compatible avec la notion de connexité*

Pour résoudre ce problème on choisira, pour distance entre deux points dans un ensemble E le plus petit des sauts maximums effectués lorsqu'on lie les points par un chemin de points de E :

$$d_c^E(x_i,x_j)=\min_{\pi\in chem(x_i,x_j)}\left(\max_{1\le i<N}(d(x_{\pi(i)},x_{\pi(i+1)})\right)$$

avec $chem(x_i,x_j)=\{\pi,\ \text{application de } \{1,....N\} \text{ dans } \{1,....N\} \text{ avec : } \pi(1)=i,\pi(N)=j\}$

En notant $Cl_k(i)$ la classe de l'élément x_i de E obtenue lors d'une partition minimale en k classes connexes on a :

$d_c^E=\delta_{\max}(\arg\min\{k / Cl_k(i)=Cl_k(j)\})$ soit le plus petit seuil de connexité pour lequel les points x_i et x_j sont agrégés.

2.2. *Distance intra classe*

Une fois définie d_c on obtient aisément une distance intra-classe pour une segmentation en k classes par : $D_{\text{intra}}(k)=\dfrac{1}{N_k}\sum_i\sum_{\substack{j\neq i\\ x_j\in C_k(i)}}d_c(x_i,x_j)$ avec $N_k=\sum_i\sum_{\substack{j\neq i\\ x_j\in C_k(i)}}1$

On a aussi : $D_{\text{intra}}(k)=\dfrac{\sum_{i=k}^{N-1}\hat{N}_k\delta_{\max}(k)}{\sum_{i=k}^{N-1}\hat{N}_k}$

Avec $\widehat{N}_k=\sum_{i=1}^{N}\sum_{j=1}^{N}1_{\{d_c^E(x_i,x_j)=\delta_{\max}(k)\}}$ soit le nombre de couples distants de $\delta_{\max}(k)$

Une telle écriture de la distance intra classe permet le calcul de toutes les distances intra classes des segmentations successives à l'issue de la classification hiérarchique sans augmentation significative du temps de calcul.

Pour déterminer le nombre de classes optimum on se propose de rechercher la segmentation correspondant à la plus grande rupture de distance intra classe

3. Résultats

3.1. *Classes connexes et convexes*

Dans les exemples suivants : séparation de gaussiennes, base de Ruspini ou Iris de Fischer, les classes sont à la fois connexes et convexes et peuvent, en conséquence, être séparée par d'autres méthodes de classification (voir figure 2)

3.2. *Classes connexes et non convexes*

Dans cet exemple : reconnaissance de classes formées par des cercles imbriqués et bruités la seule caractéristique des classes est la connexité, le regroupement autour de barycentres, comme dans les $K-means$ sera inopérante. On voit ici que la classification hiérarchique permet de retrouver les classes initiales tant que le bruitage n'est pas trop important et que, de plus, le critère du maximum de saut de distance intra est un bon indicateur du nombre de classes (voir figure 3)

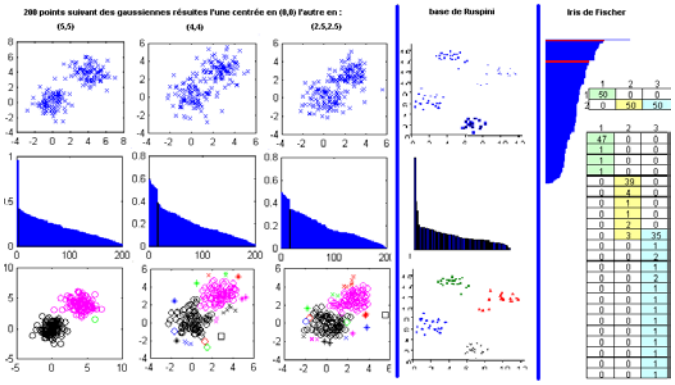


Figure 2.

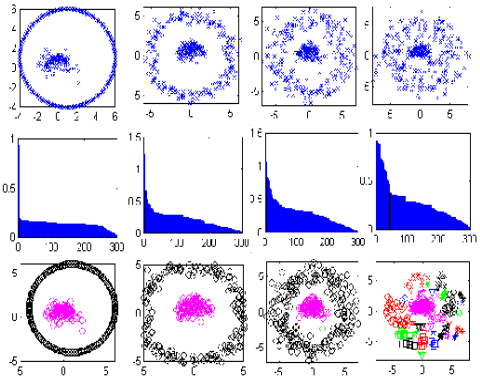


Figure 3.

Exemples de résultats sur des classes connexes, pour chaque exemple excepté les Iris de Fisher on lit de haut en bas : représentation de la base initiale, diagramme des distances intra et résultats de la classification. Dans le cas des Iris de Fischer, la dimension de l'espace étant 4 on ne représente que le diagramme des distances intra et le croisement entre les classifications retenues et les « vraies classes »

4. Construction d'un test de significativité sur la rupture de distance intra classe dans le cas gaussien

Du fait de la complexité de l'expression de la distance intra classe, le taux de significativité du saut maximum de distance intra classe : $saut = \max_{k\in\{1\ldots N/2\}} \frac{D_{\text{intra}}(k-1) - D_{\text{intra}}(k)}{D_{\text{intra}}(k-1)}$ [2] va être estimé par simulation. Dans le cas présent (test d'existence d'une unique classe de forme gaussienne) on a effectué, pour des tailles d'échantillon (N) variant de 5 à 100 individus et pour des dimensions (P) variant entre 1 et 10, 500 tirages sur lesquels on a estimé la valeur du saut.

On remarque empiriquement que :

- Les moyennes et écarts types des sauts diminuent en fonction de N et P. Des estimations de la moyenne et de l'écart type par des fonctions du type :

$$x(N,P) = \exp[-(a\ln(N) + b(\ln(P) + c\ln(N)\ln(P) + d)]$$

donnent de très bons résultats (R2 de 99,8% pour la moyenne et de 98% pour l'écart type, T de Student supérieurs à 9 pour la moyenne, à 4,6 pour l'écart type). Voire tableau 1 pour les résultats numériques.

[2] On prend le maximum sur les premières classifications (de 1 à N/2) pour éviter les problèmes de bords qui apparaissent lorsqu'on scinde la base en un nombre trop élevé de classes.

- La distribution des sauts (centrés normés) semble vite converger en N et en P avec. On estime alors ces différents quantiles d'ordre α (voir tableau 2)

	a	b	c	d
Moyenne	*0.21*	*0.26*	*0.24*	*0.42*
Ecart type	*0.18*	*0.31*	*0.38*	*1.1*

Tableau 1. *Estimation de la moyenne et de l'écart type du saut*

α	90%	95%	97%	99%
Q_α	*1.5*	*2*	*2.4*	*3.2*

Tableau 2. *Quantiles d'ordre alpha de la répartition du saut centré normé en fonction de N et P*

On construit ainsi un test de rejet de l'hypothèse d'existence d'une unique classe de forme gaussienne à $\alpha\%$ si la valeur centrée réduite du saut dépasse le quantile d'ordre α

5. limites de la méthode

Dans le cas où les « vraies » classes C_i vérifieraient des conditions de non séparabilité conjointe, c'est à dire que pour isoler une classe on est obligé d'en scinder une autre, soit encore que :

$$\exists i \neq j \,/\, \exists k \,/ \min_{x_i \in C_i, x_j \in C_j} (d(x_i, x_j)) < \max_{x_{k_1} \in C_k} (\min_{x_{k_2} \in C_k} (d(x_{k_1}, x_{k_2}))$$

Alors la classification hiérarchique ne pourra, en aucuns cas, retrouver conjointement les bonnes. Dans les meilleurs cas (par exemple les exemples des trois premières séparations de gaussiennes, les iris de Fischer ou les premiers cercles concentriques) on isolera les points les plus éloignés de leur classe pour obtenir, au final, une classification satisfaisante. Dans le pire des cas, si : $\exists i \neq j \,/\, \exists k \,/ \min_{x_i \in C_i, x_j \in C_j} (d(x_i, x_j)) < \min_{x_{k_1} \in C_k} (\min_{x_{k_2} \in C_k} (d(x_{k_1}, x_{k_2}))$ alors les classes i et j ne pourront être séparées que si la classe k est scindée en singletons. Ce cas extrême peut arriver dans des cas de très grandes disparités

6. Perspectives

Du point de vue de la modélisation d'une liaison du type $y = f(x_1, ..., x_n)$ avec f continue, la connexité de $(X_1, ..., X_n)$ est nécessaire, dans le cas contraire des raccordements seront à envisager. De plus la non-connexitéé de l'espace $(Y, X_1, ..., X_n)$ mettra en défaut l'existence de la fonction unique f continue.

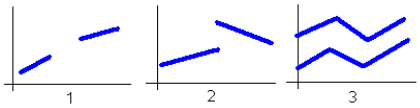


Figure 4. *Exemples pour l'application à la modélisation, dans le premier cas l'espace des X n'est pas connexe dans les autres c'est l'espace (X,Y)*

3. Bibliographie

[BAL 65] BALL G., HALL D., *ISODATA a novel method of a data analysis and pattern classification*, rapport, 1965, Stanford Research Institute.

[SAL 96]., SALEMBIER P., OLIVERAS A., "practical extension of connected operators" *Mathematical Morphology and its application to image and signal processing*, Kluwer, 1998, p 191-206

[WEM 99] WEMMERT C., GANCARSKII J., KORCZAK J ," Un système de raffinement non-supervisé d'un ensemble de hiérarchie de classes ", *Sebag*, 1999, p. 153-160.

Analyse Discriminante Multiple Relationnelle

Rafik Abdesselam

Laboratoire GEMMA (UMR-CNRS 6154)
Université de Caen
Esplanade de la Paix, F-14032 CAEN Cedex
abdesselam@econ.unicaen.fr

RÉSUMÉ. *Une méthode factorielle d'analyse de données évolutives dans un but de discrimination et de classement est présentée. L'Analyse Discriminante Multiple Relationnelle (ADMR) proposée est basée sur la recherche de moments principaux dans l'espace des individus muni d'une distance relationnelle. On mesure ainsi, en terme d'inertie, la liaison observée durant une période partielle ou globale, entre les ensembles évolutifs de variables explicatives et l'ensemble des variables indicatrices associées aux modalités d'une variable à expliquer. Il s'agit ici d'un problème relevant d'une analyse discriminante où la contrainte consiste à imposer aux facteurs discriminants d'appartenir au sous-espace engendré par l'ensemble des observations mesurées durant la période considérée. Un exemple sur données réelles est présenté.*

MOTS-CLÉS : *Analyse discriminante, distance relationnelle, rapport de corrélation, données évolutives.*

1. Introduction

Une analyse factorielle discriminante sur données évolutives est proposée. L'Analyse Discriminante Multiple Relationnelle (ADMR) est basée sur la notion de distance relationnelle, introduite par Y. Schektman, dans l'espace des individus. L'ADMR est ensuite comparée à une autre analyse proposée, L'Analyse Discriminante Multiple (ADM) qui est présentée comme une extension de l'Analyse Factorielle Multiple (AFM), introduite par B. Escofier et J. Pagès, sur tableaux de moyennes évolutives. Ces analyses sont spécialement conçues pour étudier une population d'individus caractérisés par un certain nombre de groupes de variables, c'est-à-dire de tableaux de données évolutives. 0n distingue deux, voire plusieurs groupes d'individus définis a priori par une variable nominale et sur lesquels ont été mesurées les mêmes variables continues à différents instants. L'ojectif est de caractériser l'évolution de ces groupes d'individus en fonction de ces mesures répétées dans le temps. Un exemple d'application sur données évolutives réelles issues du domaine de l'agronomie est présenté.

2. Distances et analyses dans un Modèle Euclidien Relationnel

On utilisera les notations suivantes pour construire la matrice M associée à la distance de référence dans l'espace E des individus dans le cas de deux sous-espaces.

$E_t = E_{x(t)} = IR^p$ étant le sous-espace des individus, associé par dualité aux p variables continues centrées $\{x^j(t); j = 1, p\}$,
$E_y = IR^q$ étant le sous-espace des individus, associé par dualité aux variables indicatrices centrées des modalités de la variable nominale y, notées $\{y^k; k = 1, q\}$,
X_t est la matrice d'ordre (n, p) des données explicatives à l'instant t associée aux variables $\{x^j(t); j = 1, p\}$,
$Y_{(n,q)}$ est la matrice des données associée à l'ensemble des variables $\{y^k; k = 1, q\}$,
M_y [resp. M_t] est la matrice de la distance dans l'espace E_y [resp. E_t], isomorphe du sous-espace de même nom, via l'injection canonique notée Iny [resp. $Inx(t)$].

On pose $M_t = {}^t\mathbf{Inx}(\mathbf{t})M\mathbf{Inx}(\mathbf{t})$ et $M_y = {}^t\mathbf{Iny}M\mathbf{Iny}$, où (M_t, M_y) est un couple de distances euclidiennes, M est une distance relationnelle dans $E = E_t \oplus E_y$, relativement aux variables $\{x^j(t)\}$ et $\{y^k\}$, et est notée $R[M_t, M_y]$, si et seulement si :

$$M_{ty} = {}^t\mathbf{Inx}(\mathbf{t})M\mathbf{Iny} = M_t[(V_tM_t)^{\frac{1}{2}}]^+V_{ty}M_y[(V_yM_y)^{\frac{1}{2}}]^+ \quad [1]$$

où, $V_{x(t)} = V_t = {}^tX_tDX_t$, $V_y = {}^tYDY$ et $V_{ty} = {}^tX_tDY$ désignent les matrices de variances et covariances, $D = (1/n)I_n$ est la matrice diagonale des poids des n individus et I_n la matrice unité d'ordre n.

Pour définir la matrice associée à la distance dans l'espace des individus $E = E_1 \oplus \ldots \oplus E_t \oplus \ldots \oplus E_T \oplus E_y$, on utilisera aussi les notations suivantes :
$X_{(n,Tp)} = [X_1, \ldots, X_t, \ldots, X_T]$ est la matrice des données évolutives : juxtaposition des T tableaux X_t à n individus et p variables. X peut être partielle [resp. globale] si T est inférieur [resp. égal] au nombre total d'ensembles de mesures répétées dans le temps,
$E_x = \oplus\{E_t\}_{\{t=1,T\}} = IR^{Tp}$ étant le sous-espace des individus, associé par dualité aux T ensembles de variables explicatives centrées $\{x^j(t); j = 1, p\}_{\{t=1,T\}}$.

Pour mesurer l'association, en terme d'inertie d'un nuage de points dans E, entre les ensembles de variables, et comme il s'agit d'analyse discriminante, on choisit des distances de Mahalanobis $M_t = V_t^-$ comme distances intra dans tous les sous-espaces explicatifs E_t, on élimine ainsi les effets variances-corrélations à l'intérieur de chaque sous-ensemble de variables, et $M_y = \chi^2_y$ la distance du khi-deux dans E_y.

Les deux analyses discriminantes multiples proposées diffèrent principalement par le choix des distances inter les sous-espaces explicatifs E_t : à effet relationnel pour l'ADMR et à effet relationnel nul pour l'ADM.

2.1. *Analyse Discriminante Multiple Relationnelle - ADMR*

L'ADMR consiste à utiliser l'expression relationnelle [1] entre tous les couples d'ensembles de variables $\{x^j(t); j = 1, p\}_{t=1,T}$ et $\{y^k; k = 1, q\}$,

$$\begin{cases} R_t = V_t^- & pour\ t = 1, T \\ R_{tt'} = R[V_t^-, V_{t'}^-] = V_t^-V_{tt'}V_{t'}^- & pour\ t \neq t' \\ R_{ty} = R[V_t^-, \chi^2_y] = V_t^-V_{ty}\chi^2_y \end{cases}$$

la matrice $R = R[V_1^-, \ldots, V_t^-, \ldots, V_T^-, \chi^2_y] = R[R_x, \chi^2_y]$ relationnelle équilibrée d'ordre $Tp+q$, est associée à la distance dans $E = E_x \oplus E_y$.

On note P_x^R l'opérateur de projection R-orthogonale sur $E_x = \oplus\{E_t\}_{\{t=1,T\}}$,
$N_g^R(x/y) = \{P_x^R(e_k(y)\,;\, k = 1, q\} \subset E$ est le nuage des q points centres de gravité,
$N_{\tilde{x}} = \{\tilde{x_i} \subset E_x\,;\, i = 1, n\}$ est le nuage points-individus actifs associé au tableau $\tilde{X} = X\ Diag[V_t^-]\ R_x^-$ et dont $N_g^R(x/y)$ est le nuage points-centres de gravité dans le sous-espace explicatif $E_x = \oplus\{E_t\}_{\{t=1,T\}}$.

Définition 1

L'ADMR proposée consiste à effectuer, dans le MER, l'ACP suivante :

$$ACP\,[\,\{P_x^R(e_k(y))\,;\, k = 1, q\}\,;\, R\,;\, D_y\,]. \quad [2]$$

REMARQUE. — Les représentations simultanées et barycentriques de l'ADMR sont les projections orthogonales, sur les plans principaux de l'ACP [2], du nuage $\{P_x^R[e_k(y)], k = 1, q\} \cup \{\,[\,\tilde{x_i}, 0\,] \subset E\,;\, i = 1, n\}$.

2.2. *Analyse Discriminante Multiple - ADM*

Pour l'ADM, la relation [1] est utilisée uniquement entre les T couples d'ensembles de variables explicatives $\{x^j(t); j = 1, p\}_{t=1,T}$ et à expliquer $\{y^k; k = 1, q\}$,

$$\begin{cases} M_t = V_t^- & pour\ t = 1, T \\ M_{tt'} = R[V_t^-, V_{t'}^-] = 0 & pour\ t \neq t' \\ M_{ty} = R[V_t^-, \chi_y^2] = V_t^- V_{ty} \chi_y^2 \end{cases}$$

la matrice $M = [R(V_1^-, \chi_y^2), \ldots, R(V_t^-, \chi_y^2), \ldots, R(V_T^-, \chi_y^2)] = M[M_x, \chi_y^2]$ semi-relationnelle d'ordre $Tp + q$, est associée à la distance dans l'espace des individus $E = E_x \oplus E_y$ où, la distance $M_x = Diag[V_t^-]$ dans $E_x = \oplus\{E_t\}_{\{t=1,T\}}$ est à effet relationnel nul.

On appliquant le lemme 1 pour $t = 1, T$, c'est-à-dire en projetant le nuage N_y sur chacun des T sous-espaces E_t, on obtient ainsi les T AFD instantanées séparées avec leurs éléments propres $\{(\lambda_r(t)\,,\, u_r(t))\}_{\{t=1,T\}}$: les moments discriminants non nuls et les vecteurs axiaux correspondants.

On note P_x^M l'opérateur de projection M-orthogonale sur $E_x = \oplus\{E_t\}_{\{t=1,T\}}$,
$N_g^M(x/y) = \{P_x^M(e_k(y))\,;\, k = 1, q\} \subset E$ est le nuage des q centres de gravité, associé au tableau de moyennes évolutives $G = [\,G_1, \ldots, G_t, \ldots, G_T, 0\,]$: juxtaposition des T tableaux $G_t = \chi_y^2 V_{yt}$.

Lemme 1

Si $M = [R(V_1^-, \chi_y^2), \ldots, R(V_t^-, \chi_y^2), \ldots, R(V_T^-, \chi_y^2)]$ alors $P_x^M[e_k(y)] = g(x/y^k)$.

Afin d'équilibrer l'influence des différents groupes de variables du tableau G des centres de gravité, on introduit la pondération proposée dans [ESC 88] sur laquelle est basée la méthode AFM. Elle consiste ici à donner à chacune des variables d'un même groupe un poids égal à l'inverse du premier moment ou pouvoir discriminant de l'AFD instantanée : le poids ou encore le coefficient de pondération des variables à l'instant t est égal à $1/\lambda_1(t)$.

Définition 2

L'ADM revient à effectuer l'ACP pondérée suivante : $ACP\,[\,\{P_x^M(e_k(y))\,;\, k = 1, q\}\ ;\ M_p\ ;\ D_y\,]$. [3]

avec $M_p = [\,R(\frac{V_1^-}{\lambda_1(1)}, \chi_y^2), \ldots, R(\frac{V_t^-}{\lambda_1(t)}, \chi_y^2), \ldots, R(\frac{V_T^-}{\lambda_1(T)}, \chi_y^2)\,]$

REMARQUE. — Les représentations simultanées et barycentriques de l'ADM sont les projections orthogonales du nuage $\{P_x^M(e_k(y))\,;\, k = 1, q\} \cup \{x_i = [\,x_i(1), \ldots, x_i(T), 0\,]\,;\, i = 1, n\}$ sur les plans principaux de l'ACP [3].

NOTE. — L'ADM est équivalente à l'AFM du tableau G des moyennes évolutives : centres de gravité dans $E_x = \oplus\{E_t\}_{\{t=1,T\}} = IR^{pT}$. L'ADMR ne nécessite aucune pondération ; la distance relationnelle positionne les sous-espaces évolutifs E_t tel que l'on puisse traduire en terme d'inertie dans E, la structure des corrélations observées entre les sous-espaces de variables.

REMARQUE. — Dans [2] et [3], les nuages des barycentres $N_g^R(x/y)$ et $N_g^M(x/y)$ sont dans le même sous-espace explicatif $E_x = \oplus\{E_t\}_{\{t=1,T\}} = IR^{pT}$ de l'espace euclidien des individus $E = E_x \oplus E_y$. Le critère de comparaison de ces deux analyses est celui du rapport maximum de l'inertie inter-classe sur l'inertie totale. Pour l'AMDR, l'inertie inter-classe $I[N_g^R(x/y)]$ est égale au rapport de corrélation généralisé.

3. Exemple d'application

Les données analysées proviennent de l'Institut National de la Recherche Agronomique (INRA - Angers). L'objectif est de caractériser l'évolution de trois variétés de pommes : (Golden, Fuji et Braeburn) en fonction de deux caractéristiques : la teneur en sucre (TS) et en acidité (TA) de pommes récoltées à environ 15 jours d'intervalle avant, pendant et après la date de maturité optimale : (prématurité, maturité et postmaturité).

Les principaux résultats de comparaison des deux analyses multiples partielles : jusqu'à maturité, sur variables centrées et réduites, sont présentés en parallèle.

TS1	TA1	TS2	TA2	ADMR ⟵ Variétés ⟶ ADM	TS1	TA1	TS2	TA2
1.312	-1.378	-0.876	-1.532	Golden	0.460	-0.265	0.385	-0.454
-0.168	0.136	1.000	1.435	Fuji	0.768	-1.015	0.859	1.331
-1.144	1.241	-0.124	0.097	Braeburn	-1.227	1.280	-1.243	-0.877

Tableau ⋁ 1. Coordonnées des centres de gravité dans $E_x = E_1 \oplus E_2$

F1	F2	ADMR ⟵ Facteurs ⟶ ADM	F1	F2
0.96	0.70	Pouvoirs discriminants	1.95	0.70
57.85	42.15	Pourcentages	73.52	26.48
F1	F2	ADMR ⟵ Contributions ⟶ ADM	F1	F2
0.01	66.66	Golden	0.85	65.81
49.53	17.13	Fuji	43.08	23.59
50.46	16.21	Braeburn	56.07	10.60

Tableau 2. Facteurs - Contributions (%) des points moyens aux facteurs

Maturité : T = 2	ADMR	ADM
B-Inter-classe	1.6564	2.6566
W-Intra-classe	0.3436	0.7082
T-Totale	2.0000	3.3274
Expliquée	82.82%	79.84%

Tableau 3. Critère d'inertie

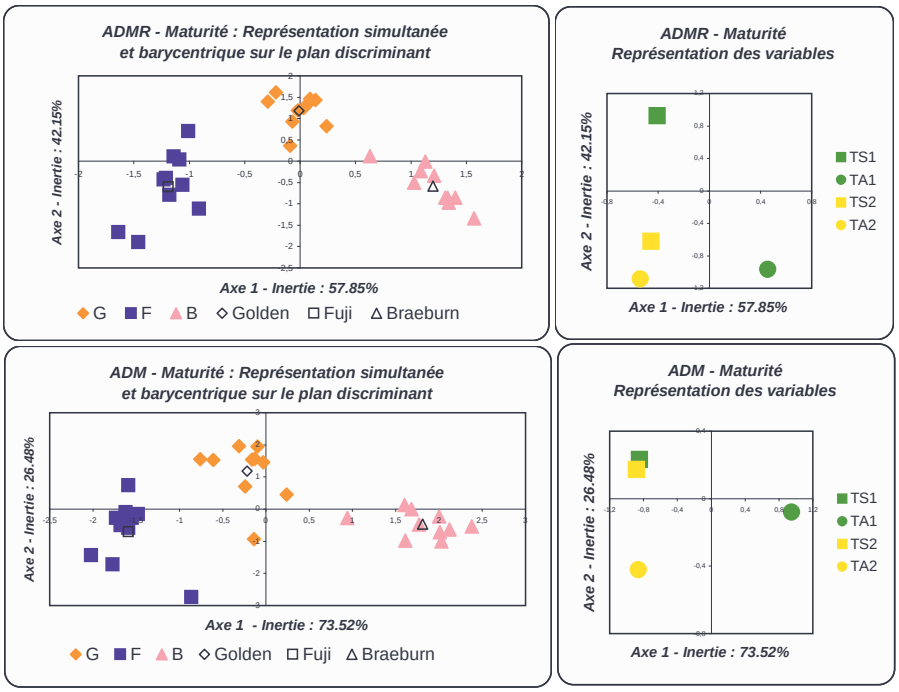


Figure 1. Représentations graphiques sur le plan discriminant

4. Bibliographie

[SCH 87] Schektman, Y., *A general euclidean approach for measuring and describing associations between several sets of variables*, Academic press, Tokyo, 1987.

[ESC 88] Escofier, B. and Pagès, J.P., *Analyses factorielles simples et multiples*, Dunod, 1988.

[CAS 95] Casin, Ph., L'analyse discriminante de tableaux évolutifs, Revue de Statistique Appliquée, vol. XLIII(3), 1995, , p. 73-91.

[DAZ 96] Dazy, F. and Lebarzic, J.F., *L'analyse des données évolutives*, Edts technip, 1996.

[ABD 96] Abdesselam, R. and Schektman, Y., Une analyse factorielle de l'association dissymétrique entre deux variables qualitatives, Revue de Statistique Appliquée, vol. XLIV(2), 1996, , p. 5-34.

[SCH 00] Schektman, Y. and Abdesselam, R., *A Geometrical Relational Model for Data Analyses*, W.Gaul, O.Opitz, M.Schader Editors, by Springer, 2000.

Comparaison de classifications non supervisées de données microbiologiques

S. REGIS, A. DONCESCU, J. AGUILAR-MARTIN et J. DESACHY

Groupe DISCO LAAS-CNRS Toulouse / LBB-CNRS UMR5504 GBA-INSA Toulouse
'Equipe GRIMAAG Université Antilles-Guyane Pointe-à-Pitre
sregis,adoncesc@laas.fr,

RÉSUMÉ. *Nous présentons dans ce papier les résultats de différentes classifications non supervisées sur des données biotechnologiques. Les méthodes utilisées sont classiques ou innovantes et nous permettent de tirer une première conclusion quant à la classe d'outils de classification réellement pertinents et efficaces pour ce type de données.*

MOTS-CLÉS : *Classifications non supervisées ; fermentation*

1. Introduction

L'avancée considérable des recherches biologiques et l'essor des biotechnologies au cours de ces dernières décennies, offrent de nouveaux champs d'étude et d'applications pour les mathématiques et l'informatique tant dans leurs domaines classiques qu'innovants. Les données que nous étudions sont issues de procédés biotechnologiques industriels réalisés dans un bioréacteur. Ces biotechnologies ont été développées au Laboratoire de Biotechnologies-Bioprocédés de Toulouse[1]. Deux voies sont utilisées pour optimiser ces bioprocédés : la modélisation mathématique non linéaire et la classification ; l'objectif de ces deux voies étant de détecter les états physiologiques des micro-organismes utilisées dans le bioréacteur. Bien que la tendance soit à la fusion ou à la coopération de ces deux voies, nous nous interesserons uniquement à la classification. La méthode utilisée pour ces bioprocédés est une méthode non supervisée baptisée LAMDA[MAR 80][VIL 00], développée au Laboratoire d'Analyse et d'Architecture des Systèmes de Toulouse. Cette méthode fournit des résultats satisfaisants, cependant elle n'a pas réellement été comparée à d'autres méthodes non supervisées, plus classiques. Nous présentons dans ce papier une comparaison empirique de plusieurs classificateurs non supervisées (LAMDA, LVQ, kernel ACP+LVQ) sur ces données biotechnologiques. La section 2 présente plus amplement les données à traiter. Puis nous présentons dans la section 3 les différentes méthodes. Enfin les sections 4 et 5 présentent les résultats et les perspectives.

2. Les données biotechnologiques

Les données biotechnologiques que nous cherchons à classifier sont des signaux numériques issus d'un procédé de fermentation alcoolique utilisant la levure Saccharomyces Cerevisiae) . Il s'agit de mesures de paramètres biochimiques effectués à des intervalles de temps réguliers. Les paramètres biochimiques les plus pertinents sont : le CO2, l'O2, le Ph et la Luminance. Ces données sont représentées sous forme de vecteur où chaque composante d'un vecteur correspond à la valeur d'un paramètre biochimique à un instant donné t. Le nombre de vecteurs qui dépend du temps total du bioprocédé et de la fréquence de mesure des paramètres, est ici égal à 1012. Les données n'ont subi aucun traitement ni filtrage et aucune hypothèse n'est faite sur la nature du bruit éventuellement présent

1. Nous tenons à remercier les équipes du LBB pour leur aide et leur soutien

au niveau de ces données. Les experts cherchent à identifier 3 états physiologiques principaux des levures qui sont dans l'ordre : la fermentation (état 1), la diauxie (oxidation/redusction)(état 2) et l'oxydation (état3). Au niveau des classifications ces états correspondent à une classe ou un regroupement de plusieurs classes.

3. Les Classificateurs

Les classificateurs sont les suivants : la méthode LVQ, le classificateur LAMDA, et la méthode kernel ACP+LVQ.

3.1. *La méthode LVQ*

La méthode LVQ (Learning Vector Quantization)[KOH 92] s'inspire de la même démarche que les cartes auto organisatrices sauf qu'ici, il n'y a qu'un seul neurone qui représente le centre de la classe. La méthode LVQ peut s'utiliser tant en classification supervisée que non supervisée mais nous ne nous intéresserons qu'à la version non supervisée. L'apprentissage est réalisé sur un certain nombre d'échantillons. Pour chaque échantillon E, on cherche le neurone C_i le plus proche, puis celui-ci est modifié comme suit :

$$C_i = C_i + \alpha.(C_i - E) \quad [1]$$

où α est un réel positif et strictement inférieur à 1 qui décroit au fur et à mesure des itérations. Par rapport aux cartes auto organisatrices cette méthode a l'avantage de simplifier les résultats puisqu'il n'y a pas de notion de voisinage entre les neurones ni de difficulté d'interprétation de l'espace de représentation. Cependant elle reste une méthode non déterministe qui dépend fortement de l'initialisation des neurones, et des échantillons choisis.

3.2. *La méthode LAMDA*

Nous ne ferons qu'une présentation succinte de LAMDA (Learning Algorithm for Multivariate Data Analysis) car celle ci a déjá été présentée plus en détails dans [WAI 98][NAK 02]. L'originalité de LAMDA est de concilier les avantages de la loi bayesienne avec la structure des methodes neuronales tout en utilisant des mesures floues. LAMDA repose sur l'aggrégation d'informations marginales, chaque information marginale étant calculée grâce à la généralisation floue de la loi binômiale suivante :

$$\rho_{ji}^{1-\alpha(x_i,c_{j,i})}(1-\rho_{ji})^{\alpha(x_i,c_{j,i})} \quad [2]$$

où $c_{j,i}$ représente la composante i du centre c_j de la classe J, x_i est la composante i de l'élément x à classer, $\rho_{i,j}$ est la probabilité qu'un élément appartienne á la classe c_j et $\alpha(x_i, c_{i,j})$ représente la distance entre x_i et $c_{i,j}$.LAMDA traite séquentiellement les éléments et n'a pas de phase d'apprentissage : LAMDA n'a donc pas besoin d'échantillons. La grande particularité de LAMDA est son interactivité avec l'utilisateur : celui-ci peut en effet regrouper certaines classes (grâce à une visualisation simple et conviviale de la classification), et réitérer une nouvelle classification pour le reste des éléments. Cette méthode se rapproche de la philosophie de certains travaux[ANK 00, POU 01] mais ici l'intervention de l'utilisateur se fait en aval de la classification et non pendant celle-ci. Cependant, dans dans notre exemple, nous n'utiliserons pas l'interactivité avec l'expert afin de rester dans le cadre d'une classification non supervisée. Le principal problème de la classification provient des incertitudes au niveau des transitions des classes : celles-ci proviennent à la fois du bruit présent et des données elles-mêmes. Une méthode utilisant la transformée en ondelettes a été proposée pour résoudre ce problème[NAK 02]. Cette amélioration consiste à introduire dans la classification les points d'inflexions les plus significatifs des signaux biochimiques. Ces points sont détecté grâce au maximum du module de la transformée en ondelettes et sont introduits sous forme de fonction par palier dans la classification. Ce ajout permet non seulement de résoudre le problème des incertitudes au niveau des transition entre classes mais aussi de réduire le nombre de classes qui peut être tres important dans la classification LAMDA 'classique'.

3.3. *La méthode kernel ACP+LVQ*

On entend par kernel ACP l'utilisation d'un noyau dans une ACP (Analyse en Composante Principale) classique. La fonction noyau permet l'immersion des données dans un espace de dimension supérieur et permet ainsi l'analyse d'éventuelles dépendances non linéaires entre ces données[MUL 01]. Cette propriété des fonctions noyau (en fait de noyaux remplissant certaines conditions) sont souvent utilisées dans les SVM (Support Vector Machines). La fonction noyau que nous utilisons est un noyau polynomial :

$$(x.y)^2 \qquad [3]$$

où x et y sont deux vecteurs représentant les données. Nous ne détaillerons pas plus la méthode kernel ACP (voir[MUL 01]), mais le principal avantage par rapport à l'ACP classique est, comme nous venons de le voir la détection de dépendances non linéaires. Son principal inconvénient est d'utiliser une matrice carrée de taille n$\times$n où n désigne un nombre d'échantillons alors que l'ACP utilise la matrice de covariance dont la taille est d$\times$d où d désigne la dimension de l'espace des données (ici d=4). La kernel ACP fournit une pré-classification pour la classification LVQ. Les directions principales de la kernel ACP servent à déterminer les centres pour l'initialisation de la classification LVQ.

4. Résultats Numériques et comparaisons

Nous donnons dans le tableau suivant le pourcentage de classifications correctes. Il est difficile de définir une classification correcte car cette notion reste relativement subjective car liée à l'expert. Nous avons choisi de prendre comme référence les résultats de la classification obtenue avec la version de LAMDA améliorée [NAK 02] : on suppose que la classification est correcte et qu'on a ainsi 100% d'éléments bien classés. En particulier, les frontières des classes sont parfaitement définies. Pour la classification LVQ et la préclassification kernel ACP, nous avons choisi d'utiliser le même jeu d'échantillons (de 155 données) représentatifs des 3 états. On utilise 3 centres pour la classification LVQ qui sont les moyennes des 3 états physiologiques fournis par LAMDA améliorée.

Méthodes	% Classication
LVQ	58,39%
LAMDA	94,47%
kernel ACP+LVQ	71,83%

La méthode LAMDA fournit les meilleurs résultats cependant le nombre de classes est important (33 classes). Les résultats de LVQ indiquent que la connaissance statistique et numérique des paramètres est insuffisante pour déterminer les états physiologiques. La méthode kernel ACP montre que l'étude des corrélations entre ces paramètres biochimiques (corrélations non analysées par LAMDA) permet de differencier (mais pas de facon précise) les états 1 et 3. L'état 2 n'est pas détecté par la kernel ACP. En fait aucune des 3 méthodes ne détecte parfaitement l'état 2. Il semble que la détection de l'état 2 soit liée aux connaissances de l'expert. Il apparait donc que la méthode LAMDA dans sa version non supervisée donne les meilleurs résultats.

5. Conclusion

Nous avons présenté les résulats d'une comparaison empirique de plusieurs classifications non supervisées. Ces résultats tendent à montrer que les connaissances de l'expert sont indispensables pour une bonne classification. Ces connaissances de l'expert peuvent être utilisées soit de facon totalement implicite grâce à l'interactivité (c'est le cas dans la méthode LAMDA) soit de facon explicite par une modélisation interprétable de ces connaissances de l'expert (grâce à des systèmes experts ou la logique inductive par exemple). Nos prochains travaux porteront sur cette dernière alternative afin de mieux expliciter les connaissances des experts.

6. Bibliographie

[ANK 00] ANKERST M., ESTER M., KRIEGEL H.-P., Towards an Effective Cooperation of the User and the Computer for Classification, *in proc. of KDD'2000*, Boston, 2000, p. 179-188.

[KOH 92] KOHONEN T., KANGAS J., LAAKSONEN J., TORKKOLA K., Lvq-pack : A program package for the correct application of learning vector quantization algorithms, *IEEE International Joint Conference on Neural Networks*, Baltimore, June 1992, p. 725-730.

[MAR 80] MARTIN J. A., BALSSA M., MANTRAS R. D., Estimation récursive d'une partition. Exemple d'apprentissage et auto apprentissage dans **R**, rapport n^{o}880139, 1980, LAAS-CNRS.

[MUL 01] MULLER K.-R., MIKA S., RATSCH G., TSUDA K., SCHOLKOPF B., An Introduction to Kernel-Based Learning Algorithms, *IEEE Transaction on Neural Networks*, vol. 12, n^o 2, 2001, p. 181-202.

[NAK 02] NAKKABI Y., REGIS S., DESACHY J., DONCESCU A., ROUX G., Apport de la Transformée en Ondelettes pour affiner les résultats de classifications, *IXe Rencontre de la Société Francophone de Classification*, Toulouse, Septembre 2002, p. 287-291.

[POU 01] POULET F., CIAD : Construction Interactive d'Arbres de Décision, *VIIIe Rencontre de la Société Francophone de Classification*, Pointe-à-Pitre, Décembre 2001, p. 275-282.

[VIL 00] VILANOVA J. W., Construction d'un modèle comportemental pour la supervision de procédés : application à une station de traitement des eaux, PhD thesis, LAAS - CNRS, November 2000.

[WAI 98] WAISSMAN-VILANOVA J., AGUILAR J., DAHHOU B., ROUX G., Généralisation de degré d'adéquation marginale dans la méthode de classification LAMDA, *VIemes Rencontres de la Société Francophone de Classification*, 1998.

Comportement d'exposition solaire : recherche d'une typologie

Julie Latreille* - Emmanuelle Mauger* - Christiane Guinot* -

Denis Malvy[$] - Laurence Ambroisine* - Erwin Tschachler*[,§]

** CE.R.I.E.S., 20, rue Victor Noir, 92521 Neuilly sur Seine, France*
Téléphone : 33 (0)1 46 43 47 71, télécopie : 33 (0)1 46 43 46 00,
E-mail : {julie.latreille, emmanuelle.mauger, christiane.guinot, laurence.ambroisine}@ceries-lab.com
[$] *EA2323, Centre René Labusquière, ISPED, Université Victor Segalen Bordeaux 2, 146 rue Léo Saignat, Bordeaux, 33076 Cedex, France*
E-mail : jean-marie-denis.malvy@chu-bordeaux.fr
[§] *Département de Dermatologie, Université de Vienne, Vienne, Autriche*
E-mail : Erwin.Tschachler@akh-wien.ac.at

RÉSUMÉ. Dans le but d'identifier des indicateurs qui peuvent permettre de quantifier l'exposition au soleil et les habitudes de protection face au soleil, un questionnaire auto-administré sur les connaissances des dangers du soleil, les habitudes d'exposition et de protection solaire a récemment été développé dans le cadre de l'étude épidémiologique SU.VI.MAX. Afin de définir une stratégie pour l'analyse des données de cette cohorte, une étude pilote a été réalisée sur un échantillon de taille restreinte. Le lien entre les variables relatives à l'exposition solaire a été étudié grâce à une analyse des correspondances multiples, puis une typologie a été recherchée à partir des composantes principales grâce à la méthode de Ward. Quatre comportements d'exposition au soleil ont été identifiés. Les résultats de cette étude pilote nous confortent dans l'idée que notre questionnaire est approprié pour classer des individus en fonction de leur connaissance des risques et de leurs habitudes d'exposition et de protection face au soleil, dans le but de pouvoir identifier et cibler des groupes d'individus pour des campagnes d'information de santé publique et/ou pour des études d'intervention. L'analyse décrite ci-dessus sera reconduite sur les données du questionnaire collecté auprès des 13017 volontaires de la cohorte nationale SU.VI.MAX, indépendamment pour les femmes et pour les hommes.

MOTS-CLÉS : Analyse des correspondances multiples, méthode de Ward, questionnaire auto-administré

1. Introduction

Les rayons ultraviolets sont connus pour jouer un rôle prépondérant dans le développement des tumeurs cutanées. Néanmoins, l'augmentation de la durée des vacances, la facilité des voyages et la mode du bronzage ont entraîné ces cinquante dernières années une plus grande exposition au soleil des populations humaines [ART 95]. Dans le but d'identifier des indicateurs permettant de quantifier l'exposition au soleil et les habitudes de protection face au soleil un questionnaire auto-administré sur les connaissances des dangers du soleil, les comportements d'exposition au soleil et d'utilisation de produits de protection solaire a récemment été développé pour l'étude SU.VI.MAX (SUpplémentation en VItamines et Minéraux Anti-oXydants), une étude épidémiologique expérimentale d'intervention nutritionnelle qui s'intéresse aux grandes pathologies chroniques caractéristiques des pays industrialisés [HER 98]. Dans le but de définir une stratégie pour l'analyse des données de

cette cohorte, une version adaptée de ce questionnaire a été utilisée sur un échantillon de femmes de taille restreinte afin d'explorer le lien entre les connaissances, les habitudes d'exposition et d'utilisation de produits de protection solaire.

2. Matériel et Méthodes

Trois cent vingt femmes caucasiennes d'Ile de France âgées de 19 à 73 ans ayant déclaré être utilisatrices de produits de protection solaire ont été incluses dans cette étude.

Une analyse descriptive de l'échantillon a tout d'abord été réalisée. Pour caractériser l'exposition solaire des femmes, une variable de synthèse définissant les habitudes d'exposition solaire a été recherchée. Une typologie de variables a d'abord été effectuée pour obtenir des groupes de variables homogènes. Une analyse en composantes principales (ACP) a ensuite été appliquée sur chaque groupe de variables, afin d'obtenir une variable de synthèse appelée score construite à partir de la première composante principale. Le coefficient alpha de Cronbach a été calculé afin de mesurer la cohérence interne du score, un coefficient supérieur à 0,7 indiquant une cohérence satisfaisante du score. Puis les liens entre les scores construits et les autres variables du questionnaire ont été étudiés. Dans le but de mieux comprendre le comportement des femmes face au soleil, une classification des femmes en fonction de leurs habitudes d'exposition a également été réalisée en procédant en deux étapes. Une analyse des correspondances multiples (ACM) sur les variables relatives aux habitudes d'exposition au soleil a tout d'abord été réalisée dans le but de construire des variables de synthèse résumant au mieux l'information. Les liens entre les variables décrivant les habitudes d'exposition solaire ont également été étudiés [JOB 92]. Une classification ascendante hiérarchique (méthode de Ward) des femmes a été réalisée à partir des composantes principales retenues à l'étape précédente [EVE 93]. La représentation du dendogramme a été réalisée. Afin de décrire la typologie obtenue, des tests de comparaison de moyennes et de pourcentages ont été effectués. Afin d'apprécier visuellement le lien entre les connaissances des femmes sur les dangers du soleil et les habitudes d'exposition solaire, une ACM a été réalisée avec en variables actives les variables relatives aux connaissances des dangers du soleil et en variables supplémentaires les classes d'habitude d'exposition au soleil obtenues à l'étape précédente. La même analyse a été réalisée en prenant les variables relatives à la protection solaire comme variables actives.

3. Résultats

3.1. *Analyse descriptive*

Vingt trois pour cent des femmes de l'échantillon connaissent la définition exacte d'un coup de soleil, 58% des femmes sont conscientes d'un lien entre l'exposition au soleil et la survenue de cancers cutanés, et 29% des femmes évoquent l'existence d'un lien entre l'exposition au soleil et le vieillissement prématuré de la peau. Bien que se déclarant utilisatrices de produits de protection solaire, uniquement 8% de ces femmes connaissent la signification des initiales SPF (Sun Protection Factor).

3.2. *Caractérisation de l'exposition au soleil*

Un score quantifiant l'intensité d'exposition solaire a été obtenu. Plus la personne s'expose au soleil, plus le score est élevé. Le score a une distribution avec un pic autour de la valeur 3 et un autre autour de 6. Sur l'échantillon, le score minimum obtenu est de 0,55 et le score maximum est de 10.

3.3. *Typologie selon les habitudes d'exposition solaire*

3.3.1. *Recherche de variables de synthèse décrivant les habitudes d'exposition*

Les deux premiers axes factoriels restituent 39% de l'information.

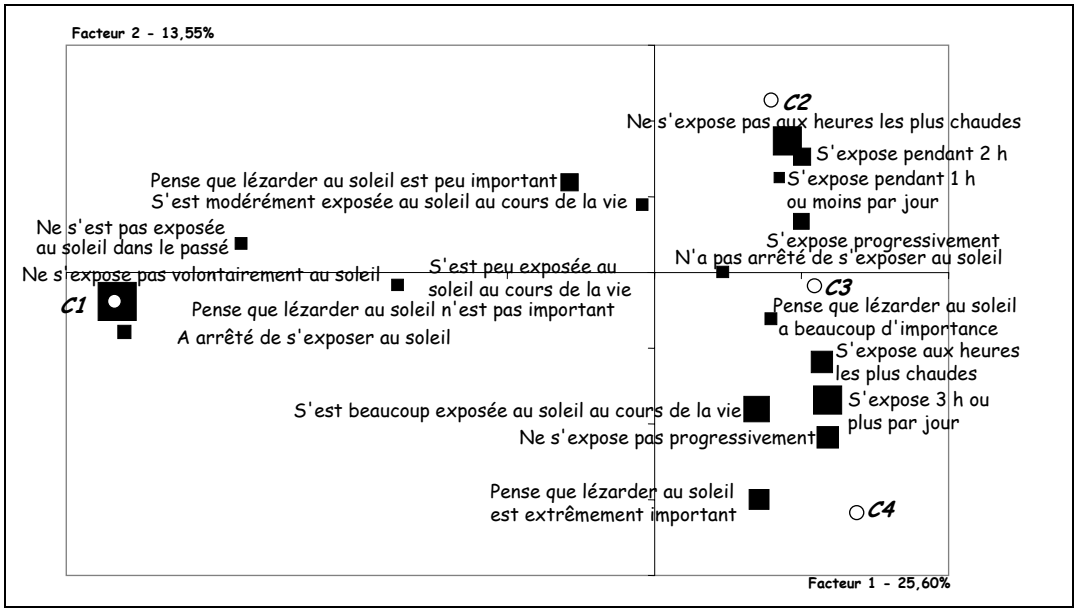


Figure 1. *Premier plan factoriel de l'analyse des correspondances multiples sur les habitudes d'exposition solaire.* ■ *Variables décrivant les habitudes d'exposition solaire,* ○ *Typologie de comportement d'exposition au soleil, en variable illustrative*

La figure 1 montre que la première composante oppose les femmes qui ont déclaré ne pas s'exposer à celles qui ont déclaré s'exposer. Le second axe oppose les femmes qui s'exposent de façon modérée à celles qui s'exposent de façon intense.

3.3.2. *Recherche d'une typologie de comportement d'exposition au soleil*

La typologie obtenue a permis d'identifier quatre classes : les femmes qui ne s'exposent pas au soleil (C1, N = 69), les femmes qui s'exposent modérément et prudemment (C2, N = 95), les femmes qui s'exposent modérément mais pas très prudemment (C3, N = 96) et les femmes qui s'exposent de façon intense et imprudente (C4, N = 54). Les femmes qui ne s'exposent pas (C1) déclarent plus fréquemment avoir entendu parler de mélanome et utiliser des produits de protection solaire avec un fort indice de protection (indice >30) ; de plus elles ne réduisent pas l'indice de protection durant la période d'exposition. A l'opposé, les femmes qui s'exposent de façon intense et imprudente (C4) déclarent utiliser des produits de protection solaire avec un faible indice de protection (indice <10) et n'avoir jamais entendu parler de mélanome. Elles sont en moyenne plus jeunes que les autres. Les groupes C2 et C3 présentent des comportements intermédiaires.

3.4. *Liens avec les connaissances des dangers du soleil et la protection solaire*

Les femmes des classes C1 et C2 ont la meilleure connaissance de la gravité des risques liés à l'exposition de la peau au soleil. Il existe un lien entre la typologie de comportement d'exposition solaire et l'utilisation de produits de protection solaire : plus l'indice de protection solaire utilisé est faible, plus l'exposition au soleil est intense (C3 et C4).

4. Conclusion

L'analyse décrite ci-dessus a été réalisée sur un échantillon restreint de femmes d'Ile de France ayant déclaré être utilisatrices de produits de protection solaire. Cette analyse sera reconduite sur les données du questionnaire actuellement collecté auprès des 13017 volontaires de la cohorte nationale SU.VI.MAX, indépendamment pour les femmes et pour les hommes. Une première étude sur les comportements d'exposition au soleil et les habitudes de protection face au soleil, mais n'investiguant pas les connaissances des dangers du soleil, a précédemment été menée au sein de cette cohorte. Elle a permis de constater des différences de comportement entre les hommes et les femmes, entre les phototypes, ainsi qu'entre les sujets les plus jeunes et les moins jeunes concernant l'intensité d'exposition et l'utilisation de moyen de protection face au soleil [GUI 01]. La littérature reporte des différences de comportements d'exposition solaire et de connaissances des dangers significatives par sexe en Australie : la connaissance des dangers du soleil est plus importante chez les femmes qui ont tendance à utiliser plus régulièrement des produits de protection solaire [ART 95]. Grâce aux données issues de la cohorte française SU.VI.MAX, nous serons alors en mesure de comparer les comportements d'exposition solaire en tenant compte des connaissances des dangers du soleil de femmes et d'hommes adultes, et d'étudier l'impact sur l'utilisation de produits de protection solaire.

Les résultats de cette étude pilote nous confortent dans l'idée que notre questionnaire est approprié pour classer des individus en fonction de leur connaissance des risques et de leurs habitudes d'exposition et de protection face au soleil, dans le but de pouvoir identifier et cibler des groupes d'individus pour des campagnes d'information de santé publique et/ou pour des études d'intervention.

5. Bibliographie

[ART 95] ARTHEY S., CLARKE VA., « Suntanning and sun protection : a review of the psychological litterature ». *Soc Sci Med* 1995, 40:265-274.

[EVE 93] EVERITT BS. Eds. *Cluster analysis*. London : Arnold, 1993.

[GUI 01] GUINOT C., MALVY D., LATREILLE J., PREZIOSI P., GALAN P., VAILLANT L., TENENHAUS M., HERCBERG S., TSCHACHLER E., « Sun exposure behaviour of a general adult population in France ». Dans : *Skin and Environment – Perception and Protection* (J. Ring, S. Weidinger, U. Darsow, éditeurs), 10e congrès de l'EADV, Munich, 10-14 octobre 2001, Bologne, Monduzzi editore S.p.A., 2001, p.1099-1106.

[HER 98] HERCBERG S., PREZIOSI P., BRIANÇON S., GALAN P., TRIOL I., MALVY D., ROUSSEL AM, FAVIER A., « A primary prevention trial using nutritional doses of antioxidant vitamins and minerals in cardio-vascular diseases and cancers in a general population: « The SU.VI.MAX study » - Design, methods and participants characteristics ». *Control Clin trials* 1998, 19:336-351.

[JOB 92] JOBSON JD. *Applied Multivariate Data Analysis.* Volume II : Categorical and Multivariate Methods. New York : Springer Verlag, 1992.

Arbre de décision pour des variables de type taxonomique

Chérif MBALLO[1,2] — Mounir ASSERAF[1] — Edwin DIDAY [2]

1 *: ESIEA Recherche*
38, Rue des Docteurs Calmette et Guérin
53000 Laval- France.
{ asseraf,mballo}@esiea-ouest.fr

2 *: LISE-CEREMADE, Université Paris IX Dauphine*
Place du Maréchal de Lattre de Tassigny
75775 Paris 16ième
diday@ceremade.dauphine.fr

RÉSUMÉ. *Plus on avance dans la technologie, plus les bases de données prennent de l'importance au niveau volume. L'approche des objets symboliques permet de réduire la taille des données à traiter en transformant les variables initiales classiquement utilisées en variables dites symboliques. La structure de ces variables au point de vu algébrique nous impose d'adapter des métriques pour pouvoir les étudier. Notre contribution dans cet article consiste à adapter le critère de Kolmogorov-Smirnov à ce type de variables mais seulement de type taxonomique.*

MOTS-CLÉS: *Analyse de données symboliques, taxonomies, arbre de décision, parcours d'arbres, critère de Kolmogorov-Smirnov.*

1. Introduction

Dans le domaine de la discrimination par arbre de décision binaire (Breiman et al., 1984), les variables explicatives sont de types quantitatives ou qualitatives. Nous proposons dans cet article une adaptation du critère de Kolmogorov-Smirnov aux variables symboliques (Diday, 1998 ; Bock et al., 2000) de type taxonomique du fait que nous pouvons ordonner ce type de variable et que la relation d'ordre choisie est totale. Cette adaptation du critère de Kolmogorov-Smirnov s'applique aussi dans le cas d'un tableau de (p+1) variables symboliques taxonomiques : p variables explicatives et une variable à expliquer (variable classe). Mais notre exemple illustratif portera seulement sur une seule variable explicative.

2. Ordre et variable taxonomique

Une variable taxonomique (Diday, 1998) est une application de l'ensemble des individus dans un ensemble de valeurs ordonnées (totalement ou partiellement). Le domaine d'observation a une structure hiérarchique. Une variable taxonomique est aussi appelée variable hiérarchique ou structurée.

Exemples : La couleur est considérée comme étant « claire » si elle est « blanche » ou « jaune ».

La racine, unique nœud n'ayant pas de père, correspondra ici au nom de la variable taxonomique considérée et donc ne sera prise par aucun individu de la population. La description d'un individu sera un nœud de la taxonomie. Le problème consiste à ordonner les individus à travers leur description afin d'adapter le critère de Kolmogorov-Smirnov. On parlera de taxonomie binaire si chaque nœud a au plus deux fils. Dans le cas où un nœud peut avoir plus de deux fils, on parlera de taxonomie n-aire.

Pour étudier ce problème, nous nous basons sur le principe de parcours d'arbres (parcours en profondeur infixe et parcours en largeur) et d'arbre binaire de recherche (Gondran et al, 1985). Un parcours d'arbre est une façon d'ordonner les nœuds d'un arbre afin de les parcourir. Dans notre cas, il n' y aura pas de traitement de la racine car, comme indiqué ci-dessus, elle représente la variable et donc n'est pas une valeur (description) à traiter. Deux nœuds quelconques admettent toujours un nœud père commun (au moins la racine). Soient x et y deux nœuds de la variable taxonomique. Désignons par $x\Re y$ pour dire que le nœud x « ***est avant*** » le nœud y. Cette relation sera définie différemment selon que l'on ordonne en profondeur (parcours en profondeur infixe) ou en largeur (ou parcours par niveau).

2.1. *Ordonner en « profondeur »*

Nous considérons tout d'abord le cas où la taxonomie est binaire. Le parcours de la variable taxonomique, dans le but de trouver un ordre croissant des individus à travers leur description, consistera alors à parcourir récursivement les sous arbres gauche et droit de la racine (variable) de la façon suivante :

– traiter d'abord le sous arbre (ou nœud) gauche ;

– traiter la racine de ce sous arbre (ou père de ce nœud) gauche;

– et enfin traiter le sous arbre (ou nœud) droit de la même façon.

Ce traitement se fera en respectant la propriété suivante connue des arbres binaires de recherche : tous les nœuds du sous arbre gauche d'un nœud ont un numéro d'ordre inférieur au sien et tous les nœuds du sous arbre droit ont un numéro d'ordre supérieur au sien. Le principe de numérotation des nœuds est le suivant : on part de la racine (variable) et on continue à suivre la branche gauche jusqu'au dernier nœud n'ayant pas de fils. On le numérote et c'est ainsi le premier nœud numéroté. On numérote le nœud père et on passe au fils droit s'il existe en appliquant le même principe. Ensuite on remonte au nœud père du nœud père précédent et on applique encore le même principe. L'ordre ainsi obtenu est un ordre croissant. Ce principe nous permet de définir la relation $\Re$ de la façon suivante :

x$\Re$y ⇔ [x provient de la branche gauche du nœud père commun avec y]

ou **[y provient de la branche droite du nœud père commun avec x]**

Dans le cas d'une taxonomie n-aire, le principe consiste à choisir, parmi les m fils d'un nœud (m >2), les p premiers à partir de la gauche comme groupe de fils gauche (les (m-p) restants forment alors le groupe de fils droit) et d'appliquer le même déroulement que précédemment. Les numéros d'ordre d'un groupe sont consécutifs (si un élément du groupe n'a pas d'autres fils).

On montre sans difficulté majeure que $\Re$ est une relation d'ordre total (anti-réflexive et transitive).

2.2. ***Ordonner « niveau par niveau » ou en « largeur »***

Nous définissons ici des niveaux de la taxonomie à partir du bas. Le premier niveau (niveau 1) de la taxonomie correspond au(x) plus bas nœud(s) n'ayant pas de fils. Ensuite on passe au niveau suivant et ainsi de suite jusqu'à la racine. Mais le niveau correspondant à la racine n'est pas numéroté du fait que la racine n'intervient pas dans le traitement. Avec ce principe, nous définissons la relation $\Re$ de la façon suivante :

– Si les deux nœuds x et y sont au même niveau, alors :

x$\Re$y ⇔ x provient de la branche gauche du nœud père commun avec y

– Si les deux nœuds x et y ne sont pas au même niveau, alors : **x$\Re$y ⇔ niveau(x) < niveau(y).**

Ici aussi, on montre sans difficulté majeure que $\Re$ est une relation d'ordre total.

L'utilisateur est libre de choisir l'ordre le plus convenable selon ses données.

3. Arbre de décision

Pour construire un arbre de décision, nous avons besoin d'un critère d'évaluation de la coupure d'un nœud en deux nœuds fils. Le critère de Kolmogorov-Smirnov (*KS)* permet de séparer une population en deux sous populations plus homogènes en se basant uniquement sur les deux fonctions de répartition induites par le regroupement des classes a priori en deux super-classes. La méthode nommée « twing » (Breiman et al., 1984) est utilisée pour générer, à partir de m classes, deux super-classes C_1, C_2 auxquelles sont associées les deux fonctions de répartition F_1, F_2 d'une variable aléatoire symbolique *X* de type taxonomique. Cette méthode examine tous les cas possibles pour regrouper m classes en deux classes appelées super-classes. La fonction de répartition empirique $\hat{F}_i$ (*i*=1,2) qui estime F_i est donnée par :

$$\hat{F}_i(x) = \frac{card\{(X \le x) \cap C_i\}}{cardC_i} \quad .$$

Ainsi la distance de Kolmogorov-Smirnov *KS* sera :

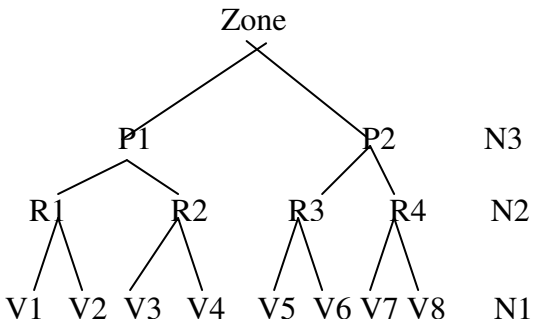
$$KS = \sup \left| \hat{F}_1(x) - \hat{F}_2(x) \right|$$

C'est une extension naturelle de la distance *KS,* seulement l'argument sélectionné pour le seuil est un nœud de la variable taxonomique considérée et non un réel comme dans le cas classique. On peut donc utiliser toutes les autres étapes pour construire l'arbre de décision qui sont communes à tout type de variable.

Nous proposons dans la suite un exemple d'application pour illustrer le principe de la construction d'un arbre de décision pour une variable ***Zone*** de type taxonomique. Le tableau de données ci-dessous (tableau 1) indique

une première colonne représentant les dix individus, une deuxième colonne donnant la description de chaque individu au départ et une troisième colonne répartissant les individus en deux classes a priori.

Individu	Description (Zone)	Classe a priori
Ind1	V1	1
Ind2	V2	2
Ind3	V3	1
Ind4	V3	1
Ind5	R1	2
Ind6	P1	2
Ind7	V5	1
Ind8	R3	2
Ind9	V7	1
Ind10	V8	2



P : Pays ; R : Région ; V : Ville ; N : Niveau

En ordonnant les individus à travers leur description par les deux méthodes exposées au § 2, on obtient le tableau 2 ci-dessous (la description correspondante de l'individu est entre parenthèses).

Tableau 1

Désignons par Z cette variable taxonomique explicative. Considérons par exemple l'ordre donné par la « *profondeur* » et calculons les valeurs correspondantes du *KS* pour chaque description. Pour la construction de l'arbre de décision binaire, un nœud interne est représenté par une ellipse contenant à gauche l'effectif de la classe C1 (classe 1) et à droite celui de la classe C2 (classe 2). Un nœud n'ayant pas de fils (feuille) est représenté par un rectangle contenant l'effectif de la classe (la classe origine est précisée entre parenthèses).

Ordre en profondeur	Classe	Ordre par niveau	Classe
Ind1 (V1)	1	Ind1 (V1)	1
Ind5 (R1)	2	Ind2 (V2)	2
Ind2 (V2)	2	Ind3 (V3)	1
Ind6 (P1)	2	Ind4 (V3)	1
Ind3 (V3)	1	Ind7 (V5)	1
Ind4 (V3)	1	Ind9 (V7)	1
Ind7 (V5)	1	Ind10 (V8)	2
Ind8 (R3)	2	Ind5 (R1)	2
Ind9 (V7)	1	Ind8 (R3)	2
Ind10 (V8)	2	Ind6 (P1)	2

Ordre en *Profondeur* :

KS(V1)=0.2 ; KS(R1)=0 ; KS(V2)=0.2 ; KS(P1)=0.4 ; KS(V3)=0 ; KS(V5)=0.2 ; KS(R3)=0 ; KS(V7)=0.2 ; KS(V8)=0 ;

Donc le seuil de coupure pour séparer les deux classes est le nœud x=P1 correspondant à l'individu Ind6.

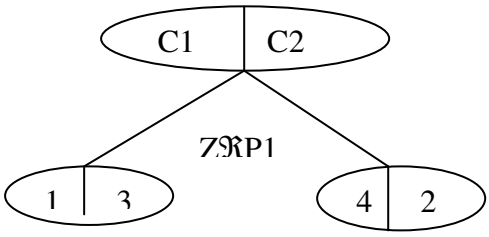
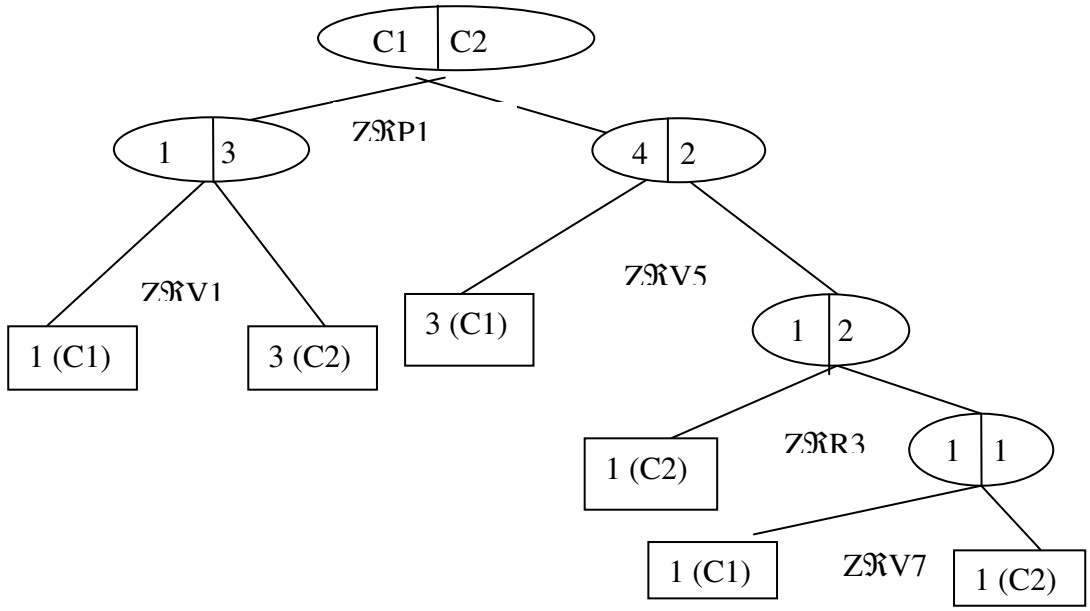


Tableau 2

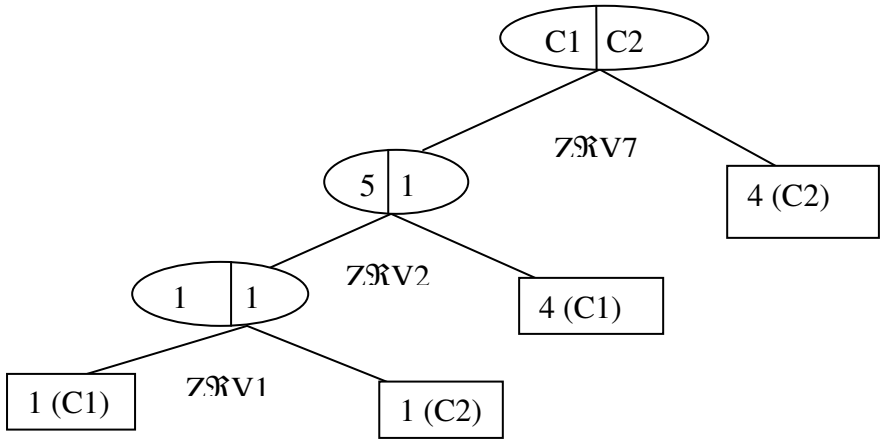
Pour chacun des deux nœuds ainsi formés, on calcule encore les valeurs des KS de chacune des descriptions.

Pour le nœud gauche, on a : KS(V1)=1, valeur maximale de ce critère donc x=V1 est le seuil de coupure de ce nœud et on obtient ainsi deux feuilles : un individu de C1 à gauche et trois individus de C2 à droite.

Pour le nœud droit, on a : KS(V3)=0.5 ; KS(V5)=0.75 ; KS(R3)=0.25 ; KS(V7)=0.5 et enfin KS(V8)=0 ; donc le seuil de coupure est x=V5 et le fils gauche de cette coupure est une feuille de trois individus de C1 (voir figure ci-dessous). Il reste alors à couper le fils droit contenant un individu de C1 (V7) et deux de C2 (R3 et V8). Le calcul du KS donne : KS(R3)=0.5 ; KS(V7)=0.5 et KS(V8)=0. Le seuil de coupure est x=R3. Nous obtenons une feuille à gauche (R3) et un nœud de deux individus (V7 et V8). Le calcul du KS donne : KS(V7)=1, donc le seuil de coupure est x=V7 et nous obtenons deux feuilles contenant chacun un seul individu (figure ci-dessous).



En considérant maintenant l'ordre donné par la largeur et en appliquant le même principe, on obtient l'arbre de décision binaire suivant :



Nous remarquons avec ce petit exemple que l'arbre de décision obtenu par l'ordre en « *largeur* » est moins « développé » que celui donné par l'ordre en « *profondeur* ». Mais ce constat sur ce petit exemple se vérifie t-il sur de gigantesques bases de données ?

4. Bibliographie

Asseraf, M. ; Mballo, C. (2003), Arbre de décision pour des variables de type intervalle. *XXXV^{ième} Journées de Statistique, du 02 au 06 Juin 2003,* Lyon, France.

Bock H. H.& Diday E. (2000), *Analysis of symbolic data : Exploratory methods for extracting statistical information from complex data*, Springer-Verlag, Berlin-Heidelberg.

Breiman, L.; Friedman, J. H.; Ohlsen R. A.; Stone C. J. (1984), *Classification and regression trees*, The Wadsworth Statistics/Probability Series, Belmont, CA.

Diday, E. (1998), L'analyse des données symboliques : un cadre théorique et des outils, Cahier de recherche du CEREMADE, UMR 7534, Numéro 9821, Université Paris Dauphine.

Gondran, M. ; Minoux , M. (1985), *Graphes et algorithmes*, 2^{e} édition, Eyrolles, Paris.

Classification arborescente de données par auto-assemblage de fourmis artificielles

H. Azzag[*], **N. Monmarché**[*],**C. Guinot**[**], **M. Slimane**[*], **G. Venturini**[*]

[*]*École Polytechnique de l'Université de tours, Laboratoire d'Informatique,*
64, Avenue Jean Portalis, 37200 Tours, France
hanene.azzag@etu.univ-tours.fr, monmarché,slimane,venturini@univ-tours.fr

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

RÉSUMÉ. *Nous présentons dans cet article un nouveau modèle pour la classification hiérarchique inspiré du comportement de fourmis réelles. Il s'agit de simuler la manière dont les fourmis construisent des assemblages complexes en s'accrochant les unes aux autres. Motivés à moyen termes par la construction de sites portails pour le Web, nous montrons comment ce comportement peut servir à construire un arbre regroupant des données en fonction de leurs similarités. Nous détaillons plusieurs algorithmes utilisant ce modèle et nous les évaluons sur une vingtaine de bases de données. Nous comparons les résultats obtenus avec les k-means et la classification ascendante hiérarchique.*

MOTS-CLÉS : *Classification hiérarchique, arbres, fourmis artificielles, sites portails*

1. Introduction

La motivation initiale et le but à moyen terme de cette recherche est la construction automatique de sites portails pour le Web. Un site portail peut être vu comme une structure classificatoire hiérarchique d'un ensemble de documents qui diffère cependant d'un dendogramme (des données sont présentes dans les noeuds internes). Un des problèmes majeurs posé dans ce cadre est celui de la définition automatique de la hiérarchie de documents qui dans les systèmes actuels est donnée à l'avance par un expert humain [KUM 01]. Si l'on travaille avec un grand nombre de documents, ou si l'on souhaite que la machine puisse de manière autonome construire un tel site, les approches existantes sont inopérantes.

Nous proposons une nouvelle approche construisant une classification de données sous la forme d'un arbre. Ce nouveau modèle se place dans la lignée de nos travaux précédents sur les algorithmes de classification s'inspirant du comportement des fourmis réelles et plus généralement des systèmes biologiques. Ces algorithmes peuvent bénéficier de propriétés intéressantes comme l'optimisation locale et globale de la classification, l'absence d'information sur une classification initiale des données, le parallélisme, etc.

La section 2 décrit d'une part le modèle biologique portant sur la manière dont les fourmis construisent des structures vivantes en s'accrochant les unes aux autres. et d'autre part comment ce comportement peut être utilisé pour construire des arbres classifiant les données. La section 3 présente les différents choix algorithmiques possibles où nous évaluons chacune des options sur des bases de données classiques et réelles. Nous présentons un résumé des résultats expérimentaux obtenus avec notamment une comparaison avec les k-means et CAH. La section 4 conclut sur les perspectives qui découlent de ce travail.

2. Modèles biologique et informatique

Les capacités des fourmis réelles ont inspirées depuis plus d'une décennie les chercheurs dans la conception de nouveau algorithmes pour la classification. Le modèle le plus étudié initialement est celui modélisant la manière dont les fourmis trient le couvain [GOS 91][LUM 94][KUN 99][MON 99].

Nous traitons dans cet article d'un nouveau comportement que l'on observe chez plusieurs espèces de fourmis que nous décrivons brièvement ici [LIO 01][THE 01] : il s'agit de la manière dont ces insectes s'accrochent les uns aux autres pour construire des structures vivantes ayant différentes fonctionnalités. Les fourmis peuvent ainsi construire des "chaînes de fourmis" leur permettant de passer d'un point à un autre ou de rapprocher des bords d'une feuille pour y placer leur nid, ou encore des "gouttes de fourmis" qui semble être une fonctionalitée encore inexpliqué. Les principes de ce comportement sont les suivants : les fourmis partent d'un point initial (le support) sur lequel elles s'accrochent. Lorsqu'une fourmi s'est accrochée, elle fait partie de la structure et les autres fourmis peuvent alors se déplacer sur elle. La structure croît au cours du temps selon les actions locales des fourmis qui se déplacent pour choisir un emplacement où s'accrocher. Elles sont influencées par la forme locale de la structure et également par le point à atteindre. Les fourmis accrochées au milieu de la structure ne peuvent se dégager à moins d'entraîner un éventuel décrochage. On observe donc aussi un phénomène de décroissance de la structure.

A partir de ces éléments, nous définissons les grandes lignes de notre modèle informatique simulant ce comportement. Les fourmis $f_1, ..., f_n$ représentent chacune une des données de la base et sont placées initialement sur le support f_0. Ensuite, nous simulons successivement une action pour chaque fourmi. Une fourmi peut avoir deux états : elle est soit libre de se déplacer ou de se connecter, soit assemblée à la structure sans la possibilité de se déplacer mais seulement de se décrocher. Les fourmis ne perçoivent la structure que localement. Pour une fourmi f_i en déplacement et positionnée sur une fourmi f_{pos} assemblée à la structure, le voisinage V_{pos} perceptible par f_i est limité à f_{pos}, à la fourmi mère de f_{pos} (du niveau précédent dans l'arbre), aux fourmis filles de f_{pos}. La fourmi f_i peut donc percevoir les valeurs de similarité entre la donnée qu'elle représente et les données représentées par les fourmis de V_{pos}. En fonction de ces valeurs de similarité, elle peut soit se connecter à f_{pos}, soit se déplacer sur une des fourmis de V_{pos}. Pour les fourmis accrochées à la structure, il existe une possibilité de décrochage. Nous limitons le champ de cette première étude à des algorithmes ne faisant pas de décrochage de manière à bien comprendre la croissance de l'arbre (voir la section 4). Ainsi, une fois qu'il n'y a plus de fourmis en déplacement et que toutes les fourmis se sont accrochées les unes aux autres (ou sur le support), l'algorithme s'arrête. L'arbre résultant représente une classification des données. Les propriétés visées pour une classification des données représentant un site portail sont les suivantes : chaque sous-arbre A représente une catégorie composée de toutes les fourmis de A. Soit f la fourmi qui est à la racine d'un sous-arbre A. Nous souhaitons que 1) f soit représentative de cette catégorie (les fourmis placées dans A sont le plus similaires possible à f), 2) les fourmis filles de f qui représentent des sous-catégories soient les plus dissimilaires possible entre elles.

3. Algorithmes et tests

Il faut maintenant trouver des règles de comportement des fourmis artificielles permettant d'obtenir les propriétés mentionnées précédemment. Nous avons étudié plusieurs algorithmes :

- AntTree$_{DISSIM}$: chaque fourmi f_i descend dans l'arbre en suivant le chemin de similarité maximum avec les données déjà accrochées. Si sur ce chemin f_i est à un niveau donné suffisamment dissimilaire aux fourmis qu'elles rencontrent, elle se connecte pour créer une nouvelle classe. Si f_i atteint une feuille, elle se connecte à celle-ci. Le seuil de dissimilarité nécessaire à la création d'une sous-catégorie est calculé globalement selon différentes méthodes testées. Il peut être égal à la moyenne des similarités, à la moyenne "inférieure" des similarités, etc.
- AntTree$_{SIM-DISSIM}$: il s'agit du même algorithme mais utilisant en plus un seuil de similarité. f_i ne suit le chemin de similarité maximale tant qu'elle est suffisamment similaire. Sinon, elle est replacée sur le support. La terminaison de cet algorithme peut être assurée par différentes méthodes.
- AntTree$_{SIM-DISSIM-GLOBAL}$: les seuils de similarité et de dissimilarité sont adaptés au cours du temps aux données traitées (car il existe des caractéristiques très différentes dans les données testées).

Bases	Classes	AntTree$_{SIM-DISSIM-LOCAL}$			CAH		
	réelles	C [σ_c]	P [σ_P]	Ec [σ_{Ec}]	C	P	Ec
IRIS	3	3,0 [0,00]	86,40 [0,33]	0,15 [0,00]	2,00	66,67	0,22
WINE	3	9,0 [0,00]	94,44 [0,17]	0,18 [0,00]	3,00	94,94	0,07
GLASS	7	14,6 [1,20]	57,10 [3,73]	0,34 [0,02]	2,00	45,79	0,40
PIMA	2	18,3 [0,64]	70,91 [0,98]	0,51 [0,00]	2,00	65,10	0,48
SOYBEAN	4	4,0 [0,00]	83,19 [6,83]	0,14 [0,05]	4,00	100,00	0,00
THYROID	3	7,0 [0,00]	89,30 [0,00]	0,18 [0,00]	3,00	90,23	0,16
CERIES	6	8,9 [0,83]	65,33 [7,79]	0,19 [0,03]	2,00	46,72	0,33
ZOO	7	5,0 [0,00]	88,51 [0,49]	0,04 [0,01]	4,00	82,18	0,05
HAYES-ROTH	3	4,9 [0,30]	57,58 [2,27]	0,37 [0,00]	2,00	38,64	0,52
LYMPHOGRAPHY	4	7,0 [0,00]	72,50 [0,43]	0,39 [0,00]	2,00	56,08	0,49
HEART	2	8,2 [0,75]	76,19 [2,08]	0,44 [0,01]	2,00	75,56	0,37
SEGMENT	7	10,3 [0,9]	67,02 [2,96]	0,13 [0,01]	2,00	28,57	0,39
VEHICLE	4	5,0 [0,00]	39,29 [1,20]	0,38 [0,00]	2,00	36,76	0,49
TIC-TAC-TOE	2	6,0 [0,00]	65,39 [0,08]	0,52 [0,00]	2,00	65,34	0,50
HOUSE-VOTES-84	2	5,0 [0,00]	91,72 [0,00]	0,23 [0,00]	2,00	81,61	0,30
WAVEFORM1000	3	5,8 [0,75]	58,58 [1,78]	0,33 [0,00]	3,00	68,80	0,29

TAB. 1. *Extraits des résultats obtenus. C est le nombre de classe, P la pureté et Ec une mesure d'erreur de classification fondées sur les couples de données.*

– AntTree$_{SIM-DISSIM-LOCAL}$: les seuils s'adaptent comme précédemment mais sont localisés pour chaque fourmi qui ajuste ainsi ses seuils en fonction du résultats de ses actions et de la donnée qu'elle représente. L'adaptation n'est plus globale pour toutes les données mais locale à chaque donnée.

Par ailleurs, puisque les premières fourmis de la base de données ont tendance à se connecter en premier dans l'arbre, nous étudions différentes manières de trier les données : aléatoirement, selon l'ordre croissant (respectivement décroissant) des valeurs de similarité moyenne.

Nous avons évalué ces algorithmes sur un ensemble de 24 bases de données, représentant des données numériques et symboliques allant de 47 à 2310 données et sur les données issues du CE.R.I.E.S. [GUI 03]. Les classifications obtenus sont évaluées à la fois en terme de nombres de classes, de pureté des classes, et d'erreur sur les couples de données. Les meilleurs résultats sont obtenus pour un tri croissant des données initiales et des seuils calculés localement par chaque fourmi.

Les temps de calcul pour AntTree$_{SIM-DISSIM-LOCAL}$ sont de l'ordre de quelques secondes à un minute sur un PC standard. L'arbre résultant est à la fois engendré sous la forme de pages HTML dans lesquelles l'utilisateur peut naviguer et ouvrir les documents (ou données) correspondantes. Nous utilisons également un affichage hyperbolique et dynamique de l'arbre de manière à laisser l'utilisateur se focaliser interactivement sur une partie précise sans perdre le contexte de l'ensemble de l'arbre.

Pour les bases de données numériques, nous comparons les résultats obtenus avec les k-means (initialisé à 10 classes). Dans la majorité des bases, nos algorithmes sont nettement meilleurs que les k-means. Pour toutes les bases, nous comparons les résultats avec la classification ascendante hiérarchique. Les résultats obtenus par notre approche sont de qualité similaire à ceux de CAH sur 10 bases, inférieurs sur 4 bases et supérieurs sur 10. Les résultats obtenus sont donc très encourageants (voir extrait des résultats dans la table 1) compte tenu également du fait que nos algorithmes sont environ 20 fois plus rapides que CAH.

4. Conclusion

Nous avons présenté dans cet article un nouveau modèle inspiré de la biologie pour la classification de données sous une forme arborescente. Il donne des résultats très positifs notamment en comparaison avec des méthodes classiques. Dans les travaux en cours, nous étudions comment implémenter le décrochage des fourmis. Des sous-parties de l'arbre peuvent ainsi se détruire et les fourmis correspondantes se retrouvent sur le support. Ce décrochage est très important du point de vue de l'optimisation car c'est une manière de revenir sur des décisions précédemment prises par l'algorithme. Il s'agit aussi d'éviter de trier initialement les données puisque le décrochage doit permettre d'effectuer seul des choix judicieux pour les données de départ. Nous sommes actuellement en train d'appliquer cet algorithme à un problème réel de construction automatique de site portail.

5. Bibliographie

[GOS 91] GOSS S., DENEUBOURG J.-L., Harvesting by a group of robots, VARELA, Ed., *Proceedings of the First European Conference on Artificial Life*, Sydney, Australia, 1991, Toward a Practice of Autonomous Systems, p. 195–204.

[GUI 03] GUINOT C., MALVY D. J.-M., MORIZOT F., TENENHAUS M., LATREILLE J., LOPEZ S., TSCHACHLER E., DUBERTRET L., Classification of healthy human facial skin, Textbook of Cosmetic Dermatology Third edition (to appear), 2003.

[KUM 01] KUMAR R., RAGHAVAN P., RAJAGOPALAN S., TOMKINS A., On semi-automated Web taxonomy construction, *WebDB*, Santa Barbara, May 2001.

[KUN 99] KUNTZ P., SNYERS D., LAYZELL P., A stochastic heuristic for visualising graph clusters in a bi-dimensional space prior to partitioning, *Journal of Heuritics*, vol. 5, n^o 3, 1999.

[LIO 01] LIONI A., SAUWENS C., THERAULAZ G., DENEUBOURG J.-L., The dynamics of chain formation in Oecophylla longinoda, *Journal of Insect Behavior*, vol. 14, 2001, p. 679-696.

[LUM 94] LUMER E., FAIETA B., Diversity and Adaptation in Populations of Clustering Ants, *Proceedings of the Third International Conference on Simulation of Adaptive Behaviour*, 1994, p. 501–508.

[MON 99] MONMARCHÉ N., On data clustering with artificial ants, FREITAS A., Ed., *AAAI-99 & GECCO-99 Workshop on Data Mining with Evolutionary Algorithms : Research Directions*, Orlando, Florida, July 18 1999, p. 23-26.

[THE 01] THERAULAZ G., BONABEAU E., SAUWENS C., DENEUBOURG J.-L., LIONI A., LIBERT F., PASSERA L., SOLÉ R.-V., Model of droplet formation and dynamics in the Argentine ant (Linepithema humile Mayr), *Bulletin of Mathematical Biology*, vol. 63, 2001, p. 1079-1093.

Classification hiérarchique ascendante avec imputation multiple de données manquantes*

Ana Lorga da Silva
Laboratório de Estatística e Análise de Dados
Faculdade de Psicologia e Ciências da Educação, Universidade de Lisboa
Alameda da Universidade
1694-013 Lisboa Portugal
aigcls@iseg.utl.pt

Gilbert Saporta
Chaire de Statistique Appliquée
Conservatoire National des Arts et Métiers
292 rue Saint Martin
75141 Paris cedex 03 France
saporta@cnam.fr

Helena Bacelar-Nicolau
Laboratório de Estatística e Análise de Dados (LEAD)
Faculdade de Psicologia e Ciências da Educação, Universidade de Lisboa
Alameda da Universidade
1694-013 Lisboa Portugal
hbacelar@fpce.ul.pt

RÉSUMÉ. En présence de données manquantes, et dans le cadre de la classification hiérarchique de variables on étudie deux méthodes utilisant les matrices obtenues après imputation multiple.
MOTS-CLÉS :Données Manquantes, Imputation Multiple, Classification Hiérarchique Ascendante

1. Introduction

Lorsque certaines valeurs sont manquantes dans un tableau de données, il est souvent nécessaire de les estimer avant d'appliquer une méthode statistique. L'imputation multiple [RUB 87] permet de restituer la variabilité des données par la substitution de *m* matrices de données complètes à une matrice comportant des données manquantes. Quand on fait ensuite

* Ce travail a été partiellement supporté par le Programme de Coopération Scientifique et Technique Luso-Française MSPLDM-542-B2 (Ambassade de France au Portugal et Ministère de la Science et de l'Enseignement Supérieur - ICCTI) co-dirigé par H. Bacelar-Nicolau e G. Saporta et par le Project d'Analyse des Données Multivariées (CEAUL-FCT) dirigé par H. Bacelar Nicolau.

une classification hiérarchique des variables, on cherche à obtenir à la fin une seule structure (consensus ou résumé) des *m* structures hiérarchiques obtenues.

2. Imputation Multiple, Matrices de similarité et Algorithmes de Classification

Soit $m > 1$ matrices de données, $X^1, X^2, ..., X^m$, obtenues par application d'une méthode d'imputation multiple à une matrice avec données manquantes. Ces matrices ont ceci de commun que les *n* individus et les *p* variables sont les mêmes, la différence concernant un pourcentage *i* des observations, celles qui correspondent aux données manquantes de la matrice initiale.

Quand on utilise une méthode de classification sur les variables ou les individus, se pose alors la question: «Comment faire la combinaison des *m* matrices de façon à conclure sur la structure des données complètes?». On a développé deux méthodes qu'on comparera et évaluera en recourant à des matrices de données simulées.

2.1.1..*Première Méthode - Combinaison des matrices de similarité*

Cette méthode est basée sur la combinaison des matrices de similarité

Après avoir obtenu les *m* matrices par une méthode d'imputation multiple on détermine:

i. pour chaque matrice, $X^1, X^2, ..., X^m$, la matrice de similarité S_k, k=1,2,..., *m*,

ii. la moyenne des matrices de similarité *S* tel que, $S = \left(\sum_{k=1}^{m} S_k \right) \Big/ m$

iii. sur *S* , on utilise la méthode d'agrégation choisie,

iv. on détermine la structure hiérarchique correspondante, qu'on considérera représentative des *m* structures hiérarchiques correspondantes (associées à chaque $X^1, X^2, ..., X^m$).

2.1.2.. *Deuxième Méthode - Combinaison des matrices des ultramétriques: «Consensus Ordinal IM»*

Cette méthode est basée sur la combinaison des structures hiérarchiques, considérant *l'ordre de l'agrégation*, (pas les niveaux d'agrégation), on la considère comme une méthode de consensus.

La procédure est la suivante:

i. pour chaque matrice, $X^1, X^2, ..., X^m$, on détermine sa matrice de similarité S_k, k=1,2,..., *m*,

ii. sur chaque matrice de similarité S_k, k=1,2,..., *m*, on utilise la méthode d'agrégation choisie, en obtenant pour chaque S_k une structure hiérarchique $H_k, k = 1, 2, ..., m$ $3 \le m \le 5$ (représenté par un dendrogramme),

iii. à chaque $H_k, k = 1, 2, ..., m$, est associée une matrice ultramétrique $U_k, k = 1, 2, ..., m$, qu'on détermine,

iv. on calcule les coefficients de Spearman r_s entre toutes les paires d'ultramétriques. On considére les cas où $r_s = 1$ qui correspondent à deux structures identiques. On cherche alors la structure ordinale majoritaire

v. le nombre de structures égales, n_i, doit être tel que

1) $n_i \in \left]\frac{m}{2}, m\right]$

2) si 1) n'est pas satisfaite on refait l'imputation on considérant m=10, avec le but de trouver un n_i, satisfaisant 1) si ce n'est pas le cas, on dira qu'il n'y a pas d'arbre représentatif (pas de consensus).

C'est une condition semblable aux méthodes de consensus comme décrites par exemple en [GOR 99]. Si on obtient une hiérarchie représentative, on pourra parler d'une règle de consensus d'arbres de classification.

On appellera cette méthode «***consensus ordinal IM***».

2.2. *Application*

Comme dans des publications antérieures ([SIL 02] et [SIL 03]), on utilise des matrices de données complètes 1000×5, issues de 5 distributions multinormales correspondant aux structures suivantes

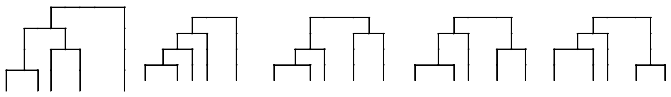


Figure 1 : Les 5 dendrogrammes

On enlève ensuite des données selon un modèle MAR - «Missing at Random».

On effectue ensuite 100 simulations de chaque cas.

On utilise comme coefficient de ressemblance le coefficient d'affinité, $c_a = \sum_{i=1}^{n} \sqrt{\frac{x_{ij}}{x_{.j}} \frac{x_{ij'}}{x_{.j'}}}$, où $x_{.j} = \sum_{i=1}^{n} x_{ij}$ et $x_{.j'} = \sum_{i=1}^{n} x_{ij'}$, défini par exemple en [BAC 85] et [BAC 02] et le coefficient de corrélation de Bravais-Pearson.

Comme méthodes d'agrégation on utilise les trois critères d'agrégation classiques: "average linkage", "single linkage" et "complete linkage".

On efface ensuite 10%, 15% et 20% des données de deux variables. Les données manquantes présentent un schéma majoritairement monotone - «*A monotone missing data pattern occurs when the variables can be ordered, from left to right, such that a variable to*

the left is at least as observed as all variables to the right» [STA 01] - avec un petit pourcentage de données manquantes représentées par un schéma non monotone.

On fait l'étude des résultats obtenus en utilisant les deux méthodes décrites, en utilisant la méthode d'imputation (m=5) pour toutes les données manquantes et en supprimant les lignes qui contiennent des données manquantes qui n'appartiennent pas au schéma monotone.

Le modèle d'imputation utilisé est basé sur la théorie Bayesienne [STA 01]. D'abord le modèle prédictif de régression OLS est estimé a partir des données complètes, comme d'habitude. On utilise ce modèle pour en générer d'autres où les valeurs des paramètres sont tirées au hasard dans la distribution *a posteriori*. *"The randomly drawn values are used to generate imputations, which include random deviations from the model's predictions"* ([STA 01]). De cette façon on garde plus de variabilité, car les paramètres sont estimés *a posteriori*.

Pour comparer les résultats aux structures originelles complètes on utilise le coefficient de Spearman entre ultramétriques.

Conclusion

Nous montrons sur nos simulations que les meilleurs résultats sont obtenus avec la première méthode associée au coefficient d'affinité avec les critères d'agrégation "average linkage" et "single linkage" et que le coefficient d'affinité est plus robuste que le coefficient de corrélation.

Bibliographie

[BAC 85] BACELAR-NICOLAU, H. "The Affinity Coefficient in Cluster Analysis", *Methods of Operation Research*, vol.53, 1985, p. 507-512, Martin J. Bekman *et al.* (ed), Verlag Anton Hain, Munchen.

[BAC 02] BACELAR-NICOLAU, H. "On the Generalised Affinity Coefficient for Complex Data", *Byocybernetics and Biomedical Engineering*, vol.22, nº 1, p. 31-42

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

[LIT 87] LITTLE, R. J. A. & RUBIN, D. B. *Statistical Analysis With Missing Data*, John Wiley & Sons, New York, 1987.

[RUB 87] RUBIN, D.B. *Multiple Imputation for Nonresponse in Surveys*, John Wiley & Sons, New York, 1987.

[SIL 02] SILVA, A.L, BACELAR-NICOLAU, H. & SAPORTA, G. "Missing Data in Hierarchical Classification of Variables - a Simulation Study" in *Classification Clustering and Data Analysis*, 2002, p.121-128, Springer.

[SIL 03] SILVA, A.L, BACELAR-NICOLAU, H. & SAPORTA, G. "Efeito de um Método de Imputação Múltipla em Classificação Hierárquica de Variáveis", *Proceedings JOCLAD 2003 X Jornadas de Classificação e Análise de Dados*, 2003 p. 130-142

[STA 01] STATISTICAL SOLUTIONS, Lda. "*SOLAS for Missing Data Analysis*, 3.0". Cork, Ireland: Statistical Solutions, 2001

Etude d'un comportement paramétré de CAHCVR sur des données réelles en imagerie numérique

K. Bachar[1] **et I-C. Lerman**[2]

(1) ESSCA - 1, Rue Lakanal – Angers – France
(2) IRISA - Campus de Rennes 1 – Rennes – France

RÉSUMÉ.
Nous analysons l'effet des paramètres de l'indice de la vraisemblance des liens dans l'algorithme CAHCVR et comparons ses résultats avec ceux obtenus utilisant l'inertie expliquée dans les applications liées à la segmentation d'images numériques réelles. Les résultats observés, comme l'accélération de l'algorithme due au choix de la stratégie de l'agrégation multiple et d'une nouvelle définition de la contiguïté, ainsi que l'absence d'inversion dans l'indice de la vraisemblance des liens, se trouvent formellement justifiés.

MOTS-CLÉS : CAH, inversion, graphe de contiguïté, agrégation multiple, images numériques.

1. Introduction

CAHCVR est la **C**lassification **A**scendante **H**iérarchique sous **C**ontrainte de contiguïté et procédant par agrégations successives des paires de classes **V**oisines **R**éciproques, au sens d'un indice de dissimilarité entre classes [BAC94]. Ce travail correspond à l'un des développements les plus récents de celui dans [LER01]. Ce développement est de nature méthodologique, mais, directement induit par les résultats expérimentaux de traitement de l'ensemble des pixels de différentes images réelles de tailles importantes et suffisamment difficiles. La notion de contiguïté qui interviendra dans nos expériences est, pour un pixel considéré, relative au voisinage de ce dernier. Cette notion s'étend naturellement aux classes connexes de pixels. Plusieurs paramètres sont à prendre en compte dans CAHCVR. Le premier est le plus important : il s'agit du choix de l'indice de dissimilarité entre classes qui permet de repérer les paires de classes voisines réciproques qu'il y a lieu de fusionner.
Deux types d'indices ont été étudiés. Le premier est celui classique de l'inertie expliquée et le second, est celui de la **V**raisemblance du **L**ien maximal (critère **VL**). Ce dernier est lui-même paramétré.
Plus précisément, un des résultats de cette recherche consiste à fixer un type de paramétrage pour VL permettant le « meilleur » rendu aux « yeux » de l'expert.

2. Agrégation multiple et graphe de contiguïté

Nous avons montré l'efficacité de la stratégie de l'agrégation multiple dans [LER01].
Il s'agit de fusionner, dans une même étape, toutes les paires de classes contiguës et voisines réciproques et réalisant la dissimilarité minimale. Cela a pour conséquence l'obtention d'un résultat unique (le résultat par agrégation binaire n'est pas toujours unique) et surtout accélère la rapidité du traitement. Aussi, dans nos différentes expériences nous avons opté systématiquement pour le principe de l'agrégation multiple. Nous avons aussi testé deux notions différentes de contiguïté propres au contexte de nos données (contexte de l'image planaire).

3. Un indice VL dans le cas d'un graphe de contiguïté

Chaque élément de l'ensemble à classifier est un sommet d'un graphe de contiguïté connexe G_0 dont une arête traduit la contiguïté entre 2 éléments.
Pour le critère VL, on construit ([LER91], [NIC96]) des indices probabilistes $P(x,y)$; ici entre chaque paire (x,y) d'éléments contigus au sens de G_0.
On associe les indices de dissimilarité informationnelle $\delta(x,y) = -\log_2(P(x,y))$.
Si $\alpha(A,B)$ représente le *nombre d'arêtes* reliant les 2 classes A et B, le *nouvel indice* de dissimilarité entre 2 classes A et B s'exprime par :

$$\Delta_\varepsilon(A,B) = \alpha^\varepsilon(A,B) \times Min\{\ \delta(x,y)/(x,y) \in A \times B, (x,y) \in G_0\ \}$$

Cas particulier : si G_0 est complet, $\alpha(A,B) = |A| \times |B|$; dans ce cas on retrouve l'indice plus classique VL.

4. Les paramètres ε et π

ε est un paramètre réel positif et compris entre 0 et 1. Deux valeurs de ε peuvent être distinguées : $\varepsilon = 1$ (VL « pure ») et $\varepsilon = 0.5$. Dans le cas ou G_0 est complet la valeur $\varepsilon = 0.5$ correspond à la moyenne géométrique entre $|A|$ et $|B|$. La valeur la plus appropriée pour notre contexte (image) se situe, selon l'image, entre 0.35 et 0.5. Un autre paramètre π s'est avéré fondamental. Il s'agit de la valeur de l'indice brut $p(A,B)$ à partir de laquelle on autorise la possibilité d'une fusion de classes.
Pour aboutir à une telle stratégie, on posera pour tout couple de contigus $(x,y) \in G_0$:
$P(x,y) = P(x,y)$ *si* $P(x,y) > \pi$ et $P(x,y) = \eta$ *si* $P(x,y) \le \pi$; η étant une valeur suffisamment faible. De cette façon, on imposera de ne plus discriminer pour la classification les valeurs de l'indice probabiliste inférieures à π**.** Dans nos expériences la valeur $\pi = 0.45$ s'est avérée très adéquate.

5. Les notions de contiguïté « cont1 » et « cont2 » et l'inversion

Dans le plan image deux pixels sont contigus au sens de « cont1 », si et seulement s'il n'y a pas de pixel intermédiaire. Ainsi au sens de « cont1 », un pixel (sommet d'un graphe de contiguïté connexe G_0 dont une arête traduit la contiguïté entre 2 pixels) a au maximum 8 pixels qui lui sont contigus.
Au sens de « cont2 », deux pixels sont contigus si de plus la distance dite de Hamming (dans le plan image) est égale à 1. Ainsi au sens de « cont2 », un pixel a au maximum 4 pixels qui lui sont contigus (s'il est au bord de l'image il a 3 pixels contigus ; et s'il est au coin de l'image il a 2 pixels contigus). La transmissibilité de « cont2 » est assurée par transitivité. *(fig. 11)*

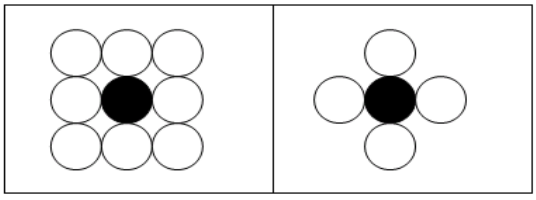


Fig. 11 *Graphes : « cont1 » et « cont2 »*

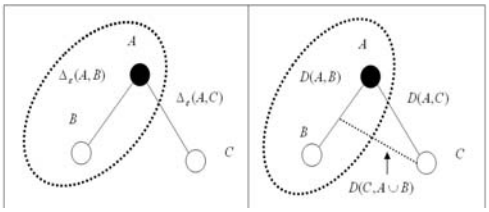


Fig. 12 *Inversion VL et Inertie*

L'indice VL de dissimilarité Δ_ε entre 2 classes, A et B, dépend du nombre d'arêtes reliant les 2 classes. Il ne dépend pas des masses de ces 2 classes ; contrairement à celui de l'inertie expliquée (noté D). Cela montre que l'indice VL ne produit pas d'inversion (*fig. 12*).

6. Les données traitées, les résultats obtenus et les temps de traitement

Il s'agit de 2 images satellitaires (*****) ; l'une « sat3 » (fig. 1) (758x419 pixels) et l'autre « sat2 » (fig. 8) (459x686 pixels). L'unique attribut retenu pour la segmentation est la variable luminance du pixel (niveau de gris). *(CAHCVR, avec Δ_ε ou D, peut être utilisé pour un nombre quelconque d'attributs sur des blocs de pixels.).* Les résultats obtenus avec les 2 indices sont tout à fait comparables. Notons quelques différences « visuelles » : dans l'image « sat3 » (fig. 1bis), avec 32 classes, la zone 1 est mieux définie par Δ_ε (fig. 3 et 4) que par D (fig. 5). L'augmentation du nombre de classes (700 classes ici) montre un éparpillement plus accentué avec D (fig. 7) qu'avec Δ_ε (fig. 6). Dans l'image « sat2 » (fig. 8bis) les zones 3, 4 et 5 sont bien identifiées avec D (fig. 10) ; les zones 3 et 4 se trouvent dans une même classe avec Δ_ε (fig. 9). En revanche on a une meilleure définition de la zone 1 avec Δ_ε (fig. 9).

	Image Sat2	*Image Sat3*
Cont1	**113**	**98**
Cont2	**62**	**57**

Temps de traitement en secondes (NEC 1200Mhz)

7. Conclusion

Formellement la complexité temporelle (temps de traitements) de CAHCVR, avec agrégation binaire, est linéaire en nombre d'éléments [BAC 98]. La pratique a déjà confirmé ce résultat. Le choix de l'agrégation multiple et le choix du graphe de contiguïté initial ont joué un rôle significatif.
En effet, l'agrégation multiple donne un « gain » des temps de l'ordre de 50% par rapport à l'agrégation binaire ; la contiguïté « cont2 » génère un gain de temps de 77% par rapport à « cont1 ». Cette amélioration des temps de traitements est justifiée formellement. La « qualité » des résultats s'améliore aussi. Les résultats de nos expériences ont mis en évidence une sensibilité importante liée au paramétrage (nombre de classes, les paramètres ε et π **).**
La comparaison « rigoureuse » des résultats, en terme de « qualité », avec Δ_ε et D, n'est pas un problème facile car nous n'avons pas de connaissance a priori claire sur les images testées (ces images sont « naturelles »). Une telle connaissance a priori pourrait se matérialiser par un étiquetage systématique de tous les pixels. C'est ce que nous avons essayé de faire en simulant des images simples où cette connaissance a priori est présente ; tous les « objets » définis dans ces images « artificielles » sont alors parfaitement identifiés (avec Δ_ε et avec D).

(*****) Nous sommes redevables à l'IGN de nous avoir transmis ces données images.

Bibliographie

[BAC 94] BACHAR K, *«Contribution en analyse factorielle et en CAH sous contraintes de contiguïté »*. Thèse (université de Rennes 1), 1994.
[BAC 98] BACHAR K, LERMAN I-C, *«Statistical conditions for a linear complexity for an algorithm of hierarchical classification under constraint of contiguity»*, in A. Rizzi, M. Vichi, H.-H. Bock (Eds) : Advances in Data Science and Classification / IFCS'98, Spinger-Verlag, pp. 131-136, 1998.
[LER 91] LERMAN I-C, *«Foundations of the Likelihood Linkage Analysis (LLA) Classification method»* . Applied Stochastic Models and Data Analysis, Vol.7, #1, p.63-76, John Wiley, march 1991.
[LER 01] LERMAN I-C, BACHAR K *«Agrégations multiples et contraintes de contiguïté dans la CAH utilisant les voisins réciproques et le critère VL »* 8ème Rencontres de la SFC, Guadeloupe, 2001.
[NIC 96] COSTA NICOLAU F, BACELAR-NICOLAU H, *«Some trends in the classification of variables»*, in Data Science, Classification and Related Methods / IFCS'96, Springer-Verlag, pp. 89-98, 1996.

Fig. 1 *Image « sat3 »*

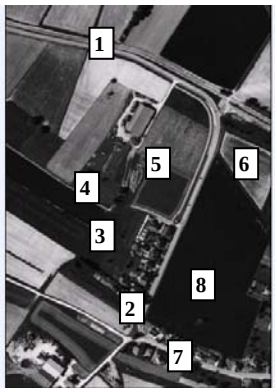


Fig. 1bis *Image « sat3 »*

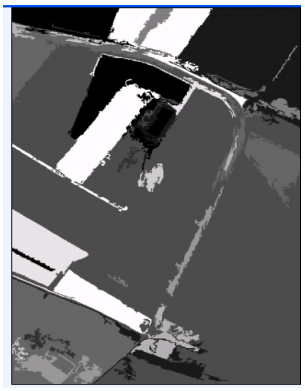


Fig. 2 *3sg432-11-v5045*

Fig. 3 *3sg440-11-v5040*

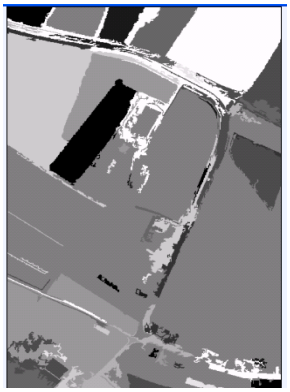


Fig. 4 *3sg432-11-v5050*

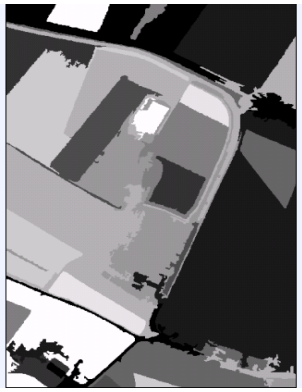


Fig. 5 *3sg432-11-i*

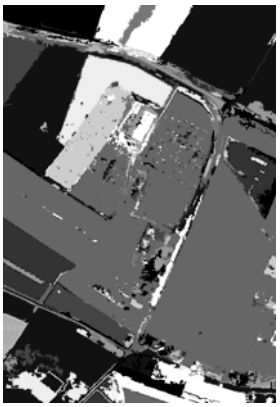


Fig. 6 *3sg4700-11-v5045*

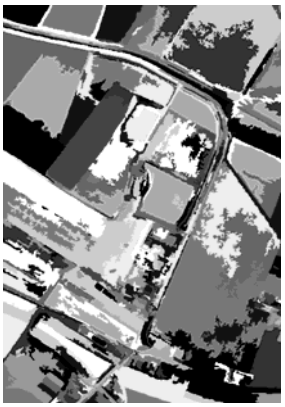


Fig. 7 *3sgin4700-11-i*

Fig. 8 *Image « sat2 »*

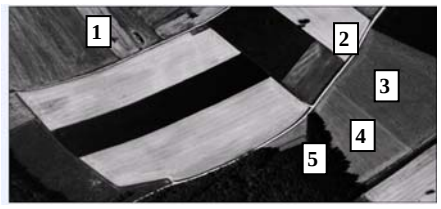


Fig. 8bis *Image « sat2 »*

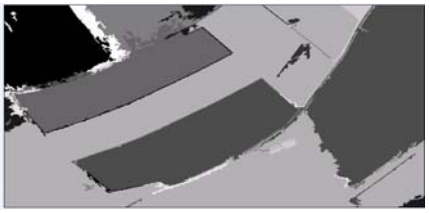


Fig. 9 *2sg438-11-v5040*

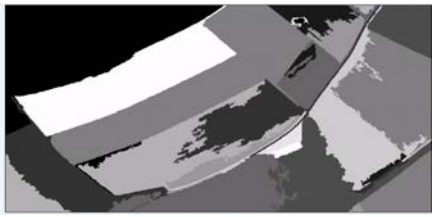


Fig. 10 *2sg432-11-i*

Classifications binaires et Quasi-hiérarchies

Jean-Pierre Barthélemy

ENST-Bretagne BP 832, 29285 Brest cedex (France) et
CAMS UMR CNRS 8557, EHESS, 54 Boulevard Raspail 75270 Paris cedex 06 (France)

RÉSUMÉ. *Dans cette contribution, nous étudions les systèmes classificatoires où toute classe se laisse engendrer par deux éléments seulement. Nous établissons deux théorèmes de bijection qui conduisent à des caractérisations combinatoires des quasi-hiérarchies.*

MOTS-CLÉS : *classifications binaires, quasi-hiérarchies, dissimilarités booléennes, relations ternaires*

1. Systèmes de classes.

Soit X un ensemble fini. Un *système de classes* (SC) sur X est un sous-ensemble $\mathcal{K}$ de l'ensemble 2^X de l'ensemble de tous les sous-ensembles de X qui vérifie :

(C1) $X \in \mathcal{K}$, $\phi \notin \mathcal{K}$ et pour tout $x \in X$, $\{x\} \in \mathcal{K}$

Les éléments de $\mathcal{K}$ sont appelés les classes de $\mathcal{K}$ et l'on dit que l'ensemble X et les singletons sont les classes triviales. Pour $A \subseteq X$, on désigne par $\mathcal{K}(A)$ l'ensemble des classes de $\mathcal{K}$ qui contiennent A.

On dit que le SC $\mathcal{K}$ est :

fermé lorsque : (F) $C, C' \in \mathcal{K}$ et $C \cap C' \neq \phi$ implique $C \cap C' \in \mathcal{K}$ (autrement dit $\mathcal{K}$ est fermé si et seulement si pour tout sous-ensemble, non vide, A de X, l'ensemble $\mathcal{K}(A)$, ordonné par inclusion, possède un élément minimum) ;

binaire lorsque :

(B1) pour tout $A \subseteq X$, avec $|A| = 2$, l'ensemble $\mathcal{K}(A)$, ordonné par inclusion, admet un élément minimum ;

(B2) pour toute classe C de $\mathcal{K}$, il existe $A \subseteq X$, avec $|A| = 2$ tel que C soit l'élément minimal de $\mathcal{K}(A)$.

Si $\mathcal{K}$ est un SC vérifiant (B1) on note, pour $A = \{x, y\}$, par $\delta_{\mathcal{K}}(x, y)$ la classe minimale de $\mathcal{K}(A)$ (cette notation sera justifiée plus loin).

pleinement binaire lorsque, outre (B1), on a :

(PB) pour tout $A \subseteq X$, non réduit à un singleton, il existe $u, v \in A$, tels que $\delta_{\mathcal{K}}(u, v) \in \mathcal{K}(A)$;

séparé lorsqu'il existe u et v tels que $\{u, v\}$ n'est contenu dans aucune classe non triviale.

Autrement dit "binaire" signifie que toute classe non réduite à un singleton est engendrée par exactement deux éléments. "Pleinement binaire" signifie que toute classe minimale contenant un sous-ensemble donné (non réduit à un singleton) est engendrée par deux éléments de celui-ci. Un SC pleinement binaire est binaire et un SC binaire est séparé. Un SC binaire n'est pas nécessairement fermé (et réciproquement). En revanche, nous verrons qu'un SC pleinement binaire est toujours fermé. Remarquons qu'un SC binaire possède au plus $\frac{n(n+1)}{2}$ classes (i.e. au plus $\frac{n^2-n+2}{2}$ classes non triviales.

Une *hiérarchie* est un SC $\mathcal{K}$ tel que : (H) Pour tout $A, B \in \mathcal{K}$, $A \cap B \in \{A, B, \phi\}$.

Une quasi-hiérarchie (Bandelt et Dress, 1989, Diatta et Fichet, 1994) est un SC fermé $\mathcal{K}$ tel que : (Q) Pour tout $A, B, C \in \mathcal{K}$, $A \cap B \cap C \in \{A \cap B, A \cap C, B \cap C\}$. On remarque que toute hiérarchie est une quasi-hiérarchie. Le résultat ci-dessous étend une observation de Bandelt et Dress (1989).

Proposition 1 *Soit $\mathcal{K}$ un SC, les deux assertions ci-dessous sont équivalentes :*

(i) $\mathcal{K}$ est pleinement binaire,

(ii) $\mathcal{K}$ est une quasi-hiérarchie.

2. Dissimilarités booléennes

Une *dissimilarité booléenne* sur X est une fonction δ de $X \times X$ dans 2^X telle que :

(DB1) pour tout $x \in X$, $\delta(x, x) = \{x\}$

(DB2) pour tout $x, y \in X$, $\delta(x, y) = \delta(y, x)$

(DB3) pour tout $x, y \in X$, avec $x \neq y$, $\{x, y\} \in \delta(x, y)$.

On dit qu'une dissimilarité booléenne δ est

séparée lorsque :

(BD4) il existe $u, v \in X$ tels que $X = \delta(u, v)$;

convexe lorsque :

(BD5) pour tout $x, y, z, t \in X$, si $z, t \in \delta(x, y)$, on a $\delta(z, t) \subseteq \delta(x, y)$.

Soit $\mathcal{K}$ un SC vérifiant la condition (B1), Posons $\delta_{\mathcal{K}}(x, x) = \{x\}$. Il est clair que $\delta_{\mathcal{K}}$ est une dissimilarité booléenne convexe. De plus $\delta_{\mathcal{K}}$ est séparée si et seulement si $\mathcal{K}$ est séparé. Réciproquement si δ est une dissimilarité booléenne convexe et séparée, l'ensemble $\mathcal{K}[\delta] = \{\delta(x, y) \mid (x, y) \in X \times X\}$ est un SC binaire.

Une dissimilarité booléenne δ, convexe et séparée est appelée :

une quasi-ultramétrique booléenne lorsque :

(BD6) pour tout $x, y, z \in X$, $\delta(x, y) \cap \delta(x, z) \cap \delta(y, z) \cap \{x, y, z\} \neq \phi$;

une ultramétrique booléenne lorsque :

(BD7) pour tout $x, y, z \in X$, $\delta(x, y) \cap \delta(y, z) \in \{\delta(x, y), \delta(y, z)\}$.

Il est clair qu'une ultramétrique booléenne est une quasi-ultramétrique booléenne.

On désigne par φ l'application qui à tout système de classe binaire $\mathcal{K}$, sur X, associe la dissimilarité booléenne (convexe et séparée) $\delta_{\mathcal{K}}$ et par ψ l'application qui à toute dissimilarité booléenne convexe et séparée δ associe le SC binaire $\mathcal{K}[\delta]$.

Proposition 2 *φ est une bijection de l'ensemble de tous les systèmes de classes binaires sur X dans l'ensemble de toutes les dissimilarités booléennes convexes et séparées sur X. L'application ψ est l'inverse de φ. De plus :*

- $\mathcal{K}$ est une quasi-hiérarchie si et seulement si $\varphi(\mathcal{K}) = \delta_{\mathcal{K}}$ est une quasi-ultramétrique booléenne ;

- $\mathcal{K}$ est une hiérarchie si et seulement si $\varphi(\mathcal{K}) = \delta_{\mathcal{K}}$ est une ultramétrique booléenne.

3. Quasi-hiérarchies et relations ternaires

Rappelons qu'une *relation ternaire* sur X est un sous-ensemble T de $X \times X \times X$. Si T est une relation ternaire, on écrira $T(x, y, z)$ plutôt que $(x, y, z) \in T$.

On dit qu'une relation ternaire sur X est :

(1,2)-symétrique lorsque pour tout $x, y, z \in X$, $T(x, y, z)$ implique $T(y, x, z)$;

3-réflexive lorsque pour tout $x, y \in X$, $T(x, y, x)$;

convexe lorsque pour tout $x, y, z, t \in X$, $T(x, y, z)$ et $T(x, z, t)$ implique $T(x, y, t)$;

séparée lorsqu'il existe u et v tels que pour tout $t \in X$, $T(u, v, t)$;

complète lorsque pour tout $x, y, z \in X$, on a $T(x, y, z)$, ou $T(x, z, y)$, ou $T(y, z, x)$;

emboîtée lorsque, pour tout $x, y, z \in X$, s'il existe $t \in X$ vérifiant $T(x, y, t)$ et non $T(y, z, t)$, on a pour tout $t' \in X$, $T(y, z, t')$ implique $T(x, y, t)$.

On dit qu'une relation ternaire T sur X est :

bi-classifiante lorsqu'elle est (1,2)-symétrique, 3-réflexive, convexe et séparée ;

quasi-hiérarchique lorsqu'elle est bi-classifiante et complète ;

hiérarchique lorsqu'elle est bi-classifiante et emboîtée.

Il est trivial – et bien connu – que la donnée d'une relation ternaire sur X est équivalente à la donnée d'une application de $X \times X$ dans 2^X. Cette équivalence se manifeste, via la bijection ξ qui associe, pour toute relation ternaire T, à $(x, y) \in X \times X$, l'ensemble des $z \in X$ tels que $T(x, y, z)$.

Proposition 3 *L'application ξ définit une bijection de l'ensemble des relations ternaires bi-classifiantes sur X vers l'ensemble de tous les systèmes de classes binaires sur X. De plus, pour une relation ternaire T :*

- $\xi(T)$ est une quasi-hiérarchie si et seulement si T est quasi-hiérarchique,

- $\xi(T)$ est une hiérarchie si et seulement si T est hiérarchique.

La dernière assertion de cette proposition retraduit, à une dualité près, la caractérisation des hiérarchies obtenue par Colonius et Schulze (1981), la première porte sur la "duale" de la relation de séparation introduite par Bandelt (1992). Le recours aux dissimilarités booléennes (et aux propositions 1 et 2) permet cependant de grandement simplifier les démonstrations. L'utilité des classifications binaires et leur application à l'analyse des dissimilarité est discuté dans Brucker (2003).

4. Bibliographie

[BAN 89] BANDELT H.-J., DRESS W. M., Weak hierarchies associated with similarity measures – an additive clustering technique, *Bulletin of Mathematical Biology*, vol. 51, n° 1, 1989, p. 133-166.

[BAN 92] BANDELT H.-J., Four-point characterization of the dissimilarity function obtained from indexed closed weak hierarchies, Mathematisches Semminar, 1992.

[BRU 03] BRUCKER F., Réalisations de dissimilarités, *Actes des Rencontres de la Société Francophone de Classification*, 2003, ce volume.

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

[DIA 94] DIATTA J., FICHET B., *New approaches in classification and data analysis*, Chapitre From Asprejan hierachies and Bandelt-Dress weak-hierarchies to quasi-hierarchies, p. 111-118, Springer-Verlag, Berlin, 1994.

Sur la comparaison et la visualisation des partitions floues

François Bavaud

Université de Lausanne
CH-1015 Lausanne-Dorigny
Francois.Bavaud@imm.unil.ch

RÉSUMÉ. *Une partition floue assigne à chaque objet (parmi n objets) une distribution sur a catégories. Par des méthodes d'algèbre linéaire élémentaire, on définit et étudie des propriétés telles que l'emboîtement des partitions, leur itération, ou leur stabilité en relation avec une autre partition. L'introduction de mesures de similarités "naturelles" entre objets, non pondérées (R) ou pondérées (T, P) permet de définir des distances euclidiennes entre objets, mais aussi entre partitions, lesquelles peuvent alors être représentées comme des points dans un espace factoriel de basse dimensionalité par MDS classique. Les versions pondérées T et P diffèrent pour les partitions floues, et engendrent diverses constructions formelles n'ayant pas d'équivalent au niveau des partitions ordinaires (=déterministes). Ce travail suggère une certaine vue de l'analyse multivariée de variables catégorielles floues, autrement dit de l'*analyse des correspondances floues multiples.[1].

MOTS-CLÉS : *Partitions floues, partitions itérées, similarités, MDS, visualisation d'objets, visualisation de partitions, matrice de transition, projection, emboîtement, comparaison de partitions.*

1. Matrices d'appartenance

Définition 1 Une partition *floue* $\mathcal{A}$ de n objets en a groupes est définie par une matrice $(n \times a)$ d'appartenance $Z^{\mathcal{A}} = (z_{ij})$ telle que $z_{ij} \geq 0$, $\sum_{j=1}^{a} z_{ij} = 1$ ($\forall i = 1, \ldots, n$) et $n_j^{\mathcal{A}} := \sum_{i=1}^{n} z_{ij}^{\mathcal{A}} > 0$ ($\forall j = 1, \ldots, a$). La notion d'appartenance en jeu peut s'interpréter comme z_{ij} = "probabilité que l'objet i appartienne au groupe j". La partition est *déterministe* si $z_{ij} = 1$ ou $z_{ij} = 0$ pour tous i, j, i.e. si $z_{ij}^2 = z_{ij}$. La partition est *pleine* si $\text{rang}(Z) = a$, et *défective* si $\text{rang}(Z) < a$.[1]

Une partition déterministe est pleine. En général, Z est formé de $c(\mathcal{A}) \leq a$ sous-blocs irréductibles indicés par $J = 1, \ldots, c(\mathcal{A})$. Chaque composante J est constituée (en lignes) de groupes j tels que $z_{ij'} = 0$ si $z_{ij} > 0$, pour $j \in J$ et $j' \notin J$; de même, chaque composante J est constituée (en colonnes) d'objets i tels que $z_{ij} = 0$ si $j \notin J$. On définit les matrices $(a \times a)$

$$B := Z'Z \quad \text{i.e.} \quad b_{jj'} := \sum_i z_{ij} z_{ij'} \qquad\qquad N := \text{diag}(\mathbf{1}'Z) \quad \text{i.e.} \quad n_{jj'} := \delta_{jj'}\, n_j \qquad [1]$$

1. Le travail a bénéficié de discussions stimulantes avec M. Rajman dans le cadre du projet UNIL-EPFL "Clavis" (2001).
1. voir par exemple MIRKIN, B. (1996) : "Mathematical Classification and Clustering", Kluwer, pp. 229-246 ou SAPORTA, G. (1990) : "Probabilités, analyse de données et statistique", Editions Technip, Paris, pp. 210-224 pour l'approche déterministe, et BEZDEK,J. (1981) "Pattern Recognition with Fuzzy Objective Function Algorithms". Plenum Press, New York, pour l'approche floue.

On a $b_{jj'} = 0$ si j et j' appartiennent à des composantes J et J' différentes. Aussi

$$\begin{array}{ccccc} c(\mathcal{A}) = m & \Leftrightarrow & \mathcal{A} \text{ est déterministe} & \Leftrightarrow & B = N \\ \text{rang}(Z) = m & \Leftrightarrow & \mathcal{A} \text{ est pleine} & \Leftrightarrow & B^{-1} \text{ existe} \end{array}$$

Lorsque $\mathcal{A}$ n'est pas déterministe, $G := N^{-1}B$ diffère de l'identité et génère des *partitions iterées* d'ordre r définies comme $Z^{(r)} := ZG^{r-1}$, avec limite $z_{ij}^{(\infty)} = n_j I(i \in J(j))/n_{J(j)}$ (où $I(E)$ est la fonction caractéristique de l'événement E). La partition $Z^{(\infty)}$ est pleine ssi $\mathcal{A}$ est déterministe.

Exemple 1 : on considère la partition floue $\mathcal{A}$ de $n = 5$ objets dans $a = 4$ groupes avec

$$Z^{\mathcal{A}} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0.2 & 0.8 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0.6 & 0.4 \\ 0 & 0 & 0.2 & 0.8 \end{pmatrix} \quad N = \begin{pmatrix} 1.2 & 0 & 0 & 0 \\ 0 & 1.8 & 0 & 0 \\ 0 & 0 & 0.8 & 0 \\ 0 & 0 & 0 & 1.2 \end{pmatrix} \quad B = \begin{pmatrix} 1.04 & 0.16 & 0 & 0 \\ 0.16 & 1.64 & 0 & 0 \\ 0 & 0 & 0.4 & 0.4 \\ 0 & 0 & 0.4 & 0.8 \end{pmatrix}$$

2. Similarités et distances entre objets

Soit $S = (s_{ii'})$ une matrice générale de similarités entre objets, définie non négative, et telle que $s_{ii'} \geq 0$, $s_{ii'} = s_{i'i}$ et $s_{ii'} \leq \sqrt{s_{ii}\, s_{i'i'}}$. Les matrices $R := ZZ'$, $T := ZN^{-1}Z'$ et (pour une partition pleine) $P := ZB^{-1}Z'$ constituent trois candidats naturels pour S :

$$r_{ii'} := \sum_{j=1}^{a} z_{ij}\, z_{i'j} \qquad t_{ii'} := \sum_{j=1}^{a} \frac{z_{ij}\, z_{i'j}}{n_j} \qquad p_{ii'} := \sum_{j,j'=1}^{a} z_{ij}\, b_{jj'}^{(-1)}\, z_{i'j'} \qquad [2]$$

$R = (r_{ii'})$ définit, pour une partition déterministe, la relation d'équivalence "i et i' appartiennent au même groupe". $T = (t_{ii'})$ est une matrice de transition markovienne de distribution stationnaire $\pi_i = 1/n$, vérifiant $\sum_{i'} t_{ii'} = 1$; $P = (p_{ii'})$ (pour lequel $p_{ii'} \geq 0$ peut être violé, $|p_{ii'}| \leq \sqrt{p_{ii}\, p_{i'i'}}$ restant valide) est une matrice de projection vérifiant $P^2 = P$. On a $T = P$ ssi la partition est déterministe. Aussi, les itérés de T et P construits à partir de $Z^{(r)} := Z\, G^{r-1}$ vérifient $T^{(r)} = T^{2r-1}$ et $P^{(r)} = P$.

Par le théorème de Schoenberg[2] (MDS classique), les quantités $D_{ii'}^{S} := (d_{jj'}^{S})^2 = s_{ii} + s_{i'i'} - 2s_{ii'}$ constituent des (carrés de) distances euclidiennes, pour lesquelles une configuration exacte de dimension $\leq n-1$ peut être reconstituée par MDS classique (diagonalisation de S). Explicitement

$$D_{ii'}^{R} = \sum_{j} (z_{ij} - z_{i'j})^2 \qquad D_{ii'}^{T} = \sum_{j} \frac{(z_{ij} - z_{i'j})^2}{n_j} \qquad D_{ii'}^{P} = \sum_{jj'} (z_{ij} - z_{i'j}) b_{jj'}^{(-1)} (z_{ij'} - z_{i'j'}) \qquad [3]$$

Ainsi, pour une partition déterministe, on a

$$\begin{array}{lll} r_{ii'} = 1 \quad t_{ii'} = p_{ii'} = \dfrac{1}{n_j} & D_{ii'}^{R} = D_{ii'}^{T} = D_{ii'}^{P} = 0 & \text{pour } i, i' \in j \\ r_{ii'} = t_{ii'} = p_{ii'} = 0 & D_{ii'}^{R} = 2 \quad D_{ii'}^{T} = D_{ii'}^{P} = \frac{1}{n_j} + \frac{1}{n_{j'}} & \text{pour } i \in j,\, i' \in j' \text{ avec } j \neq j' \end{array}$$

Exemple 1, suite : les matrices de similarité et distances *entre objets* associées sont

$$R = \begin{pmatrix} 1 & .2 & 0 & 0 & 0 \\ .2 & .68 & .8 & 0 & 0 \\ 0 & .8 & 1 & 0 & 0 \\ 0 & 0 & 0 & .52 & .44 \\ 0 & 0 & 0 & .44 & .68 \end{pmatrix} \quad T = \begin{pmatrix} .83 & .17 & 0 & 0 & 0 \\ .17 & .39 & .44 & 0 & 0 \\ 0 & .44 & .56 & 0 & 0 \\ 0 & 0 & 0 & .58 & .42 \\ 0 & 0 & 0 & .42 & .58 \end{pmatrix} \quad P = \begin{pmatrix} .98 & .12 & -.10 & 0 & 0 \\ .12 & .40 & .48 & 0 & 0 \\ -.10 & .48 & .62 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

2. SCHOENBERG, I.J. (1935) : "Remarks to Maurice Fréchet's article "Sur la définition axiomatique d'une classe d'espaces vectoriels distancés applicables vectoriellement sur l'espace de Hilbert"" *Annals of Mathematics, 36*, 724-732

$$D^R = \begin{pmatrix} 0 & 1.28 & 2 & 1.52 & 1.68 \\ 1.28 & 0 & .08 & 1.2 & 1.36 \\ 2 & .08 & 0 & 1.52 & 1.68 \\ 1.52 & 1.2 & 1.52 & 0 & .32 \\ 1.68 & 1.36 & 1.68 & .32 & 0 \end{pmatrix} D^T = \begin{pmatrix} 0 & .89 & 1.39 & 1.42 & 1.42 \\ .89 & 0 & .06 & .97 & .97 \\ 1.39 & .06 & 0 & 1.14 & 1.14 \\ 1.42 & .97 & 1.14 & 0 & .33 \\ 1.42 & .97 & 1.14 & .33 & 0 \end{pmatrix} D^P = \begin{pmatrix} 0 & 1.14 & 1.79 & 1.98 & 1.98 \\ 1.14 & 0 & .07 & 1.40 & 1.40 \\ 1.79 & .07 & 0 & 1.62 & 1.62 \\ 1.98 & 1.40 & 1.62 & 0 & 2 \\ 1.98 & 1.40 & 1.62 & 2 & 0 \end{pmatrix}$$

3. Partitions emboîtées

Définition 2 La partition $\mathcal{B}$ (définie par la matrice $(n \times b)$ d'appartenance $Z^{\mathcal{B}}$) est *plus grossière* que la partition $\mathcal{A}$ (définie par la matrice $(n \times a)$ d'appartenance $Z^{\mathcal{A}}$), i.e. $\mathcal{A}$ est *plus fine* que $\mathcal{B}$, noté $\mathcal{B} \leq \mathcal{A}$, si $Z^{\mathcal{B}} = Z^{\mathcal{A}}\, W^{\mathcal{AB}}$ où $W^{\mathcal{AB}} = (w_{jk}^{\mathcal{AB}})$ est une matrice $(a \times b)$ avec $w_{jk}^{\mathcal{AB}} \geq 0$ et $\sum_{k=1}^{b} w_{jk}^{\mathcal{AB}} = 1$.

La relation "$\mathcal{B} \leq \mathcal{A}$" est un ordre partiel, d'élément minimal $\mathcal{O}$ (avec $z_{ij}^{\mathcal{O}} \equiv 1$: partition à un groupe) et d'élément maximal $\mathcal{N}$ (avec $z_{ij}^{\mathcal{N}} = \delta_{ij}$: partition à n groupes). Aussi, si $\mathcal{B} \leq \mathcal{A}$ (partitions pleines), alors $P^{\mathcal{A}} P^{\mathcal{B}} = P^{\mathcal{B}} P^{\mathcal{A}} = P^{\mathcal{B}}$. La séquence $\mathcal{A}^{(r)}$ de partitions associées à $Z^{(r)}$ est décroissante : $\mathcal{A}^{(r+1)} \leq \mathcal{A}^{(r)}$.

4. Distances entre partitions

$S^{\mathcal{A}} = (s_{ii'}^{\mathcal{A}})$ et $S^{\mathcal{B}} = (s_{ii'}^{\mathcal{B}})$ étant les similarités associées aux partitions $\mathcal{A}$ et $\mathcal{B}$, une distance euclidienne (quadratique) entre ces dernières peut être définie comme

$$D^S_{\mathcal{A},\mathcal{B}} := \sum_{ii'} (s_{ii'}^{\mathcal{A}} - s_{ii'}^{\mathcal{B}})^2 = \mathrm{Tr}(S^{\mathcal{A}} - S^{\mathcal{B}})^2 = \mathrm{Tr}((S^{\mathcal{A}})^2) + \mathrm{Tr}((S^{\mathcal{B}})^2) - 2\mathrm{Tr}(S^{\mathcal{A}}\, S^{\mathcal{B}}) \qquad [4]$$

Par construction, le MDS classique appliqué à un ensemble de partitions munies de la distance $D^S_{\mathcal{A},\mathcal{B}}$ permet de visualiser exactement la configuration de façon euclidienne, chaque partition étant représentée par un point (voir figure 1).

Exemple 2 Soient $n = 5$ objets. On définit $\mathcal{A}$ comme dans l'exemple 1 ; $\mathcal{B}$ comme la partition $(123; 45)$; $\mathcal{C}$ comme$(12; 345)$; $\mathcal{D} \equiv \mathcal{N}$ comme $(1; 2; 3; 4; 5)$; $\mathcal{E} \equiv \mathcal{O}$ comme (12345), et $\mathcal{F} \equiv \mathcal{A}^{(\infty)}$ comme la partition itérée limite :

$$Z^{\mathcal{B}} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} Z^{\mathcal{C}} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} Z^{\mathcal{D}} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} Z^{\mathcal{E}} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} Z^{\mathcal{F}} = \begin{pmatrix} 0.4 & 0.6 & 0 & 0 \\ 0.4 & 0.6 & 0 & 0 \\ 0.4 & 0.6 & 0 & 0 \\ 0 & 0 & 0.4 & 0.6 \\ 0 & 0 & 0.4 & 0.6 \end{pmatrix}$$

qui définissent les dissimilarités *entre partitions* suivantes (dans l'ordre alphabétique) :

$$D^T = \begin{pmatrix} 0 & 0.63 & 1.37 & 1.74 & 1.74 & 0.63 \\ 0.63 & 0 & 1.11 & 3 & 3 & 0 \\ 1.37 & 1.11 & 0 & 3 & 3 & 1.11 \\ 1.74 & 3 & 3 & 0 & 0 & 3 \\ 1.74 & 3 & 3 & 0 & 0 & 3 \\ 0.63 & 0 & 1.11 & 3 & 3 & 0 \end{pmatrix} \qquad D^P = \begin{pmatrix} 0 & 2 & 2.63 & 1 & 1 \\ 2 & 0 & 1.11 & 3 & 3 \\ 2.63 & 1.11 & 0 & 3 & 3 \\ 1 & 3 & 3 & 0 & 0 \\ 1 & 3 & 3 & 0 & 0 \end{pmatrix}$$

En général, pour des partitions floues et pleines, $\mathcal{B} \leq \mathcal{A}$ entraîne $D^P_{\mathcal{A},\mathcal{B}} = \mathrm{Tr}(P^{\mathcal{A}}) - \mathrm{Tr}(P^{\mathcal{B}}) = a - b$, et donc $D^P_{\mathcal{A},\mathcal{C}} = D^P_{\mathcal{A},\mathcal{B}} + D^P_{\mathcal{B},\mathcal{C}}$ pour $\mathcal{C} \leq \mathcal{B} \leq \mathcal{A}$ (ou $\mathcal{A} \leq \mathcal{B} \leq \mathcal{C}$) .

Dans le cas de partitions $\mathcal{A}$ et $\mathcal{B}$ déterministes à a et b groupes, on a $D^R_{\mathcal{A},\mathcal{B}} = N_{\mathcal{A},\mathcal{A}} + N_{\mathcal{B},\mathcal{B}} - 2N_{\mathcal{A},\mathcal{B}}$ et $D^T_{\mathcal{A},\mathcal{B}} = D^P_{\mathcal{A},\mathcal{B}} = (a-1) + (b-1) - \frac{2}{n}\, \chi^2_{\mathcal{A},\mathcal{B}}$, où $N_{\mathcal{A},\mathcal{B}}$ est le nombre de paires d'objets classés dans le même groupe j de $\mathcal{A}$ *et* k de $\mathcal{B}$, et $\chi^2_{\mathcal{A},\mathcal{B}}$ est le chi2 usuel associé à la table de contingence n_{jk} (=nombre d'objets dans le groupe j de $\mathcal{A}$ et k de $\mathcal{B}$). De plus, $D^R_{\mathcal{A},\mathcal{B}} = \sum_{j=1}^{a} \rho_j^{\mathcal{B}}$ et $D^T_{\mathcal{A},\mathcal{B}} = D^P_{\mathcal{A},\mathcal{B}} = \sum_{j=1}^{a} \tau_j^{\mathcal{B}}$, où

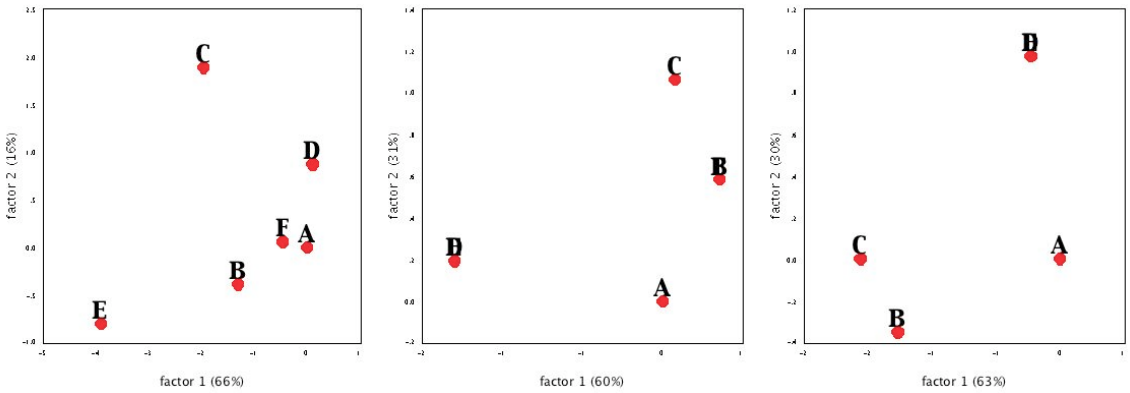


FIG. 1. *Visualisation euclidienne des distances entre partitions $D^S_{\mathcal{A},\mathcal{B}}$, pour les 6 partitions de l'exemple (2), dans les versions $S = R$ (gauche), $S = T$ (milieu) and $S = P$ (droite). $\mathcal{D}$ and $\mathcal{E}$ sont confondues dans les versions T et P ; $\mathcal{B}$ et $\mathcal{F}$ sont confondues dans la version T. Enfin, $\mathcal{F}$ est défective et donc non représentable dans la version P.*

$\rho_j^{\mathcal{B}} := n_j^2 - 2\sum_{i,i' \in j} r_{ii'}^{\mathcal{B}} + \sum_{i \in j; i'} r_{ii'}^{\mathcal{B}} \geq 0$ et $\tau_j^{\mathcal{B}} := 1 - \frac{2}{n_j}\sum_{i,i' \in j} p_{ii'}^{\mathcal{B}} + \sum_{i \in j} p_{ii}^{\mathcal{B}} \geq 0$. Les quantités $\rho_j^{\mathcal{B}}$ et $\tau_j^{\mathcal{B}}$ s'interprètent comme mesures d'instabilité du groupe j de $\mathcal{A}$ relativement à la partition $\mathcal{B}$. En particulier, $\rho_j^{\mathcal{N}} = n_j\,(n_j - 1)$ et $\tau_j^{\mathcal{N}} = n_j - 1$, qui souligne l'instabilité des grands groupes face à $\mathcal{N}$; aussi, $\rho_j^{\mathcal{O}} = (n - n_j)\,n_j$ (groupes moyens instables face à $\mathcal{O}$) et $\tau_j^{\mathcal{O}} = \frac{n - n_j}{n}$ (petits groupes instables face à $\mathcal{O}$).

Comment détecter le transfert latéral de gènes dans la classification des espèces

Alix Boc, Abdoulaye Baniré Diallo et Vladimir Makarenkov

Département d'informatique
Université du Québec à Montréal
Case postal 8888, succursale Centre-ville
Montréal (Québec) Canada - H3C 3P8
boc.alix@courrier.uqam.ca, diallo.abdoulaye_banire@courrier.uqam.ca et makarenkov.vladimir@uqam.ca

RÉSUMÉ. Cet article adresse le problème de détection de transferts latéraux de gènes (TLG) dans une phylogénie donnée. Nous décrivons ici une nouvelle méthode permettant la prédiction de probables TLG durant l'évolution d'un groupe d'organismes considérés. La méthode proposée procède par l'établissement des différences topologiques entre la phylogénie d'espèces et celle du gène en question. Elle utilise une procédure d'optimisation basée sur le critère des moindres carrés pour tester la possibilité d'un transfert latéral de gène entre tous les couples de branches de l'arbre d'espèces. Dans la partie application, nous montrons comment cette méthode permet de prédire d'éventuels transferts du gène rubisco rcbL dans une phylogénie incluant des algues, des cyanobactéries et des protéobactéries.

MOTS-CLÉS : arbre et réseau phylogénétique, transfert latéral de gènes, critère des moindres carrés

1. Introduction

Le processus d'évolution d'espèces a longtemps été modélisé à l'aide d'arbres phylogénétiques. Dans de tels arbres, chaque espèce ne peut être liée qu'avec son ancêtre le plus proche et autres relations inter-espèces telles que, par exemple, le transfert latéral de gène ne sont pas permises. Cependant, le transfert latéral de gènes joue un rôle clé dans l'évolution des espèces, en particulier des bactéries. En effet, de nombreux projets de séquençages de bactéries ont renforcé l'idée que l'analyse phylogénétique d'un groupe d'espèces doit tenir compte d'évènements évolutifs tels que la convergence, la duplication, la perte et le transfert latéral de gènes. Ces importants mécanismes ne peuvent alors être représentés qu'avec un model en réseau. Plusieurs tentatives d'utiliser des modèles en réseaux pour représenter le transfert latéral peuvent être trouvées dans la littérature scientifique, voir par exemple Hein (1990) ou Page et Charleston (1998). Récemment, un nouveau modèle de transfert latéral permettant d'inscrire l'arbre de gènes dans l'arbre d'espèces correspondant a été proposé par Hallet et Lagergreen (2001). Dans ce papier nous définissons un autre modèle de classification qui permet d'inscrire l'arbre de gènes dans l'arbre d'espèces en utilisant les moindres carrés. Nous montrerons comment les différences topologiques entre ces deux arbres peuvent être exploitées pour décrire de possibles scénarios du transfert latéral d'un gène considéré survenus au cours de l'évolution.

2. Description de la méthode

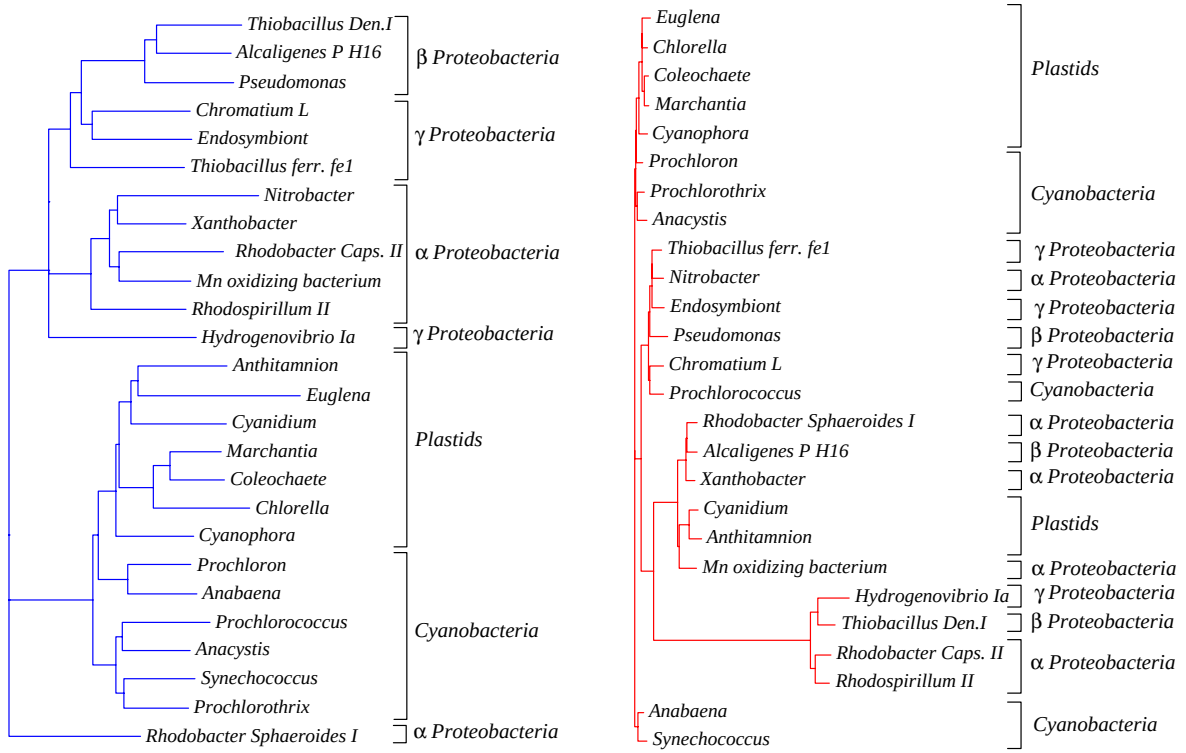
Dans cette section, nous décrivons une nouvelle méthode pour la détection des transferts latéraux. Elle procède par la réconciliation des topologies d'arbres de gènes et d'espèces (ou taxa). Cette méthode permet d'incorporer de nouvelles branches orientées qui représenteront les transferts latéraux dans l'arbre d'espèces. La méthode consiste en trois étapes principales décrites ci-dessous.

Étape 1. Soit T une phylogénie d'espèces dont les feuilles sont étiquetées selon un ensemble X de n taxa. L'arbre T peut être inféré à partir de séquences de nucléotides ou de protéines ou d'une matrice de distance en utilisant une méthode d'ajustement appropriée. T est un arbre binaire qui comprend $2n-3$ branches et dont les nœuds internes sont de degré 3. Cet arbre doit être explicitement enraciné car la position de la racine est importante dans notre modèle.

Étape 2. Soit T_1 une phylogénie de gène dont les feuilles sont étiquetées selon le même ensemble X de n taxa utilisé pour étiqueter l'arbre d'espèces T. T_1 peut aussi être inféré à partir de séquences biologiques ou de matrice de distance caractérisant ce gène en particulier. Si les topologies de T et de T_1 sont identiques alors aucun transfert latéral n'est présent. Par contre, si les deux arbres sont topologiquement différents, ceci peut être le résultat de transferts latéraux. Dans ce cas, l'arbre T_1 peut être inscrit dans l'arbre T en ajustant par les moindres carrés les longueurs des branches de T aux paires de distances entre les feuilles dans T_1.

Étape 3. Le but de cette étape est d'obtenir une liste ordonnée L de toutes les connexions possibles de TLG entre les pairs de branches de T. Cette liste comprendra $(2n\text{-}3)(2n\text{-}4)$ entrées ce qui correspond au nombre de connexions différentes entre les branches d'un arbre binaire à n feuilles. Chaque entrée de L est associée à une valeur de critère des moindres carrés obtenu par l'ajout d'une branche de connexion entre un couple de branches considérée dans T. La première entrée de L, celle qui minimise le plus le critère des moindres carrés, correspondra au cas le plus probable du transfert latéral, suivie par la deuxième connexion la plus probable et ainsi de suite.

Maintenant, nous montrons comment calculer le coefficient des moindres carrés pour une branche de TLG (a,b) ajoutée à T. Dans un arbre phylogénétique, il existe toujours un chemin unique entre une paire de sommets. L'ajout d'une nouvelle branche peut créer un autre chemin. Nous avons donc défini des règles qui permettent ou interdisent le passage par la nouvelle branche (a,b) lors de l'établissement d'un chemin entre deux sommets quelconques de l'arbre. Pour plus de détails sur ces règles, voir Boc et Makarenkov (2003). L'estimation du critère des moindres carrés se fait en quatre pas. Premièrement, on détermine toutes les paires de taxa telles que le chemin entre eux peut passer par la nouvelle branche (a,b). Deuxièmement, on sélectionne dans cet ensemble toutes les paires de taxa dont la distance peut diminuer après l'ajout de la branche (a,b). Troisièmement, on calcule une valeur optimale de la branche (a,b) selon le critère des moindres carrés. Et finalement, on réévalue au moins une fois la longueur de toutes les branches de l'arbre T. Ces calculs sont répétés pour toutes les paires de branches dans l'arbre T. Une liste ordonnée de tous les TLG possibles est alors produite.



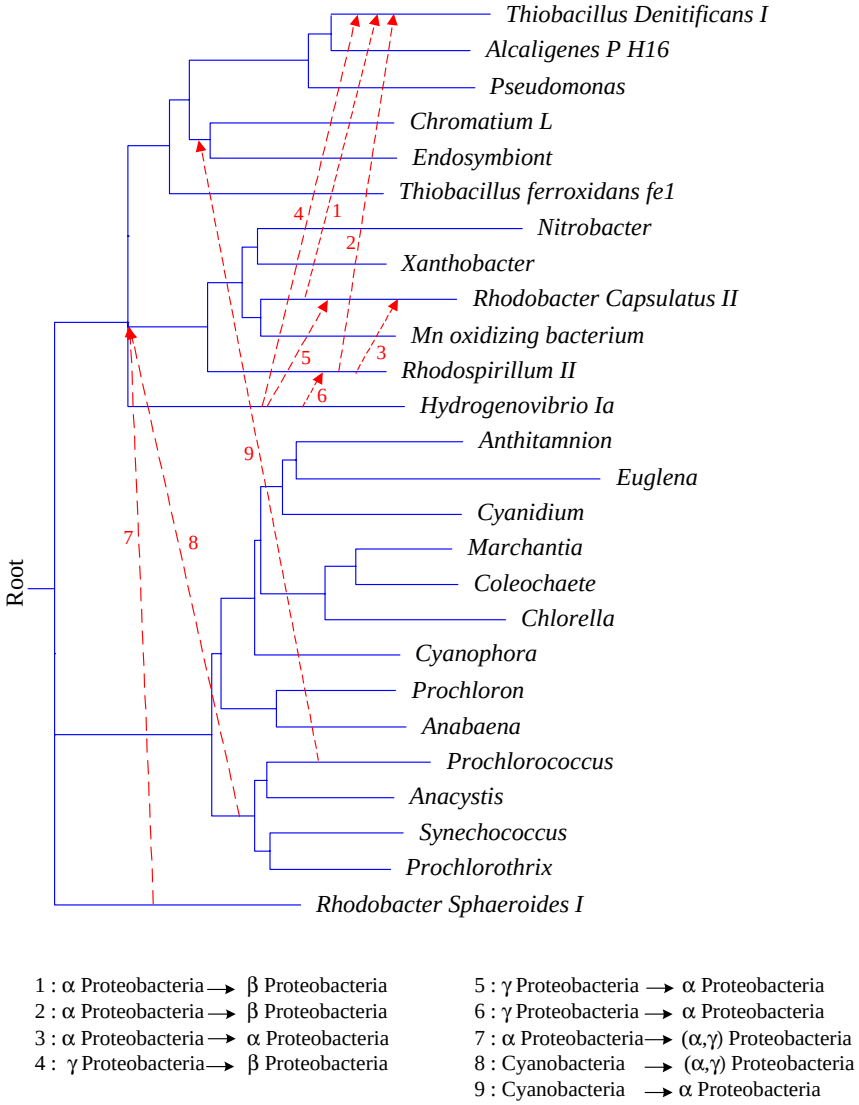
(a) Arbre d'espèces (inféré par NJ)

(b) Arbre du gène *rbcL* (inféré par NJ)

Figure 1. *(a) Phylogénie de 26 organismes construite avec la méthode Neighbour-Joining (Saitou et Nei, 1987) à partir des séquences du gène 16S rRNA. (b) Phylogénie du gène rbcL pour les mêmes 26 organismes inférée par la méthode NJ.*

3. Transfert latéral du gène *rbcL*

La méthode introduite dans cet article a été appliquée à l'analyse des données d'algues marines, de cyanobactéries et de protéobactéries considérées également par Delwiche et Palmer (1996). Les derniers auteurs ont examiné les différentes hypothèses du transfert latéral du gène rubisco entre ces trois groupes espèces. Delwiche et Palmer ont inféré, en utilisant la méthode de maximum de parcimonie, une phylogénie du gène *rbcL* (rubisco) de 48 espèces et ont trouvé que cette classification présentait un certain nombre de conflits en comparaison avec la phylogénie d'espèces inférée à partir du gène ribosomal 16S rRNA et d'autres évidences.



1 : α Proteobacteria → β Proteobacteria
2 : α Proteobacteria → β Proteobacteria
3 : α Proteobacteria → α Proteobacteria
4 : γ Proteobacteria → β Proteobacteria
5 : γ Proteobacteria → α Proteobacteria
6 : γ Proteobacteria → α Proteobacteria
7 : α Proteobacteria → (α,γ) Proteobacteria
8 : Cyanobacteria → (α,γ) Proteobacteria
9 : Cyanobacteria → α Proteobacteria

Figure 2. *Arbre d'espèces de la Figure 1a avec 9 flèches représentant les possibles transferts latéraux du gène rbcL trouvés par notre méthode. Le numéro sur les flèches indique leur ordre d'apparition (i.e. ordre d'importance) dans la liste de tous les TLG obtenus.*

Pour appliquer notre méthode, nous avons tout d'abord essayé de reconstruire l'arbre pour les 48 espèces à partir des séquences ribosomiques disponibles dans la base de données NCBI et dans Ribosomal Database Project. Cependant, les séquences recherchées étaient retrouvées pour seulement 26 des 48 espèces. La Figure 1 montre les phylogénies d'espèces (1a) et du gène *rbcL* (1b) inférées avec la méthode NJ. L'arbre d'espèces étant basé sur le gène ribosomal 16S rRNA. En observant les topologies de ces deux arbres, on note d'importantes différences entre elles. Par exemple on peut voir un cluster comprenant trois γ-protéobactéries entre lesquelles on trouve une α-protéobactérie, une β-protéobactérie et une cyanobactérie. Ces contradictions peuvent être expliquées soit par des transferts

latéraux de gènes qui auraient pu se produire entre les espèces indiquées, soit par la duplication du gène *rbcL*. Ces deux hypothèses ne sont pas mutuellement exclusive, voir Delwiche et Palmer (1996).

Notre méthode a été appliquée aux deux phylogénies de la Figure 1(a et b) et a produit une liste de transferts latéraux éventuels ordonnée selon leur importance. La solution obtenue est décrite par les neuf transferts latéraux de gènes illustrés sur la Figure 2. Le transfert entre *Rhodobacter Capsulatus II* et *Thiobacillus Denitrificans I* est le plus significatif suivi du transfert entre *Rhodopirillum II* et *Thiobacillus Denitrificans I* et ainsi de suite. Delwiche et Palmer (1996, figure 4) indiquent quatre TLG du gène rubisco entre les cyanobactéries et les γ-protéobactéries, les γ-protéobactéries et les α-protéobactéries, les γ-protéobactéries et les β-protéobactéries et finalement, entre les α-protéobactéries et les algues rouges et brunes. Notre méthode a permit de retrouver tous les transferts latéraux indiqués par Delwiche et Palmer (1996) sauf celui des α-protéobactéries vers les algues rouges et brunes.

4. Conclusion

Nous avons développé une nouvelle méthode pour détecter des transferts latéraux de gènes dans une phylogénie d'espèces. Cette méthode exploite les différences topologiques entre un arbre d'espèces et un arbre de gènes construits pour un même ensemble d'organismes. L'arbre de gènes est d'abord inscrit dans l'arbre d'espèces et puis la possibilité du transfert latéral entre chaque paire de branches de l'arbre d'espèces est estimée. Notre méthode génère une liste ordonnée de transferts latéraux entre les branches de la phylogénie d'espèces. Les éléments de cette liste doivent être analysés avec précaution prenant en compte toutes les informations disponibles sur les données traitées. Dans cet article, nous avons développé un model basé sur le critère des moindres carrés. Il serait intéressant d'étendre et de tester cette procédure dans le cadre des models de maximum de vraisemblance et de maximum de parcimonie. La méthode de détection de transferts latéraux de gènes décrite dans ce papier sera incluse (dans la version de mai 2003) au package *T-Rex* de Makarenkov (2001). Ce programme, disponible pour les plates-formes Windows et Macintosh, est mise à la disposition des chercheurs à l'adresse URL suivante : <http://www.fas.umontreal.ca/biol/casgrain/en/labo/t-rex>.

5. Bibliographie

[BOC 03] BOC, A., MAKARENKOV, V., "New Efficient Algorithm for Detection of Horizontal Gene Transfer Events", soumis à WABI 2003.

[DEL 96] DELWICHE, C.F., PALMER, J.D., "Rampant horizontal transfer and duplication of Rubisco genes in Eubacteria and Plastids", *Mol Biol. Evol*., vol. 13(6), 1996, p. 873-882.

[HAL 01] HALLET, M., LAGERGREN, J., "Efficient algorithms for lateral gene transfer problems", RECOMB 2001, Montreal, Canada, p. 149-156

[HAS 93] von HASELER, A., CHURCHILL, G. A., "Network models for sequence evolution", *J. Mol. Evol*, vol. 37, 1993, p. 77-85.

[HEI 90] HEIN, J., "A heuristic method to reconstructing the evolution of sequences subject to recombination using parsimony", *Math. Biosci.*, 1990, p. 185-200.

[MAK 01] MAKARENKOV, V., "T-Rex: reconstructing and visualizing phylogenetic trees and reticulation networks", *Bioinformatics*, vol. 17, 2001, p. 664--668.

[PAG 98] PAGE, R. D. M., CHARLESTON, M. A., "From gene to organismal phylogeny: Reconciled trees", *Bioinformatics*, vol. 14, 1998, p. 819-820.

[SAI 87] SAITOU, N., NEI, M., "The neighbour-joining method: a new method for reconstructing phylogenetic trees", *Mol. Biol. Evol.*, 1987, p. 406-425.

[THE 02] The NCBI handbook [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; 2002 Oct. Chapter 17, The Reference Sequence (RefSeq) Project.

Une Méthode Type Nuées Dynamiques pour des Données Symboliques Quantitatives

F. A. T. de Carvalho
Centro de Informática – Cin / UFPE
Av. Prof. Luiz Freire, s/n, Cidade Universitária
CEP: 50740-540 Recife-PE, BRASIL
fatc@cin.ufpe.br

P. Brito
Faculdade de Economia
Universidade do Porto
Rua Dr. Roberto Frias
4200-464 Porto, PORTUGAL
mpbrito@fep.up.pt

H. H. Bock
Technical University of Aachen
Institute of Statistics
D-52056 Aachen, GERMANY
bock@stochastik.rwth-aachen.de

RÉSUMÉ. Ce travail présente une méthode de type nuées dynamiques pour des données symboliques de type intervalle. On s'intéresse en particulier le problème de la standardisation; trois méthodes distinctes de standardisation, adaptées á ce type de données, sont proposées. La méthode est évaluée sur la base d'exemples simulés et de données réelles.

MOTS-CLÉS : Classification, nuées dynamiques, données intervalles

1. Les données

Dans ce travail, on présente une méthode permettant de construire une partition d'un ensemble d'objets décrits par des variables symboliques quantitatives, dont les valeurs peuvent être soit des intervalles de R soit des points de R (c'est-à-dire, des intervalles degenerés).

Les données en entrée constituent un tableau de données symboliques, où, pour chaque objet ω_i en ligne, et chaque variable y_j en colonne, on observe soit une valeur de R, comme

dans le cas classique, soit un intervalle. En conséquence, dans une même colonne, on peut trouver à la fois des valeurs réelles et des intervalles :

	y_1	…	y_j	...	y_p
ω_1	$[a_{11}, b_{11}]$	...	$[a_{1j}, b_{1j}]$	...	$[a_{1p}, b_{1p}]$
...	...		...		...
ω_i	$[a_{i1}, b_{i1}]$	...	$[a_{ij}, b_{ij}]$	…	$[a_{ip}, b_{ip}]$
…	…		…		…
ω_n	$[a_{n1}, b_{n1}]$	…	$[a_{nj}, b_{nj}]$	…	$[a_{np}, b_{np}]$

où l'on peut avoir $a_{ij}= b_{ij}$ pour quelques i,j. À chaque ω_i est canoniquement associé un objet symbolique, $[y_1 \in I_{i1}] \wedge \cdots \wedge [y_p \in I_{ip}]$, avec $I_{ij} = [a_{ij},b_{ij}]$, i=1,…,n, j=1,…,p.

2. La Méthode

La méthode proposée est basée sur la méthodologie des nuées dynamiques, c'est-à-dire que, ayant définit une méthode de représentation des classes, on construit la partition en appliquant itérativement une fonction d'allocation et une fonction de représentation. À la convergence, on atteint un optimum local d'un critère qui évalue l'ajustement entre les classes obtenues et leurs représentants.

Les fonctions d'allocation et de représentation sont basées sur une distance dans l'espace de description. La méthode utilise une distance de type L2 entre intervalles ou hypercubes dans R^p.

Soit $I_1 = [a_1 , b_1]$ et $I_2 = [a_2 , b_2]$, alors la distance entre des intervalles I_1 et I_2 s'écrit :

$$D(I_1,I_2) = \left[|a_1 - a_2|^2 + |b_1 - b_2|^2 \right]^{\frac{1}{2}}$$

On considère pour représentant de chaque classe le vecteur d'intervalles $(I_1, \ldots, I_p)$, avec $I_j= [a_j,b_j]$, j=1,…,p, dont les bornes a_j et b_j sont respectivement la moyenne des bornes inférieures et la moyenne des bornes supérieures calculées pour les éléments de la classe. Ce choix est en accord avec la métrique choisie, et garantit donc la convergence de la méthode.

2.1. *Standardisation*

Quand on est en présence de variables quantitatives, les différentes échelles que les variables peuvent présenter ont un effet important dans le calcul des distances et donc dans le résultat final d'une méthode de classification. Pour faire face à ce problème, trois méthodes distinctes de standardisation, adaptées à des données de type intervalle, sont proposées.

La standardisation doit être effectuée variable par variable, ainsi, la même transformation doit être appliquée aux bornes supérieure et inférieure des intervalles.

Dans les deux premiers cas, on fait la standardisation en centrant les données par rapport à la moyenne des centres des intervalles; la première méthode évalue la dispersion par la

dispersion de ces centres, alors que la deuxième méthode considère la dispersion des bornes des intervalles.

La troisième méthode proposée effectue une standardisation par rapport au maximum des bornes supérieures et au minimum des bornes inférieures.

2.2. *Les sorties*

Pour évaluer et interpréter les résultats de la méthode, on décrit chaque classe de la partition obtenue par des objets symboliques: un objet "central" prototype et un objet généralisant.

La qualité et les propriétés de la partition finale sont en outre interprétées en termes d'inertie expliquée, contribution des classes à l'inertie globale, et contributions des variables aux classes. Ces mesures sont définies comme des extensions au cas de données de type intervalle des mesures correspondantes pour des données classiques monovaluées.

Pour évaluer la performance de la méthode proposée, on l'a appliquée à des données simulées en 2 dimensions présentant différentes configurations. On illustre également la méthode par une application à des données réelles de dimension supérieure.

3. Bibliographie

[BOCK 00] BOCK, H.H.; DIDAY, E. (2000), *Analysis of Symbolic Data*, *Exploratory Methods for Extracting Statistical Information from Complex Data*. Springer Verlag, Heidelberg.

[BOCK 01] BOCK, H.H. (2001), "Clustering algorithms and Kohonen maps for symbolic data", in: *Proc. International Conference on New Trends in Computational Statistics with Biomedical Applications*, Osaka, 2001, pp. 203-215.

[CEL 89] CELEUX, G.; DIDAY, E.; GOVAERT, G.; LECHEVALLIER, Y.; RALAMBONDRAINY, H. (1989), *Classification Automatique des Données*, Bordas, Paris.

[CAR 98] DE CARVALHO, F. A. T. AND SOUZA, R. M. C. R. (1999), "New metrics for constrained Boolean symbolic objects", in: *Studies and Research: Proceedings of the Conference on Knowledge Extraction and Symbolic Data Analysis* (KESDA'98), Office for Official Publications of the European Communities, Luxembourg, 175-187.

[CAR 99] DE CARVALHO, F. A T.; VERDE, R.; LECHEVALLIER, Y. (1999), "A dynamical clustering of symbolic objects based on a context dependent proximity measure", in: *Proc. IX International Symposium on Applied Statistics and Stochastic Models – ASMDA'99*, Bacelar Nicolau et al (Eds.), LEAD, Univ. Lisboa, pp. 237-242.

[DID 72] DIDAY, E. (1972), *Nouveaux Concepts et Nouvelles Méthodes en Classification Automatique,* Thèse d'Etat, Univ. Paris VI.

[DID 89] DIDAY, E.; BRITO, P. (1989), "Symbolic Cluster Analysis", in: *Conceptual and Numerical Analysis of Data*, *Proc. of the 13th Conf. of The Gessellschaft für Klassifikation*., Univ. de Augsburg, Avril 1989, ed. Otto Opitz, Springer-Verlag, Heidelberg.

[DID 80] DIDAY, E.; GOVAERT, G.; LECHEVALLIER, Y., SIDI, J. (1980), "Clustering in Pattern Recognition", in: *Proc. of the NATO Advanced Study Institute on Digital Image Processing and Analysis*, Bonas 1980, ed. J. C. Simon

[DID 76] DIDAY, E.; SIMON, J. C. (1976), Clustering Analysis, in: *Digital Pattern Recognition*, Fu, K. S. (Eds.), Springer Verlag, Heidelberg, pp. 47-94.

[GOR 00] GORDON, A. D. (2000), “An iterative relocation algorithm for classifying symbolic data”, in: *Data Analysis: Scientific Modelling and Practical Application*, W Gaul, O Opitz and M Schader, eds., Springer, Berlin, pp. 17-23.

[OK 75] OK-SAKUN, Y. (1975), *Analyse Factorielle Typologique et Lissage Typologique*, Thèse de 3éme cycle, Univ. Paris VI.

[RAL 95] RALAMBONDRAINY, H. (1995), “A conceptual version of the K-means algorithm”, in: *Pattern Recognition Letters* 16, pp. 1147-1157.

[SCH 76] SCHROEDER, A. (1976), « Analyse d’un mélange de distributions de probabilité de même type », in : *R.S.A*. Vol. 24, nº1.

[VER 00] VERDE, R.; DE CARVALHO, F.A. T.; LECHEVALLIER, Y. (2000), “A dynamical clustering algorithm for multi-nominal data”, in: *Data Analysis, Classification and Related Methods*, Kiers, H. A. L. et al (Eds.), Springer Verlag, pp.387-394.

Ultramétrique en sandwich entre deux dissimilarités

Raphaël Bolze, Alain Guénoche

Institut de Recherche en Informatique de Nantes et Institut de Mathématiques de Luminy, Marseille
bolze.raphael@wanadoo.fr, guenoche@iml.univ-mrs.fr

RÉSUMÉ. *Le problème de l'ajustement d'une ultramétrique à une distance donnée D est traité en construisant le plus petit intervalle de distances qui contienne à la fois D et une ultramétrique U. C'est cette ultramétrique qui est la solution du problème d'ajustement. Une application est faite à des données décrites par des variables d'intervalle.*

MOTS-CLÉS : *Ultramétrique, norme infinie, variables d'intervalles*

1. Ultramétrique optimale en norme infinie

Un problème important en classification est l'approximation d'une dissimilarité donnée D, sur un ensemble X à n éléments, par une distance ultramétrique U. La qualité de cet ajustement peut être mesurée selon plusieurs normes, L_1, L_2 ou L_∞ étant les plus classiques. Pour les deux premières, le problème de calculer une ultramétrique optimale est NP-difficile [Krivanek et Moravek, 1986] alors que pour la norme infinie, Farach, Kannan et Warnow [1995] ont montré que, trouver une distance ultramétrique U telle que $\|D - U\|_\infty$ est minimum, est un problème de complexité polynômiale. Chepoi et Fichet [2000] ont grandement simplifié les choses en introduisant la notion d'ultramétrique sous-dominante dans la construction de Farach et al., qui en ignoraient l'existence. Ils ont montré que l'ultramétrique sous-dominante de D, notée $U_{sd}(D)$, augmentée pour chaque valeur d'une constante

$$k = \frac{1}{2} \max_{x,y} \|D(x,y) - U_{sd}(D)(x,y)\|,$$

soit $U^r_\infty(D) = U_{sd}(D) + k$, est optimale au sens de la norme L_∞.

Dans leurs articles respectifs, ces auteurs se sont principalement attachés à évaluer la complexité du calcul, en $O(n^2)$, sans comparer les solutions optimales qui ne sont pas toutes équivalentes. Notons $\mathbb{U}_\infty(D)$ l'ensemble des ultramétriques optimales pour D au sens de la norme infinie.

$$\forall U \in \mathbb{U}_\infty(D), \|D - U\|_\infty \text{est minimum.}$$

Dans la première partie, nous définissons une ultramétrique particulière, notée $U_\infty(D)$ qui s'ajuste le mieux à D, toujours au sens de cette même norme. C'est à dire que pour toute paire (x, y) d'éléments de X,

$$\forall U \in \mathbb{U}_\infty(D), |D(x,y) - U_\infty(x,y)| \leq |D(x,y) - U(x,y)|.$$

Par rapport à l'ultramétrique de référence, $U^r_\infty(D)$, ceci revient à déterminer non plus une constante k valable pour toute paire (x, y) de X, mais une série de constantes $k_1, ..k_{n-1}$ chacune étant appropriée à un sous-arbre du dendrogramme de $U_{sd}(D)$. Cette série de constantes se calcule directement à partir des longueurs des arêtes d'un arbre minimum de D. Le principe de l'algorithme est de rechercher, pour chaque arête, les deux classes qui sont réunies dans le dendrogramme et de calculer la constante globale sur ce sous-tableau de distance. L'algorithme est en $O(n^2)$ si toutes les longueurs sont différentes. Dans le cas de deux arêtes adjacentes et de même longueur, un pré-calcul est nécessaire. On notera que la topologie du dendrogramme associé est la même que celle du lien unique.

Exemple

D	x	y	z	t	u	v
y	6					
z	7	12				
t	5	3	4			
u	1	6	3	7		
v	3	16	6	9	2	
w	4	12	2	11	5	7

U_{sd}	x	y	z	t	u	v
y	4					
z	3	4				
t	4	3	4			
u	1	4	3	4		
v	2	4	3	4	2	
w	3	4	2	4	3	3

TAB. 1. *Une dissimilarité D et son ultramétrique sous dominante $U_{sd}(D)$; ici $k = 6$.*

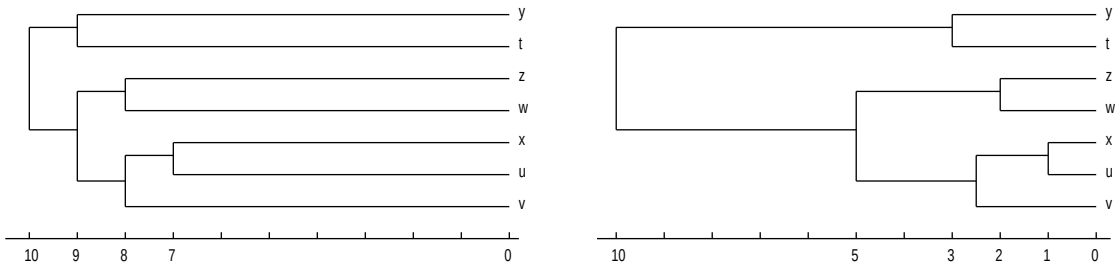


FIG. 1. *Dendrogrammes de l'ultramétrique infinie de D, à gauche l'ultramétrique U_∞^r, à droite le cas ajusté U_∞*

2. Ultramétrique entre deux dissimilarités

De façon naturelle, l'ensemble des distances sur X est muni d'une relation d'ordre partielle notée $\preceq$. On a $D \preceq D'$ ssi $\forall (x, y) \in X, D(x, y) \leq D'(x, y)$. Dans la seconde partie, nous développons un nouveau paradigme en classification : Etant donné une dissimilarité D, plutôt que de chercher à approximer D par une ultramétrique, nous construisons deux dissimilarités D_m et D_M qui encadrent D. On a donc $D_m \preceq D \preceq D_M$. Ces dissimilarités sont choisies de telle manière que cet intervalle de distances soit le plus petit possible et qu'il contienne au moins une ultramétrique U_s dite *sandwich* :

$$\exists U_s \text{ telle que } D_m \preceq U_s \preceq D_M.$$

C'est le dendrogramme de cette ultramétrique qui est considéré comme arbre de classification associé à D. Ceci revient à considérer chaque valeur $D(x, y)$ comme une estimation d'une valeur inconnue comprise entre $D_m(x, y)$ et $D_M(x, y)$. On notera que cette ultramétrique n'est généralement pas unique, même d'un point de vue topologique.

Par définition de l'ultramétrique sous-dominante, il existe une ultramétrique sandwich entre D_m et D_M si et seulement si $D_m \preceq U_{sd}(D_M)$. Etant donné une distance D, nous déterminons le plus petit $\beta \in \mathbb{R}_+$ tel que, l'intervalle $[(1 - \beta)D, (1 + \beta)D]$ contienne une ultramétrique.

Proposition

Soit D une distance sur X, un réel $\beta < 1$, $D_m = (1 - \beta)D$ et $D_M(1 + \beta)D$. La plus petite valeur β telle que $\exists U_s \in [D_m, D_M]$ est définie par :

$$\beta = || \frac{D - U_{sd}(D)}{D + U_{sd}(D)} ||_\infty.$$

Démonstration

Pour qu'il existe une ultramétrique encadrée par deux distances D_m et D_M, il faut et il suffit que l'ultramétrique sous dominante de D_M soit supérieure à D_m. On cherche donc β tel que :

$$\forall\, i,j \in X, (1-\beta)D(i,j) \leq U_{sd}((1+\beta)D(i,j))$$

$$\Leftrightarrow \forall\, i,j \in X, (1-\beta)D(i,j) \leq (1+\beta)U_{sd}(D)(i,j)$$

$$\Leftrightarrow \forall\, i,j \in X, \frac{D(i,j)-U_{sd}(D)(i,j)}{D(i,j)+U_{sd}(D)(i,j)} \leq \beta$$

Donc si on prend :

$$\beta = ||\frac{D-U_{sd}(D)}{D+U_{sd}(D)}||_\infty$$

la condition est réalisée. L'ultramétrique sandwich de référence est alors l'ultramétrique sous dominante de D_M, soit :

$$U_s^r(D) = U_{sd}(D_M) = (1+\beta)U_{sd}(D) \quad \bullet$$

De même que dans l'approximation au sens de la norme L_∞ de la première partie, cette valeur de β peut être la même pour toute paire d'éléments de X ou elle peut être ajustée, en fonction des sous-arbres du dendrogramme de $U_{sd}(D)$. On obtient ainsi l'ultramétrique sandwich notée $U_s(D)$.

Ici encore, la topologie de l'arbre obtenu est la même que celle du le lien unique (puisqu'elle est donnée par un arbre minimum). Mais, à l'intérieur du sandwich, il existe d'autres topologies, que nous nous proposons d'étudier.

Exemple

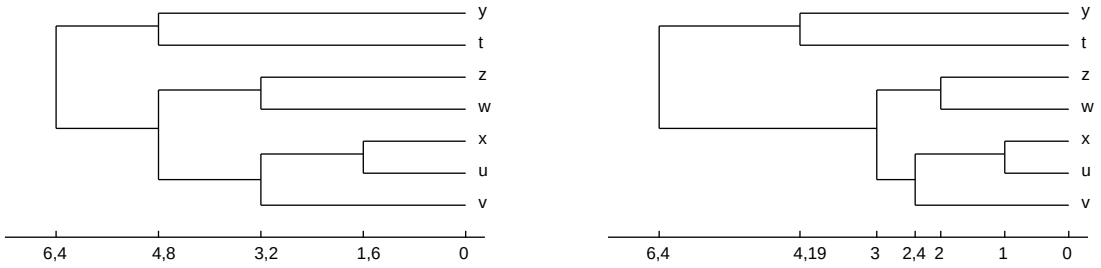


FIG. 2. *Dendrogrammes de l'ultramétrique sandwich de D, à gauche $U_s^r(D)$, à droite le cas ajusté $U_s(D)$.*

A partir d'une distance D, nous avons défini quatre ultramétriques ; deux qui sont optimales pour la norme infinie et deux qui sont comprises dans un intervalle minimum de distance qui contient D et une ultramétrique. Pour comparer ces approximations, nous avons repris des critères métriques classiques développés dans Guénoche et Garreta [2000]. Les valeurs affichées dans la table ci-dessous sont ici pour montrer que les ultramétriques ajustées sont meilleures que celles de référence. Ce n'est guère surprenant, compte tenu de la méthode. Il faut souligner que ce n'est pas au détriment de la complexité, puisque les ajustements créent une partition des valeurs de distance, et que chaque valeur n'intervient que dans un seul ajustement.

Dissimilarité D	U_{sd}	$\mathbb{U}_\infty$		$\mathbb{U}_s$	
		U^r_∞	U_∞	U^r_s	U_s
Moyenne des | différences |	2.333	2.735	2.03	2.333	2.069
Ecart quadratique moyen	8.333	11.604	7.248	8.333	7.195
Déviation standard	0.463	0.546	0.432	0.463	0.43
Stress	1.922	2.423	1.619	1.922	1.614

TAB. 2. *Comparaison des ultramétriques obtenues à partir de la dissimilarité de l'Exemple*

3. Application aux données quantitatives et aux variables d'intervalles

Dans la troisième partie, nous nous intéressons à des données dont les éléments sont décrits par des variables *d'intervalles* [Chavent et Lechevallier, 2002]. Plutôt que de calculer la distance de Hausdorff entre les éléments, nous évaluons un intervalle de distances à partir de ces données. Pour chaque $x \in X$ et pour chaque variable V^i, on connaît deux valeurs extrêmes $V^i_m(x)$ et $V^i_M(x)$ qui constituent un encadrement de la valeur hypothétique $V^i(x)$. Dans le cas où l'on n'aurait qu'une seule valeur $V^i(x)$, on peut toujours se ramener à l'aide d'un paramètre α_i, à l'intervalle $[(1-\alpha_i)V^i(x), (1+\alpha_i)V^i(x)]$ ou toute autre formule d'encadrement qui tienne compte de la distribution de V^i.

Pour ces variables d'intervalle, $D(x,y)$ peut être estimée par un intervalle $[D_m(x,y), D_M(x,y)]$. Les deux distances sont obtenues par sommation sur toutes les variables, éventuellement pondérées, des différences que l'on peut observer dans les intervalles $[V^i_m(x), V^i_M(x)]$ et $[V^i_m(y), V^i_M(y)]$. Il faut distinguer deux cas :

- Si les intervalles sont sans intersection, on admettra que $V^i_m(x) \leq V^i_M(x) < V^i_m(y) \leq V^i_M(y)$. On a alors $\delta^i_m(x,y) = V^i_m(y) - V^i_M(x)$ et $\delta^i_M(x,y) = V^i_M(y) - V^i_m(x)$.
- S'ils ont une intersection non vide, $\delta^i_m(x,y) = 0$. Pour $\delta^i_M(x,y)$, il faut distinguer le cas où les intervalles se chevauchent du cas où ils sont emboîtés ;
  - si $V^i_m(x) \leq V^i_m(y) \leq V^i_M(x) \leq V^i_M(y)$ on a $\delta^i_M(x,y) = V^i_M(y) - V^i_m(x)$ (cas du chevauchement),
  - si $V^i_m(y) \leq V^i_m(x) \leq V^i_M(x) \leq V^i_M(y)$ on a $\delta^i_M(x,y) = V^i_M(y) - V^i_m(y)$ (cas de l'emboîtement).

Les distances sont évaluées par les formules ci-dessous pour lesquelles on peut utiliser la norme de son choix :

$$D_m(x,y) = \|\delta^i_m(x,y)\| \text{ et } D_M(x,y) = \|\delta^i_M(x,y)\|.$$

On est alors ramené au problème de l'ultramétrique sandwich de la deuxième partie. S'il y a une solution, on cherche à construire une ultramétrique $\overline{U}$ centrée entre D_m et D_M, sinon, on calcule la moyenne $\overline{D} = \frac{1}{2}(D_m + D_M)$ et la solution de la première partie, $U_\infty(\overline{D})$. Ici encore, les solutions peuvent être comparées d'un point de vue métrique ou topologique

Références

M. Chavent, Y. Lechevallier (2002) Dynamical clustering of interval data : optimization of an adequacy criterion based on Hausdorff distance, in *Classification, Clustering and Data Analysis*, K. Jajuga et al. (Eds), Springer.

V. Chepoi, B. Fichet (2000) L-infinite approximation via subdominants, *J. of Mathematical Psychology*, 44, 600-616.

M. Farach, S. Kannan and T. Warnow (1995) A robust model for finding optimal evolutionary trees, *Algorithmica*, 13, 155-179.

A. Guénoche, H. Garreta (2001) Can we have confidence in a tree representation ? Proceedings of JOBIM'2000, *Lecture Notes in Computer Sciences*, vol. 2066, pp. 43-53.

M. Krivanek, J. Moravek (1986) NP-hard problems in hierarchical-tree clustering, *Acta Informatica*, 23, 311-323.

Régles d'association et interaction entre variables binaires

Martine Cadot — **Amedeo Napoli**

LORIA, UMR 7503 CNRS
BP 239 - 54506 - Vandœuvre-lès-Nancy Cedex
martine.cadot@loria.fr ; amedeo.napoli@loria.fr

RÉSUMÉ. *Les analyses multidimensionnelles permettent l'étude de plusieurs variables en tenant compte de leurs liaisons. Nous nous intéressons ici à un type de liaison particulier, qui est l'interaction statistique. A travers le modèle log-linéaire, nous exposons ce qu'est l'interaction, et les précautions d'interprétation des liens entre variables qui en découlent. Puis nous examinons le formalisme d'extraction automatique de règles d'association, afin de voir si les problèmes d'interprétation liés à l'existence d'interactions se retrouvent, et nécessitent alors une correction du jeu de règles extrait.*

MOTS-CLÉS : *Classification de règles, fouille de données, modèle loglinéaire, règles d'association, interaction*

1. Introduction

Une des techniques de la fouille de données consiste à extraire un jeu de règles d'association d'une base de données, le plus souvent sous la forme d'une matrice sujets*propriétés[6]. Puis ces règles d'association sont classées selon divers indices[7], avant d'être rendues aux experts du domaine d'où proviennent les données, pour interprétation[4]. Pour éviter l'explosion du nombre de règles, des hypothèses sont souvent faites sur les interactions entre variables, par exemple d'independance locale dans les réseaux bayésiens [9]. De telles hypothèses ne sont pas faites lors de l'extraction du jeu de règles d'association. Nous nous proposons ici d'explorer les relations entre les éléments d'un jeu de règles d'association qui seraient induites par ces interactions. Nous les définissons en utilisant le formalisme du modèle log-linéaire des statistiques inférentielles[8]. Dans un premier temps, nous exposons sur un modèle à deux variables ce qu'est l'interaction et l'effet qu'elle produit sur le jeu des 8 règles d'association correspondantes, puis nous l'étendons à trois variables.

2. Avec 2 variables

On a 2 proprietes A et B, et N sujets numérotés de 1 à N pour lesquels on connait l'absence (0) ou la présence (1) de chacune de ces 2 propriétés. On dispose d'une matrice booléenne de N lignes et 2 colonnes, c'est-à-dire

Numéro de sujet	A	B
1	0	0
2	1	0
...	...	...

AxB	B=0	B=1	total
A=0	a	b	a+b
A=1	c	d	c+d
total	a+c	b+d	N=a+b+c+d

TAB. 1. *matrice booléenne de données tableau de contingence des 2 variables observées sur N sujets*

contenant uniquement les valeurs 0 et 1. Par exemple, si A est le sexe masculin, B est la réussite à l'examen, et les

sujets sont des étudiants, on voit dans le tableau de gauche de la table 1 que l'étudiant 1 est une fille (A=0) qui a échoué (B=0), que l'étudiant 2 est un garçon (A=1) qui a échoué, etc... Le tableau de contingence permettant de résumer cette matrice de données est le tableau de droite de la table 1, de 2 lignes et 2 colonnes, où figurent les effectifs des 4 cas possibles de valeurs du couple de variables (A,B), auquel on a rajouté une colonne et une ligne de totaux, qui sont appelées les marges du tableau. On notera les modalités 0 et 1 de la variable A par A0 et A1, et pareillement pour B. On se limitera aux tableaux de contingence n'ayant aucune marge nulle, c'est-à-dire que les variables A et B ont réellement deux modalités chacune.

2.1. *Le modele log-linéaire*

Si on note n_{ij} le nombre de sujets pour lesquels on a $A = i$ et $B = j$, où $i, j \in \{0; 1\}$, et si aucun de ces 4 effectifs n'est nul, on peut écrire l'équation suivante :

$$ln(n_{ij}) = \alpha + \beta i + \gamma j + \delta ij \qquad avec\, \alpha = ln(a),\ \beta = ln(c/a),\ \gamma = ln(b/a),\ \delta = ln(ad/bc) \qquad [1]$$

Dans le cas où $\delta = 0$, on dira qu'il n'y a pas d'interaction entre A et B, ce qui correspond au cas où $a/b = c/d = (a+c)/(b+d)$, c'est-à-dire où les colonnes sont proportionnelles, et où $a/c = b/d = (a+b)/(c+d)$, c'est-à-dire où les lignes sont proportionnelles[1]. Dans ce cas d'interaction nulle, on s'interessera aux valeurs de β et de γ, qu'on appelera respectivement l'effet de A, et l'effet de B. Si β est nul, cela signifie que c=a et b=d, donc que les lignes A0 et A1 sont identiques, donc que l'effet de A est nul, s'il est positif, que la ligne de A1 l'emporte sur celle de A0, et inversement s'il est négatif. De la même façon, on examine la valeur de γ en cas d'interaction nulle, pour tirer des conclusions sur l'effet de B. Quand l'interaction n'est pas nulle, les coefficients β et γ ne sont plus alors interprétables[2]. En effet, Pour le tableau de droite de la table 2, si A=0, ce qui correspond à i=0,

1 effet principal			
AxB	B=0	B=1	total
A=0	40	460	500
A=1	40	460	500
total	80	920	1000
$ln(n_{ij}) = 3.69 + 2.44j$			

2 effets principaux			
AxB	B=0	B=1	total
A=0	60	540	600
A=1	40	360	400
total	100	900	1000
$ln(n_{ij}) = 4.09 - 0.41i + 2.20j$			

interaction			
AxB	B=0	B=1	total
A=0	790	110	900
A=1	10	90	100
total	800	200	1000
$ln(n_{ij}) = 6.7 - 4.4i - 2.0j + 4.2ij$			

TAB. 2. *tableau de contingence de 2 variables A et B sans interaction*

l'equation devient $ln(n_{ij}) = 6.7 - 2.0j$, alors que si A=1, elle devient $ln(n_{ij}) = (6.7 - 4.4) + (-2.0 + 4.2)j$ soit $ln(n_{ij}) = 2.3 + 2.2j$. Par contre, si on examine l'effet de A pour B=0 et B=1, on n'a pas de contradiction avec le coefficient β. Dans le cas où B=0, $ln(n_{ij}) = 6.7 - 4.4i$, et dans le cas où B=1, $ln(n_{ij}) = 4.7 - 0.2i$. On obtient dans les deux cas A0 qui l'emporte sur A1.

2.2. *Les règles d'association*

En logique formelle, on dit qu'on a la règle "A0 implique B0", que l'on note A0→B0, quand il n'y a aucun élément vérifiant simultanément A0 et B1 (le nombre b de la table 1 est nul). Par exemple, si A est le sexe masculin, B est la réussite à l'examen, la règle A0→B0 se traduit par "être une fille implique échouer à l'examen", et on la prouve en établissant que le nombre de personnes qui sont à la fois de sexe feminin et qui ont réussi à l'examen, est nul. Les règles d'association de la fouille de données sont des implications approchées, c'est-à-dire qu'on accepte que b soit différent de 0. Leur qualité est mesurée par de nombreux indices [7].

1. On parle également d'indépendance entre les 2 variables A et B
2. Les exemples cités dans cet article ont été choisis tels que les interactions non nulles diffèrent significativement de zéro afin de permettre au lecteur non statisticien de suivre l'exposé sans faire d'inférences statistiques, et au lecteur statisticien d'accepter de le suivre en les ayant faites.

On dira que la règle A0→B0 est de meilleure qualité que la règle A0→B1 pour un indice donné "ind" si l'expression $E(ind) = ind(A0 \rightarrow B0) - ind(A0 \rightarrow B1)$ est positive. Nous avons établi (voir preuves dans[3]) que les indices se divisent en 2 groupes. Pour le premier groupe, qui contient le support, la fréquence, la confiance et l'étonnement, cette expression est de même signe que $D = a - b$, donc que l'effet de B (en fait de B0 par rapport à B1 noté $\gamma' = ln(a/b)$), et pour le second, qui contient la différence, l'interêt, la satisfaction, la nouveauté, la conviction et l'implication[3], cette expression est de même signe que $P = ad - bc$, donc que l'interaction $\delta = ln(ad/bc)$. Quand les indices donnent des informations contradictoires, on peut choisir de rejeter les 2 règles correspondantes, ou bien de garder celles maximisant la valeur d'un indice spécifique à l'expert des données. Une autre façon d'agir assez courante est d'imposer un seuil à la confiance, de façon arbitraire ou sur des bases statistiques [2], puis regarder les valeurs les plus élevées de certains autres indices [4]. Si le seuil de confiance dépasse 0.5, comme c'est généralement le cas quand il est choisi arbitrairement (le plus souvent 0.8), la seule des deux règles qui dépasse ce seuil est gardée. Si ce seuil est plus bas, les deux règles peuvent le dépasser, et dans ce cas les autres indices peuvent intervenir dans le choix de la meilleure.

Pour le premier tableau de la table 2, comme il n'y a pas d'interaction, les indices du second groupe sont identiques. Comme il n'y a pas d'effet principal pour A, les indices sont les mêmes pour les règles 5 et 6 ainsi que pour les règles 7 et 8, on ne peut pas faire de choix, on les rejette toutes les quatre. Comme B a un effet principal, on peut faire un choix entre la règle 1 et la règle 2, en choisissant la règle 2, et entre la règle 3 et la règle 4, en choisissant la règle 3. Ce tableau nous fournit donc un ensemble de 2 règles, qui sont la règle (A=0)→(B=1) et la règle (A=1)→(B=1). Ces deux règles ont tous leurs indices identiques. Elles ne sont pas contradictoires, mais l'information qu'elles apportent à elles deux n'a guère de valeur, car l'intervention de A ne change pas le rapport des chances entre avoir B=0 et B=1. On les supprime donc et le jeu de règles est vide. Pour le deuxième tableau de la table 2, on peut se retrouver avec une seule règle, la règle 2, deux règles, la règle 2 et sa réciproque la règle 8, ou aucune règle, selon qu'une petite différence d'indices est négligée ou non. Le tableau de droite de la table 2 a

No	Partie gauche		Partie droite	Sup port	Confi ance	Etonn ement
1	A0	→	B0	40	0.08	-5.25
2	A0	→	B1	460	0.92	0.46
3	A1	→	B1	460	0.92	0.46
4	A1	→	B0	40	0.08	-5.25

No	Partie gauche		Partie droite	Sup port	Confi ance	Etonn ement
5	B0	→	A0	40	0.50	0
6	B0	→	A1	40	0.50	0
7	B1	→	A1	460	0.50	0
8	B1	→	A0	460	0.50	0

TAB. 3. *les indices des règles pour A et B sans interaction, et un effet principal de B*

une interaction qui contrarie l'effet principal de B, la règle 3 l'emporte maintenant sur la règle 4, alors que dans le couple de règles (1, 2), c'est la règle 1 qui l'emporte. Pour les 4 règles suivantes, l'interaction n'ayant pas contrarié l'effet de A, Les indices du premier groupe privilégient les règles 5 et 8. Toutefois, dans la colonne correspondant à B=0, la différence entre A0 et A1 a été accentuée alors que dans la colonne correspondant à B=1, elle a été diminuée. Ce sont les indices du second groupe qui rendent compte de cet effet, et qui privilégient la règle 5 à la règle 8. Pour conclure sur cette interaction contrariante, on peut maintenant extraire un jeu de règles qui a plus de sens, et qui contient la règle 1, la règle 3 et la règle 5.

3. Avec 3 variables

Pour le modèle log-linéaire, on note n_{ijk} le nombre de sujets pour lesquels on a $A = i$, $B = j$ et $C = k$, où $i, j, k \in \{0; 1\}$, et si aucun de ces effectifs n'est nul, on peut écrire l'équation suivante.

$$ln(n_{ijk}) = \alpha + \beta_1 i + \beta_2 j + \beta_3 k + \gamma_1 ij + \gamma_2 ik + \gamma_3 jk + \delta ijk \qquad [2]$$

avec $\alpha = ln(n_{000})$, $\beta_1 = ln(\frac{n_{100}}{n_{000}})$, $\beta_2 = ln(\frac{n_{010}}{n_{000}})$, $\beta_3 = ln(\frac{n_{001}}{n_{000}})$, $\gamma_1 = ln(\frac{n_{000}n_{110}}{n_{010}n_{100}})$, $\gamma_2 = ln(\frac{n_{000}n_{101}}{n_{100}n_{001}})$, $\gamma_3 = ln(\frac{n_{000}n_{011}}{n_{001}n_{010}})$, $\delta = ln(\frac{n_{001}n_{100}n_{010}n_{111}}{n_{000}n_{011}n_{101}n_{110}})$.

3. Pour l'implication, il faut toutefois que la valeur de $\frac{ad-bc}{N}$ soit assez grande en valeur absolue

Comme dans l'équation 1, on a l'effet global, α, l'effet de chaque variable prise séparément, $\beta 1$ pour A, $\beta 2$ pour B, $\beta 3$ pour C, les interactions des variables prises 2 à deux, γ_1 pour l'interaction A*B, γ_2 pour A*C, γ_3 pour B*C, mais on a en plus δ qui correspond à l'interaction A*B*C entre les 3 variables. Pour interpréter l'interaction de 3 variables exprimée par le coefficient δ, examinons le cas ou elle est nulle, cela donne $\frac{n_{000}n_{110}}{n_{100}n_{010}} = \frac{n_{001}n_{111}}{n_{101}n_{011}}$, ce qui signifie que l'écart à la proportionnalité du tableau AxB pour C=0 est le même que celui du tableau AxB pour la valeur 1 de C[4]. Nous montrons que si l'interprétation des relations dans le cas d'une interaction non nulle entre 3 variables est difficile, la plus grande simplicité en cas d'interaction A*B*C nulle n'empeche pas que l'apparition d'une troisieme variable peut contredire la relation trouvée sur 2 variables.

Pour les règles d'association, il en découle que peuvent apparaitre des contradictions sous la forme d'un jeu de règles contenant par exemple les 3 règles A0 C0→B0, A0 C1→B1, A0→B0. Si on imagine que A0 est "être de sexe féminin", B0 est "échouer à l'examen", et C0 est "ne pas être redoublant", on voit que ces règles se traduisent respectivement ainsi : "les non redoublantes échouent à l'examen", "les redoublantes réussissent l'examen", "les filles échouent à l'examen". La troisième règle contredit la deuxième règle. On préfère en général l'éliminer afin de garder un jeu de règles cohérent sans perdre trop d'information. Nous l'illustrons par un exemple avant de l'établir par des calculs mathématiques.

4. Conclusion et perspectives

En nous aidant du formalisme du modèle log-linéaire, nous avons essayé de comprendre les relations entre plusieurs variables binaires et leurs conséquences sur le jeu de règles d'association. Dans le modèle à deux variables, nous avons remarque la piètre qualité du jeu de règles trouvé en absence d'interaction, ou en cas d'interaction non "contrariante". Ce que nous avons retrouvé sous forme de redondance dans le modèle à 3 variables. Dans ce dernier modèle sont également apparues des contradictions dues à une règle générale A0→B0 obtenue par agrégation de 2 règles partielles contradictoires A0 C0→B0 et A0 C1→B1. Il semble intéressant de nettoyer le jeu de règles de ces contradictions et de ces redondances avant de le donner à interpreter aux experts du domaine dont sont issues les données.

5. Bibliographie

Bastide R Y., Taouil R., Pasquier N., Stumme G., Lakhal L., " Pascal : un algorithme d'extraction des motifs fréquents" , *Technique et science informatiques*, 21(1), 2002, p. 65-75.

Cadot M., Napoli A., 2003, " Une optimisation de l'extraction d'un jeu de règles s'appuyant sur les caractéristiques statistiques des données", *RSTI-RIA-ECA* - 16 /2003.

Cadot M., Napoli A., 2003, Association rules and "Simpson's Paradox", communication soumise à *JIM03*, Metz septembre 2003

Cherfi H. et Toussaint Y., " Adéquation d'indices statistiques à l'interprétation de règles d'association ", Actes des 6èmes Journées internationales d'Analyse statistique des Données Textuelles, *JADT'02*, Saint-Malo, IRISA-INRIA Vol. 1, p. 233-244, 2002.

Gras R. et collaborateurs, *L'implication statistique, une nouvelle méthode exploratoire de données*, La pensée sauvage, Grenoble, 1996.

Han J. and Kamber M., Data Mining : Concepts and Techniques, Morgan Kaufmann Publishers, San Francisco, 2001.

Kodratoff Y., " Rating the Interest of Rules Induced from Data and within texts ", *12th IEEE -International Conference on Database and Expert Systems Aplications-Dexa 2001*, Munich, sept 2001.

Morineau, A., Nakache, J.-P., Krzyzanowski, C. *Le modèle log-linéaire et ses applications*, Cisia-Ceresta, Paris 1996.

Whittaker, J. *Graphical models in applied multivariate Statistics*, John Wiley, 1990.

4. cela signifie également l'égalité des écarts à la proportion pour les tableaux AxC pour B=0 et B=1, et pour les tableaux BxC pour A=0 et A=1

Effet Guttman : son interprétation et une nouvelle méthode de redressement

Sergio Camiz[1], Giorgia Polenta

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

polenta@mat.uniroma1.it

RÉSUMÉ. On complète l'interprétation courante de l'effet Guttman par l'explication de l'épaisseur de l'arc. Ceci suggère une méthode de redressement utilisant des coordonnées polaires, qui permet de garder toute l'information contenue dans le premier plan factoriel.

MOTS-CLÉS. Analyse des correspondances, Effet Guttman, Redressement, Coordonnées polaires.

1. Introduction

L'effet Guttman se présente lors de l'application d'une analyse factorielle à une matrice en bandes diagonales. Selon les différentes techniques utilisées, sur le premier plan factoriel on peut trouver une distribution ayant la forme d'un fer à cheval (analyse en composantes principales) ou d'un arc (analyse des correspondances). Dans les plans factoriels suivants, on trouve des distributions autour de courbes plus complexes, ce qui en principe semble empêcher toute interprétation ; d'habitude on en déduit que les autres facteurs sont des transformations du premier, donc sans intérêt. Dans les applications, on trouve cet effet lorsque les données représentent un échantillon dont la diversité dépend de facteurs (non observés) selon lesquels les caractères observés assument des valeur non nulles seulement dans un intervalle, voire ont une distribution unimodale. C'est le cas des données saisies dans le cadre des sciences naturelles, car la *niche écologique*, à savoir l'environnement favorable au développement des individus d'une espèce, est normalement borné à l'intervalle des facteurs écologiques où leur vie est possible. Dans cet intervalle, l'existence de conditions optimales justifie l'assomption d'une distribution unimodale, qui par simplicité est normalement modélisée par une loi gaussienne, bien que la symétrie ne soit pas justifiée.

[1] Pour ce travail, il a été financé par le C.N.R., contrat n° CNRC00D101_001, et par l'Università La Sapienza.

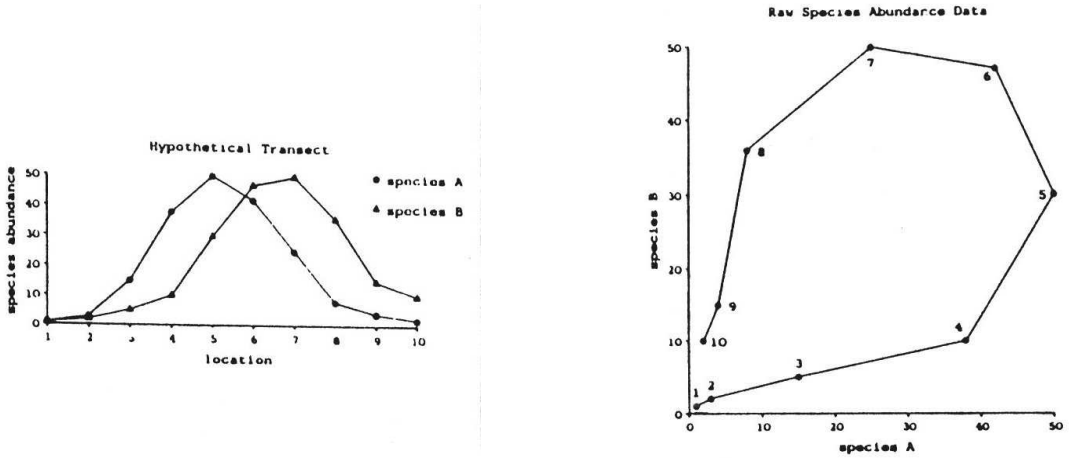


Figure 1 - Deux caractères décalés le long d'un facteur et leur représentation grafique.

L'interprétation de l'effet Guttman comme présence d'un facteur écologique sous-jacent à la distribution des espèces et des relevés dans l'étude des communautés végétales, a beaucoup intéressé les botanistes, qui ont été frappés par la possibilité de *l'analyse indirecte des gradients*, à savoir la possibilité de remonter de la composition florale des relevés à leur position le long du gradient écologique ainsi identifié. Ce but a conduit les chercheurs à des études spécifiques, suivant différentes directions : *i)* la justification de cet effet ; *ii)* la recherche de la méthode qui le minimise ; *iii)* la recherche d'une technique de redressement; *iv)* le développement de méthodes alternatives de traitement, basées sur la distribution gaussienne.

Dans toutes ces études on a utilisé, comme données simulées, des matrices à bande régulière, qui soumises aux analyses en composantes principales (*ACP*) produisent un fer à cheval, et soumises aux analyses des correspondances (*AFC*) produisent un arc. Ceci a conduit à qualifier cet effet de généralisation multidimensionnelle de ce qu'on observe dans la représentation graphique simultanée de deux variables gaussiennes décalées (Figure 1), une distorsion à éliminer sans regret. Par conséquent, Gauch et Hill [GAU 80] ont proposé un redressement soit par soustraction des coordonnées des objets sur le second facteur de leur moyenne, calculée par intervalles du premier facteur (Figure 2), soit par soustraction de la moyenne estimée par une courbe de régression de second degré. Dans les deux cas, le redressement cause une forte distorsion des distances entre objets contigus, à cause de la variation arbitraire des coordonnées sur le second facteur. De telles techniques ne fournissent pas de grands avantages, car ce redressement ne respecte pas l'inflexion des objets disposés aux extrêmes du fer à cheval. Comme autre essai, les techniques basées sur un modèle gaussien ([IHM 75], [GOO 80]) se bornent à ran-

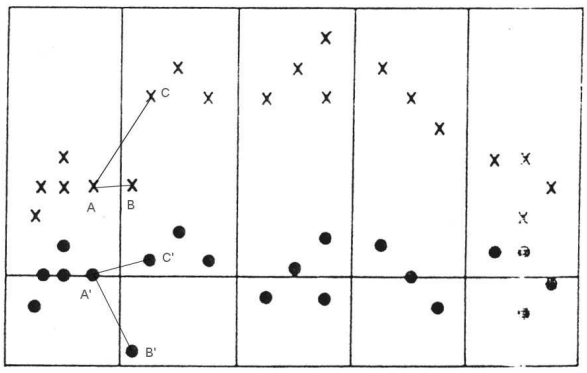


Figure 2 - Redressement selon Gauch et Hill.

ger correctement les objets le long d'un gradient unidimensionnel, au prix de la perte de toute autre information contenue dans les données.

Dans ce travail, on propose une interprétation de l'effet Guttman basée sur des matrices ayant plusieurs bandes de différente largeur, ce qui permet d'expliquer l'*épaisseur* de l'arc comme *étendue* de la distribution des caractères et on présente une méthode de redressement des facteurs qui sauvegarde toute l'information contenue dans le premier plan factoriel.

2. L'effet Guttman : son interprétation

L'interprétation de l'effet Guttman est très simple et se base sur la technique courante d'interprétation des analyses factorielles : si on suppose de disposer d'une matrice en bande, il est normal que les caractères ayant une valeur dans les unités extrêmes soient opposées sur le premier axe factoriel. Si l'étendue des caractères est limitée, il arrive que les unités centrales soient aussi opposées aux deux extrêmes. Les trois ensembles se situeront alors, dans le premier plan factoriel, aux sommets d'un *triangle*. La disposition des unités intermédiaires entre centre et extrêmes sera le long d'un arc, mais l'opposition entre elles sera représentée par l'opposition sur un troisième axe, orthogonal aux deux précédents, et ainsi de suite. Des considérations analogues peuvent se faire pour les caractères. Pour ce qui concerne la position des objets dans l'épaisseur de l'arc ainsi obtenu, il faut distinguer entre *ACP*, où les caractères ayant une étendue plus courte sont ramenés vers l'origine des axes, car leur liaison avec les autres caractères est plus faible, donc mal représentée, et *AFC,* où la relation barycentrique impose, au contraire, que les caractères moins corrélés soient situés vers l'extérieur de la distribution. Par conséquent, les unités pauvres en caractères seront ramenées vers le centre par l'*ACP* et loin du centre par l'*AFC* ; le contraire arrivera pour les unités riches en caractères. Il faut encore remarquer qu'en *AFC* la disposition de caractères de l'autre côté de l'arc par rapport à l'origine concerne des caractères présents dans des unités situées aux deux extrêmes de la distribution et que la réduction de l'intensité de la liaison entre les caractères aux extrêmes de la distribution justifie l'inflexion en *ACP* et l'ouverture et rétrécissement de l'arc en *AFC.* Toutes ces considérations mènent à la conclusion que, en *AFC,* la distribution le long de l'arc est bornée par une enveloppe convexe représentée par une bande de caractères présents dans une seule unité et par une bande d'unités n'ayant qu'un caractère présent.

3. Le redressement proposé

Étant donné que, dans les conditions qu'on a décrites, les objets représentés sur le plan factoriel de l'*AFC* se disposent selon un ruban en forme d'arc qui contourne l'origine, on propose de considérer les coordonnées polaires des points dans le plan par rapport à l'origine (Figure 3). Ainsi, la position des points le long du ruban est représentée par un angle

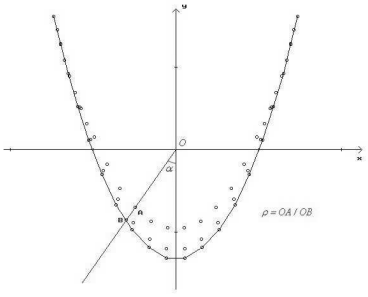


Figure 3 – Coordonnées polaires.

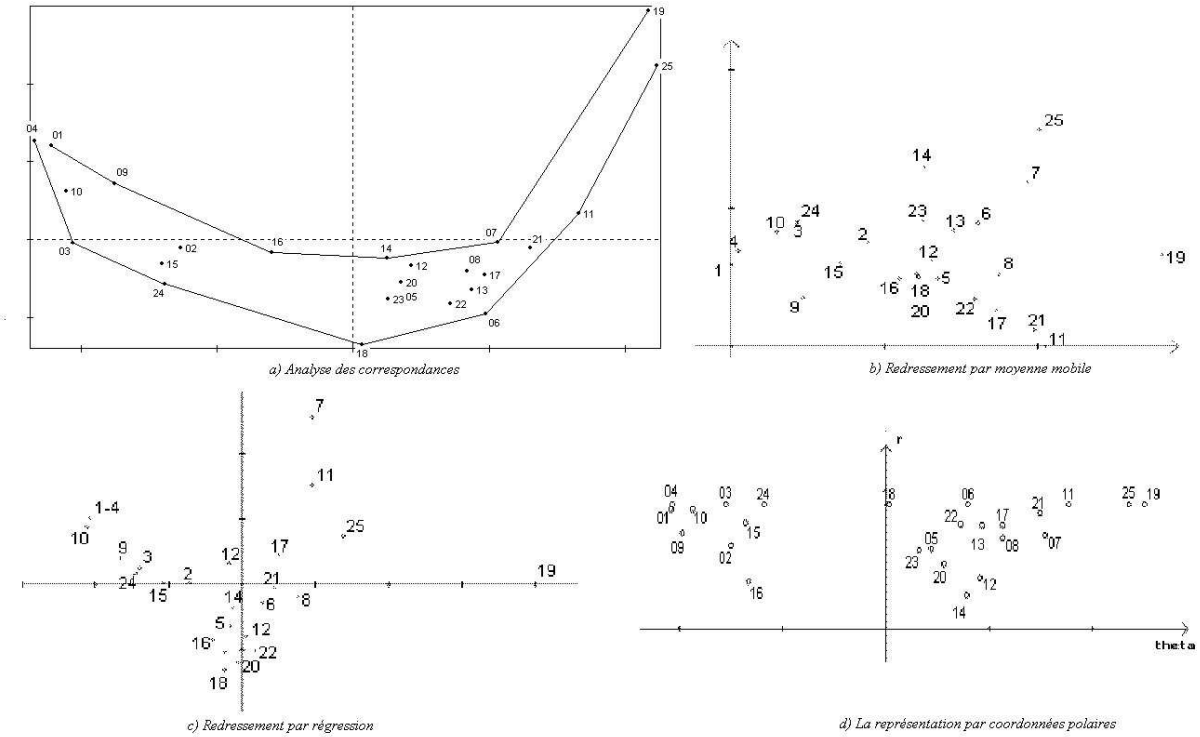


Figure 4 - Premier plan de l'AFC des données d'Ellenberg et ses redressements.

et leur position dans l'épaisseur du ruban par leur distance au centre. Comme on a vu que l'arc est en forme de parabole, on propose d'ajuster les distances des points en calculant leur rapport par la distance au centre de l'enveloppe convexe calculée sur le même rayon. Utilisant angles et rapports comme deux coordonnées cartésiennes on parvient à une représentation dans le plan où les points se disposent le long de l'axe horizontal selon le facteur curvilinéaire et le long du vertical selon leur étendue le long de ce facteur.

Dans la Figure 4 on voit une comparaison des méthodes sur l'*AFC* des données de végétation d'Ellenberg [MUL 74] : la transformation des coordonnées polaires (4*d*) s'avère beaucoup plus respectueuse des relations entre points que les redressements de [GAU 80] (4*b,* 4*c*).

4. Bibliographie

[GAU 80] GAUCH H. G., HILL M. O., « Detrended Correspondence Analysis: an Improved Ordination Technique ». *Vegetatio*, vol. 42, 1980, pp. 47-58.

[GOO 80] GOODALL D. W., JOHNSON R.W., « A Maximum Likelihood Approach to Non-linear Ordination ». *Vegetatio*, vol. 41, 1980, n. 3, pp. 133-142.

[IHM 75] IHM P., VAN GROENEWOUD H., « A Multivariate Ordering of Vegetation Data Based on Gaussian Type Gradient Response Curve ». *Journal of Ecology*, vol. 63, 1975, pp. 767-777.

[MUL 74] MÜLLER-DOMBOIS D., ELLENBERG E., *Aims and Methods of Vegetation Ecology*. 1974, NewYork, J.Wiley & Sons.

Classification automatique des données de type intervalle basée sur une distance de Hausdorff adaptative

Renata M.C.R. de Souza[1], Francisco de A.T. de Carvalho[1] et Yves Lechevallier[2]

[1]Centro de Informática – CIn/UFPE, Av. Prof. Luiz Freire, s/n – Cidade Universitária, CEP :50740-540 Recife-PE, Brésil, {rmcrs,fatc}@cin.ufpe.br
[2]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

RÉSUMÉ. Dans ce travail nous présentons une méthode de classification automatique adaptative utilisant la distance de Hausdorff et qui s'appliquera aux tableaux contenant des intervalles.

MOTS-CLÉS : Analyse des Données Symboliques, Nuées Dynamiques, Distances Adaptatives, Distance de Hausdorff, vecteur d'intervalles.

1. Introduction

L'algorithme de Nuées Dynamiques [DID 71] cherche a obtenir simultanément, une partition d'un ensemble d'individus en *k* classes et un ensemble de *k* prototypes, en optimisant un critère mesurant l'adéquation entre chacune des classes de cette partition et son prototype. Dans une version adaptative de cet algorithme [GOV 75], à chaque itération la mesure, comparant la classe avec sa représentation, est différente. Son principal avantage est sa capacité de trouver des classes de formes et de tailles différentes. L'analyse des données symboliques [BOCK 00] est un nouveau domaine de l'extraction des connaissances qui fourni des extensions appropriées aux méthodes usuelles de classification automatique pour des données de type intervalle. En particulier, dans [CHAV 02] les auteurs ont introduit une méthode de classification utilisant la distance de Hausdorff dans le cadre des données de type intervalles. Ce travail introduit une version adaptative de cette méthode. Plusieurs jeux de données simulées contenant des intervalles nous permettent de comparer ces diverses approches.

2. Une méthode de classification avec une distance fixée

Soit $\Omega = \{\omega_1,\ldots, \omega_n\}$ un ensemble de *n* individus décrits par *p* variables de type intervalle. Une *variable intervalle Y* est une correspondance de l'ensemble Ω des objets dans $\Re$ qui vérifie la propriété suivante sur son graphe : pour tout individu $s \in \Omega$ le sous-ensemble

$[a,b]=Y(s)$ est un intervalle fermé de $\Re$. On notera $\Im$ l'ensemble des intervalles fermés de $\Re$. Chaque individu ω_s de Ω est représenté par un vecteur d'intervalles $\mathbf{x}_s = (x_s^1, \ldots, x_s^p)$, où $x_s^j = [a_s^j, b_s^j] \in \Im$. Le tableau de données est constitué de *n* lignes (les individus à classer) et de *p* colonnes (les variables), chaque case de ce tableau contenant un intervalle fermé de $\Re$.
Le problème de classification est de trouver un couple constitué d'une *partition P** de Ω en *k* classes non vides et d'un *vecteur L** de *k* prototypes $(G_1,\ldots,G_i,\ldots,G_k)$ qui vérifie :

$$\Delta(P^*, L^*) = Min\left\{\Delta(P,L) \,\middle|\, P \in P_k, L \in \Lambda^k\right\}$$

où P_k est *l'ensemble des partitions* de Ω en *k* classes non vides et où Λ est *l'espace de représentation des prototypes*. Ce critère Δ est défini par :

$$\Delta(P,L) = \sum_{i=1}^{k} D(C_i, G_i) = \sum_{i=1}^{k} \sum_{s \in C_i} D(\{\omega_s\}, G_i) \quad C_i \in P, G_i \in \Lambda$$

avec D comme mesure de proximité entre une partie de Ω et un élément de Λ. L'algorithme d'optimisation utilisé est de type Nuées Dynamiques et il procède alternativement par une étape de *représentation* suivie d'une étape d'***allocation.***

a) *étape de représentation* : (la partition *P* est fixée)
Pour i=1 à *k*, on recherche le prototype G_i de Λ minimisant $D(C_i, G_i)$.

b) *étape d'allocation :* (le vecteur des prototypes est fixé)
Pour s=1 à *n*, on recherche pour l'individu ω_s sa nouvelle classe C_l d'affectation, avec $l = \arg\min_{i=1,\ldots,k} D(\{\omega_s\}, G_i)$

c) si aucun individu change de classe d'affectation alors on arrête le processus d'optimisation autrement on recommence les étapes ***a*** et ***b***.

Comme le critère Δ est additif en fonction des *k* classes et des *n* individus de Ω, la recherche de la classe d'affectation *l* de l'objet *s* dépend uniquement de la fonction de comparaison D. Ainsi la décroissance du critère Δ est obtenue sous les conditions suivantes :

- unicité du choix de la classe d'affectation pour chaque individu de Ω;
- unicité du prototype *G* minimisant $D(C,G)$ pour toute classe *C* de Ω.

Dans [CHAV 02] l'espace Λ est l'ensemble $\Im$ des intervalles fermés de $\Re$. Donc la mesure de proximité D entre un ensemble *C* de Ω et un élément de Λ est une somme des distances entre deux vecteurs d'intervalles. Cette distance est égale à $d(\mathbf{x}_s, \mathbf{y}_i) = \sum_{j=1}^{p} d_H(x_s^j, y_i^j)$ avec d_H, la distance de Hausdorff entre les intervalles $x_s^j = [a_s^j, b_s^j]$ et $y_i^j = [\alpha_i^j, \beta_i^j]$, qui est égale à : $d_H(x_i^j, y_k^j) = \max\left\{|a_i^j - \alpha_k^j|, |b_i^j - \beta_k^j|\right\}$

La distance étant maintenant choisie, le problème est trouver le prototype G_i de la classe C_i et sa représentation $\mathbf{y}_k$ qui minimise la mesure de proximité $D(C_i, G_i)$. Selon [CHAV 02], la solution est $\alpha_i^j = \mu_i^j - \theta_i^j$ et $\beta_i^j = \mu_i^j + \theta_i^j$, $j = 1, \ldots, p$ et $i = 1, \ldots, k$, où μ_i^j est la médiane de

l'ensemble des milieux des intervalles $x_s^j = [a_s^j, b_s^j]$, $s \in C_k$, et θ_i^j est la médiane de l'ensemble des demi-longueur des intervalles x_s^j.

3. Une méthode de classification avec une distance adaptative

Dans le cas précédent la distance entre deux intervalles est définie de manière unique. Dans ce paragraphe nous proposons d'utiliser une distance adaptée à chacune des classes de la partition recherchée, cette approche se trouve dans une généralisation de l'algorithme des distances adaptatives décrite dans [GOV 75]. Dans ce cas le critère Δ est maintenant défini par :

$$\Delta(P, L, d) = \sum_{i=1}^{k} \sum_{s \in C_i} d_i(\mathbf{x}_s, \mathbf{y}_i)$$

La valeur $d_i(\mathbf{x}_s,\mathbf{y}_i)$ mesure la distance entre le vecteur d'intervalles $\mathbf{x}_s$ (l'objet ω_s étant mis dans la classe C_i) et le vecteur d'intervalles $\mathbf{y}_i$ ($\mathbf{y}_i$ étant la représentation du prototype de la classe C_i). Les distances adaptatives d_i sont paramétrées par le vecteur $\lambda_i = (\lambda_i^1,...,\lambda_i^p)$ de la façon suivante :

$$d_i(\mathbf{x}_s, \mathbf{y}_i) = \sum_{j=1}^{p} \lambda_i^j \; d_H(x_s^j, y_i^j) = \sum_{j=1}^{p} \lambda_i^j \; \max\left\{ |a_s^j - \alpha_i^j|, |b_s^j - \beta_i^j| \right\} \qquad \text{avec } \lambda_i^j > 0 \text{ et } \prod_{j=1}^{p} \lambda_i^j = 1.$$

Pour résoudre ce problème, [GOV 75] propose de décomposer l'étape de représentation en deux nouvelles étapes :

a1) (la partition P et le vecteur des distances d_i sont fixés)

Pour i=1 à k, on recherche la représentation $\mathbf{y}_i$ du prototype G_i de Λ minimisant :

$$\sum_{s \in C_i} d_i(\mathbf{x}_s, \mathbf{y}_i) = \sum_{s \in C_i} \sum_{j=1}^{p} \lambda_i^j \; d_H(x_s^j, y_i^j) = \sum_{j=1}^{p} \lambda_i^j \sum_{s \in C_i} d_H(x_s^j, y_i^j)$$

La solution est identique à la solution de l'étape de représentation de la méthode précédente.

a2) (la partition P et le vecteur des prototypes sont fixés)

Pour i=1 à k, on recherche le vecteur de pondérations $\lambda_i = (\lambda_i^1,...,\lambda_i^p)$ minimisant :

$\sum_{j=1}^{p} \lambda_i^j \sum_{s \in C_i} d_H(x_s^j, y_i^j)$. Comme, décrit dans [GOV 75] (page 57), ces paramètres sont calculés par la méthode des multiplicateurs de Lagrange et on obtient :

$$\lambda_i^j = \frac{\left[\prod_{h=1}^{p} \left(\sum_{s \in C_i} \max\left\{ |a_s^h - \alpha_i^h|, |b_s^h - \beta_i^h| \right\} \right) \right]^{\frac{1}{p}}}{\sum_{s \in C_i} \max\left\{ |a_s^j - \alpha_i^j|, |b_s^j - \beta_i^j| \right\}}, \; j = 1, \ldots, p \text{ et } i = 1, \ldots, k$$

Comme dans la méthode précédente, les étapes de représentation et d'allocation sont répétées itérativement jusqu'à la convergence du critère d'adéquation. L'initialisation, l'étape d'allocation et le critère d'arrêt sont identiques. La différence entre ces deux méthodes apparaît uniquement lors de l'actualisation des pondérations λ_i^j.

4. Exemple

Pour évaluer l'intérêt de cette nouvelle méthode nous avons construit deux jeux de données présentant de différentes degrés de difficultés pour la reconnaissance des classes a priori (classes avec formes et tailles différentes, …). Ces deux configurations sont obtenues à partir de deux ensembles de points dans R^2, la première présente des classes de points bien séparées et la seconde des classes empiétantes. Ces ensembles de points sont répartis dans 4 classes de formes différentes (2 classes de forme sphérique et 2 classes de forme elliptique). L'évaluation de ces méthodes a été réalisée dans le cadre d'une expérience Monte Carlo : 100 réplications ont été réalisées pour chacune des deux configurations et nous avons calculé la moyenne de l'indice de Rand corrigé (RC). Dans chaque réplication la méthode de classification est répétée 50 fois et le meilleur résultat obtenu est retenu.

	Ensemble 1	*Ensemble 2*
Méthode non adaptative	0.64	0.40
Méthode adaptative	0.85	0.69

Tableau de comparaison avec l'indice RC

La comparaison entre ces deux méthodes a été réalisée par le test des espérances de Student avec un risque de 5%. D'après ce test, l'hypothèse que la performance moyenne de la méthode adaptative est supérieure à celle de la méthode non adaptative est acceptable.

5. Conclusions.

Dans cet article, nous avons présenté une nouvelle méthode de Nuées Dynamiques sur un tableau d'intervalles utilisant une version adaptative de la distance de Hausdorff. Cette méthode a été comparée dans le cadre d'une expérience de Monte Carlo avec la version non adaptative introduite par Chavent and Lechevallier (2002) selon la moyenne de l'indice de Rand corrigé. Les résultats obtenus montrent l'intérêt de cette nouvelle approche.

6. Références

[BOCK 00] BOCK, H. H., DIDAY, E., *Analysis of Symbolic Data, Exploratory methods for extracting statistical information from complex data*. Studies in classification, Data Analysis and Knowledge Organization, Springer-Verlag, 2000.

[CHAV 02] CHAVENT, M., LECHEVALLIER, Y. "Dynamical Clustering of Interval Data : Optimisation of an Adequacy Criterion based on Hausdorff Distance". In : K. Jajuga, A. Sokolowsky and H.-H. Bock (eds.) : *Classification, Clustering and Data Analysis. Recent Advances and Applications*. Springer-Verlag, pp 53-60, 2002.

[DID 71] DIDAY, E.,"La méthode des Nuées dynamiques ".*Rev. Stat. Appliquée*, Vol XIX, p19-34, 1971.

[GOV 75] GOVAERT, G., *Classification automatique et distances adaptatives*. Thèse, Paris 6,1975.

Adaptation des cartes topologiques auto-organisatrices aux tableaux de dissimilarités

Aïcha El Golli[*,**] — **Brieuc Conan-Guez**[*,**]

∗ *Projet AXIS, INRIA-Rocquencourt Domaine De Voluceau,*
BP 105 Bâtiment 18
78153 Le Chesnay Cedex, France
{*aicha.el_golli, brieuc.conan-guez*}*@inria.fr*
∗∗ *Université de Paris IX Dauphine, Lise Ceremade,*
Place du Maréchal De Lattre de Tassigny,
75775 Paris Cedex 16, France

RÉSUMÉ. *Le traitement des données complexes (données symboliques, données semi-structurées, données fonctionnelles) ne peut être réalisé facilement par les méthodes classiques de classification basées sur le calcul d'un centre de gravité. On propose dans ce travail une adaptation des cartes topologiques auto-organisatrices aux tableaux de dissimilarités. Cette approche générale permet le traitement aisé de nombreux types de données.*

MOTS-CLÉS : *carte topologique auto-organisatrice, nuées dynamiques, classification, tableau de dissimilarités*

1. Introduction

Les cartes topologiques auto-organisatrices (SOM : *Self Organising Map*), introduites par Kohonen [KOH 97], sont des algorithmes de classification automatique, qui cherchent à projeter des données multidimentionnelles sur un espace discret de faible dimension, appelé "carte". Cette projection permet d'obtenir un partitionnement des individus en groupements "similaires", tout en préservant au mieux la structure topologique des données. Les cartes topologiques s'intègrent parfaitement dans le formalisme des nuées dynamiques [THI 02], tout en permettant grâce à la conservation de l'ordre topologique, une meilleure visualisation des données.

En analyse de données, on est souvent amené à traiter des données complexes, comme par exemple les données symboliques [BOC 99a] (données ayant une structure complexe telles que les intervalles, les distributions, ...), les données semi-structurées (arbres, XML), ou même les données fonctionnelles (chaque individu est décrit par une fonction régulière discrétisée en un nombre fini de points d'observation). Dans ce cadre, les méthodes classiques de classification basées sur le calcul d'un centre de gravité ne peuvent plus être utilisées car chaque individu ne peut plus être décrit par un simple vecteur de $\mathbb{R}^n$. Afin de résoudre ce problème, plusieurs solutions sont envisageables selon la nature des données (par exemple l'utilisation de techniques de recodage des descriptions [REY 02] [REY 03] pour les données symboliques, ou l'utilisation d'un opérateur de projection dans le cas de données fonctionnelles). Cependant, ces méthodes nécessitent une bonne connaissance *a priori* des données, et imposent en général à l'utilisateur de fixer certains meta-paramètres (par exemple, le choix de la base de projection dans le cas de données fonctionnelles).

Dans ce travail, nous proposons comme solution alternative une adaptation des cartes de Kohonen aux tableaux de dissimilarités. Cette approche permet un traitement aisé des différents types de données évoqués ci-dessus, car seule la définition d'une mesure de dissimilarité est nécessaire au déroulement de la méthode (voir [BOC 99b] pour des dissimilarités sur données symboliques, et [WAN 99] pour des dissimilarités sur données semi-structurées).

2. Carte topologique auto-organisatrice sur tableaux de dissimilarités

L'algorithme d'auto-organisation cherche à projeter des données appartenant à un espace multidimentionnel (noté D) sur un espace de faible dimension (généralement de dimension 2). Cette réduction de dimension des données initiales se fait tout en préservant partiellement la topologie de l'espace des variables. L'objectif de l'optimisation du réseau est de reproduire sur une carte de sortie les "similarités" présentes entre les observations. La carte, notée $L((C,\Gamma),W)$, est constituée d'un ensemble C de neurones interconnectés (de cardinal m). Le lien entre les neurones se fait par l'intermédiaire d'une structure de graphe non orienté (C,Γ). Cette structure de graphe induit une distance discrète δ sur la carte : pour tout couple de neurones (c,r) de la carte, $\delta(c,r)$ est définie comme étant la longueur du plus court chemin entre c et r. A chaque neurone c de la carte est associé un point w_c de l'espace des données que l'on appelle *vecteur référent*. W est l'ensemble de tous les vecteurs référents. Après la phase d'apprentissage, la topologie de l'espace des données est conservée sur la carte : deux neurones c et r voisins sur la carte auront des vecteurs référents w_c et w_r proches au sens de la distance d définie sur l'espace des données.

De manière identique à l'algorithme des nuées dynamiques, la version batch des cartes de Kohonen [KOH 97] [THI 97] [THI 02] comporte deux phases distinctes : une phase d'affectation et une phase de représentation. Lors de la phase d'affectation, on définit une fonction d'affectation f de D vers C, qui à tout élément z de D associe le neurone dont le vecteur référent est le plus proche de z. Cette fonction induit une partition $P = \{P_c; c = 1,...,m\}$ de l'ensemble des observations (noté Ω) où chaque partie est définie par $P_c = \{z_i \in \Omega; f(z_i) = c\}$. Lors de la phase de représentation, l'algorithme minimise une fonction de coût convenablement choisie, notée $E(f, L(C,W))$. Cette fonction doit tenir compte de l'inertie interne de la partition P comme dans le cas des nuées dynamiques, tout en assurant la conservation de la topologie. Une manière de réaliser ce double objectif consiste à généraliser la fonction d'inertie associée à P en introduisant une notion de voisinage entre les neurones. Cette notion de voisinage, qui est attachée à la carte, est introduite à l'aide d'une fonction noyau positive et symétrique K. Cette fonction permet d'introduire des zones d'influence autour de chaque neurone : les distances $\delta(c,r)$, qui lient le neurone c aux autres neurones (r) de la carte, permettent de faire varier l'importance relative des différents neurones ; cette importance est quantifiée par la fonction $K(\delta(c,r))$ [THI 02].

Dans le cas classique, où chaque individu est décrit par un vecteur de $\mathbb{R}^n$, on préfère généralement utiliser la version stochastique de cet algorithme : les vecteurs référents sont recalculés à chaque présentation d'un individu.

2.1. *Principe*

Supposons à présent que nous ne disposions que d'un tableau de dissimilarités. On note d la mesure de dissimilarité de ce tableau. L'espace de représentation L_c du neurone c est l'ensemble des parties de Ω de cardinal fixé p : chaque neurone c est représenté par $a_c = \{z_{j_1},...z_{j_p}\}$, avec $z_{j_i} \in \Omega$. On voit donc que contrairement à l'algorithme classique, où chaque référent peut librement évoluer dans D tout entier, dans l'approche proposée ici, chaque neurone n'a qu'un nombre fini de représentations à sa disposition. L'espace L de représentation de la partition est la carte $L(C,a)$, avec $a = \{a_c; c = 1,...,m\}$ l'ensemble de tous les individus référents de la carte.

On note d^T la distance généralisée de $\Omega \times P(\Omega)$ dans $\mathbb{R}^+$ telle que $d^T(z_i,a_c) = \sum_{r\in C} K^T(\delta(c,r)) \sum_{z_j\in a_r} d^2(z_i,z_j)$, avec $K^T(\delta(c,r))$ la fonction noyau qui dépend de la distance entre le neurone c et le neurone r et du paramètre de température T (par exemple, $K^T(\delta) = e^{-\frac{\delta^2}{T^2}}$). Plus le paramètre T est petit, plus le nombre de neurones inclus dans le voisinage est réduit.

Pendant l'apprentissage, on cherche à minimiser la fonction de coût E suivante en alternant les phases d'affectation et les phases de représentation :

$$E(f,L(C,a)) = \sum_{z_i\in\Omega} d^T(z_i,a_{f(z_i)}) = \sum_{z_i\in\Omega}\sum_{r\in C} K^T(\delta(f(z_i),r)) \sum_{z_j\in a_r} d^2(z_i,z_j) \qquad [1]$$

Cette fonction mesure l'adéquation entre une partition induite par la fonction d'affectation et une carte $L(C, a)$.

Lors de la phase d'affectation, la fonction d'affectation f est celle qui affecte tout individu z_i au neurone de la carte le plus proche au sens de la distance d^T :

$$f(z_i) = arg \min_{c \in C} d^T(z_i, a_c) \quad [2]$$

On voit donc que la phase d'affectation fait décroitre le critère E.

Lors de la phase de représentation, on cherche le système d'individus référents a^* qui représente au mieux l'ensemble des observations au sens de E. Cette étape d'optimisation combinatoire peut être réalisée de manière indépendante pour chaque neurone. On minimise en effet m fonctions de la forme :

$$E_r = \sum_{z_i \in \Omega} K^T(\delta(f(z_i), r)) \sum_{z_j \in a_r} d^2(z_i, z_j) \quad [3]$$

Dans la version batch classique, la minimisation de la fonction E est immédiate car la position des vecteurs référents est donnée comme le centre de gravité du nuage de points pondérés par la fonction K.

2.2. *L'algorithme*

Initialisation : À $k = 0$, choisir un système de référents initial a^0 et la carte $L(C, a^0)$. Fixer T à T_{max} et le nombre total d'itérations à N_{iter}

Itération : À l'itération k, l'ensemble des individus référents a^{k-1} de l'étape précédente est connu. Calculer la nouvelle valeur de $T = T_{max} * (\frac{T_{min}}{T_{max}})^{\frac{k}{N_{iter-1}}}$ [THI 02]

► **Phase d'affectation :** mettre à jour la fonction d'affectation f_{a^k} associée au système a^{k-1}. On affecte chaque observation au référent défini à partir de l'équation [2].

► **Phase de représentation :** déterminer le nouveau système a^{k*} qui minimise la fonction $E(f_{a^k}, L(C, a))$. a_c^{k*} est défini de manière unique à partir de l'équation [3].

Répéter **Itération** jusqu'à ce que l'on atteigne $T = T_{min}$

L'extension des cartes topologiques aux mesures de dissimilarités a été introduite selon une méthode différente par Thore Graepel et Klaus Obermayer [GRA 99]. Dans cette extension, les auteurs se sont basés sur la version stochastique de l'algorithme du SOM. La représentation des neurones est différente de celle présentée dans ce travail : les auteurs utilisent en effet une fonction d'appartenance pour représenter les neurones.

3. Expérience

Dans cette expérience, on a opté pour une représentation du neurone par un seul individu. On dispose d'un tableau de dissimilarités de 500 observations. La carte est de taille $(9 * 3)$, l'intervalle de variation de T est $T_{max} = 6, T_{min} = 0.4$. La figure 1 présente l'ensemble des observations ainsi que l'état initial de la carte (les individus référents ont été choisis aléatoirement parmi les individus). La figure 2 montre la carte finale obtenue après apprentissage avec l'algorithme proposé. Le résultat est satisfaisant, car on obtient une bonne quantification de l'espace, tout en conservant la topologie des données.

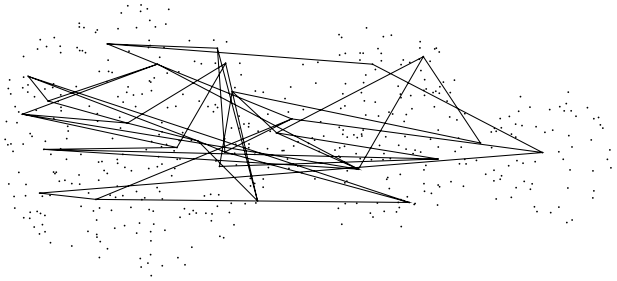


FIG. 1. *Ensemble des observations et état initial de la carte sur le plan ACP*

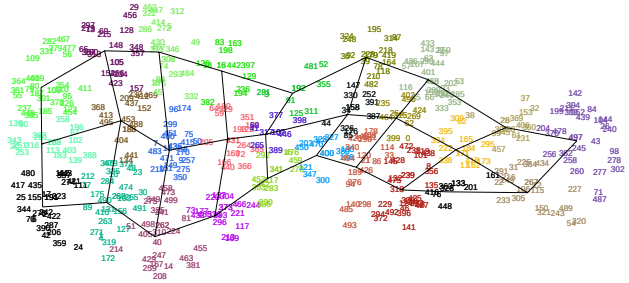


FIG. 2. *Représentation de la carte finale après 100 itérations sur le plan ACP*

4. conclusion

Les cartes topologiques permettent de concevoir des modèles de classification pour lesquels la visualisation des résultats est facilitée par l'ordre élaboré grâce à la carte. L'adaptation des cartes topologiques aux tableaux de dissimilarités permet de traiter de nombreux types de données complexes de manière aisée. La mise en oeuvre de cet algorithme sur des données réelles est en cours de réalisation.

5. Bibliographie

[BOC 99a] BOCK H. H., DIDAY E., *Analysis of symbolic Data, Exploratory methods for extracting statistical information from complex data*, Springer, 1999.

[BOC 99b] BOCK H. H., DIDAY E., *Analysis of symbolic Data, Exploratory methods for extracting statistical information from complex data*, Chapitre Similarity and dissimilarity, p. 139-197, Springer, 1999.

[GRA 99] GRAEPEL T., OBERMAYER K., A stochastic self-Organizing Map for Proximity Data, *Neural Computation*, vol. 11, 1999, p. 139-155.

[KOH 97] KOHONEN T., *Self-Organisation Maps*, Springer Verlag, New York, 1997.

[REY 02] DE REYNIÈS A., Classification de données symboliques : une extension de la méthode des nuées dynamiques, *Actes du IXème congrès de la société Francophone de Classification*, , 2002, p. 177-180.

[REY 03] DE REYNIÈS A., Classification et discrimination en analyse de données symboliques, PhD thesis, Université Paris Dauphine, 2003.

[THI 97] THIRIA S., LECHEVALLIER Y., GASCUEL O., CANU S., *Statistique et méthodes neuronales*, Dunod, Paris, 1997.

[THI 02] THIRIA S., DREYFUS G., ALL, *Réseaux de neurones méthodologie et applications*, Eyrolles, Paris, 2002.

[WAN 99] WANG J. T.-L., WANG X., LIN K.-I., SHASHA D., SHAPIRO B. A., ZHANG K., Evaluating a class of distance-mapping algorithms fordata mining and clustering, *Proceedings of the fifth ACM SIGKDD international conference on Knowledgediscovery and data mining*, 1999, p. 307-311.

Feature selection in DNA microarrays

Joaquim F. Pinto da Costa and Luis M.A. Silva

FCUP-DMA/LIACC, University of Porto, Rua Campo Alegre, 687, 4169-007 Porto, Portugal
(jpcosta@fc.up.pt ; lsilva@fc.up.pt)

ABSTRACT. *In this work, some known techniques of Data Analysis are applied in order to analyse the distinction between different classes of tumors. The data consist of the expression levels of some thousand genes (variables) in relatively few patients. This is a challenge for classical data analysis methods, as the number of variables far exceeds the number of observations. In addition to strategies based on standard statistical tests of hypothesis, some dimension reduction techniques and two clustering methods are used. These are Principal Component Analysis, Partial Least Squares and Canonical Variates for reducing the input space dimension ; CHAVL [LER 93] and Supervised Clustering of Genes [DET 02] to group the genes into clusters of similar genes.*

KEYWORDS : *Microarrays, variable selection, ACP, PLS, canonical variates, clustering*

1. Introduction

One of the main objectives of genomics is to understand how the genes function. Some questions made are : what functional papers play some genes and in which cellular processes do they participate ? how do they interact between themselves ? how does the expression of a gene changes with diseases or treatments ? The expression of the genetic information contained in the DNA molecule occurs in two phases. In the first one, the *transcription*, a part of the DNA molecule is copied into mRNA, on the basis of the complementarity property. In the second phase, the *translation*, mRNA is read to produce a protein. Knowing the transcription abundance of genes, which are contiguous parts of the DNA, became essential to answer the above questions. The transcription process of a gene sequence from DNA to a mRNA sequence (that will serve as support for the protein production) is known as gene expression. Basically, the gene expression level indicates the approximate number of copies of that gene produced in a cell ; this is believed to be correlated with the amount of corresponding protein produced. The ability of monitoring the expression of a gene in the transcription phase became possible with the technology of DNA microarrays ; this technology offers the first great hope for a global vision of the biological processes, becoming an essential tool in molecular biology and clinical diagnosis. In fact, the application of a specific regimen of chemotherapy depends on the patient correct diagnostic. With the ability of monitoring simultaneously the expression of thousand of genes, DNA microarrays represent a great step for cancer classification, that, currently, needs the intervention of different specialists. However, this technology produces information where the number of variables (genes) far exceeds the number of observations (samples), making the use of standard statistical tools difficult. In this sense, it is necessary to implement dimension reduction strategies of the predictor space. Although the number of genes available is huge (in the thousands), it is assumed that only some gene subsets or gene components determine the cancer type of a sample. The identification of predictive genes is a central objective of microarray studies applied to cancer classification. The creation of mechanisms that allow to retain only those genes or components is essential not only on a statistical level (only with dimension reduction we will be able to use standard methodologies), but also for biological interpretation (that shouldn't be disregarded).

The main contribution of our work is to introduce the use of canonical variates and CHAVL ("Classification Hiérarchique par Analyse de la Vraisemblance des Liens") [LER 93] in this type of problem and compare these methods with the methods usually used.

2. Methods Usually Used

2.1. *Tests of hypothesis*

Usually, microarray information is stored in a high dimension rectangular matrix $\mathbf{X}^T$ with 2000 to 7000 rows, corresponding to genes (plus another row containing the class information) and 40 to 100 columns, representing the samples (tumors from different patients); the ij-th cell of the matrix represents the expression level of gene i in sample j. In a problem of cancer classification, these huge datasets place serious problems in the use of standard classification tools, due to the high dimension of the predictor space (number of genes). Most of the methods used in the literature, concerning only a binary classification problem, start by doing a kind of statistical test for the difference of means [GOL 99, NGU 02] in order to select a prespecified number of genes. Then, other multivariate techniques are applied.

2.2. *Linear combinations*

Two of the multivariate techniques usually used are principal components (PCA) and partial least squares (PLS). These are essentially linear combinations of the original variables that optimize some specific criterion. PCA tries to find those linear combinations with maximum variance. If $\mathbf{x} = (x_1, x_2, \ldots, x_p)^T$ represents the observation vector, the i-th principal component solves the problem

$$\begin{aligned} \mathbf{w}_i = argmax_{\mathbf{w}}\ & Var(\mathbf{w}^T\mathbf{x}) \\ \text{subject to}\ & \mathbf{w}^T\mathbf{w} = 1 \\ & Cov(\mathbf{w}_i^T\mathbf{x}, \mathbf{w}_k^T\mathbf{x}) = 0 \qquad k < i \end{aligned} \tag{1}$$

The PLS components are very similar to PCA, but they also account for the variation of the response variable (tumor class). It has been proved [STO 90] that PLS is equivalent to find sucessive linear combinations of the original variables with maximum covariance with the response variable. It can also be algebraically represented as solving the problem

$$\begin{aligned} \mathbf{w}_i = argmax_{\mathbf{w}}\ & Corr^2(\mathbf{w}^T\mathbf{x}, \mathbf{y}) Var(\mathbf{w}^T\mathbf{x}) \\ \text{subject to}\ & \mathbf{w}^T\mathbf{w} = 1 \\ & Cov(\mathbf{w}_i^T\mathbf{x}, \mathbf{w}_k^T\mathbf{x}) = 0 \qquad k < i \end{aligned} \tag{2}$$

We have also decided to introduce the use of Canonical Variates (CV) in these data, as we have not seen its use in the literature on microarray studies. The CV components, $\mathbf{w}_i^T\mathbf{x}$, $i = 1, 2, \ldots, K-1$, are found by maximizing the criterium

$$J(\mathbf{W}) = \frac{\left|\mathbf{W}^T S_B \mathbf{W}\right|}{\left|\mathbf{W}^T S_W \mathbf{W}\right|}, \tag{3}$$

where $\mathbf{W}$ is a matrix whose columns are the unit vectors $\mathbf{w}_i$, S_B the between group dispersion matrix and S_W the within group dispersion matrix of the vector $\mathbf{x}$ whose observations are stored in a matrix $\mathbf{X}$ with n rows (number of tumors) and $p+1$ columns (p genes and the last column contains the class of each tumor).
Choosing an appropriate component subset we get a projection onto a reduced (optimized) space.

2.3. *Clustering*

Clustering techniques were probably the first to be used in microarray studies. Grouping genes with similar expression patterns turned out to be very useful in finding groups of genes with similar functions and a way to classify poorly characterized or novel genes [EIS 98].
Concerning tumor classification, Dettling *et al.* [DET 02] proposal consists on a supervised clustering technique that constructs groups of genes by directly incorporating the response variable (tumor class) into the grouping process. The process is ruled by a backward/forward mechanism and an empirical objective function that measures cluster ability for tumor discrimination. Each group $C_i,\ i = 1, \ldots, q$ is then represented by an expression value obtained by a simple linear combination $X_{C_i} = \frac{1}{|C_i|} \sum_{g \in C_i} \alpha_g X_g$ with $\alpha_g \in \{-1, 1\}$. Here we note that a particular gene g can contribute to X_{C_i} by its 'sign-flipped' expression value $-X_g$ [DET 02]. Thus, every sample (tumor) available is only represented by a q-dimensional vector, whose components are the mean values X_{C_i}.

We aim to use CHAVL ("Classification Hiérarchique par Analyse de la Vraisemblance des Liens")[LER 93] in the context of gene expression studies. This is a hierarchical clustering method developed by Lerman and co-authors which has the particularity of assigning a statistic to each level of the hierarchy, allowing thus to identify the most significant levels ; that is, those corresponding to stable partitions. The main difference between CHAVL and the usual methods is in the definition of the similarity measure. To evaluate similarities between objects, CHAVL uses probabilistic measures based on the statistical notion of likelihood [LER 93]. Instead of incorporating the response value in the clustering procedure, we start by choosing the most important genes using the methods presented in section 2.1. Then, CHAVL is used in order to cluster these genes. After running CHAVL, we represent each cluster by an expression value which is obtained by the above linear combination X_{C_i} introduced by Dettling *et al.* [DET 02]. Our aim is to compare both methodologies.

3. Some results

3.1. *Leukemia dataset*

This dataset, which can be obtained in www.genome.wi.mit.edu/MPR, contains the expression levels of $p = 3571$ genes in $n = 72$ patients suffering from acute lymphoblastic leukemia (ALL, 47 cases) and acute myeloid leukemia (AML, 25). ALL can also be divided in two subtypes : B-cell (ALL-B, 38) and T-cell (ALL-T, 9). This dataset is divided in training and test subsets. The former comprises 38 samples (19 ALL-B, 8 ALL-T and 11 AML) and the latter comprises 34 samples (19 ALL-B, 1 ALL-T and 14 AML)

Figures **1(a)-1(b)** show the projection of the 72 samples of LEUKEMIA dataset (ALL-B/ALL-T/AML distinction) onto the two first principal components. On the left we have used all the available genes (3571) resulting on a poor discriminative projection ; on the right, we have pre-selected 50 of the most predictive genes according to ANOVA (these genes had a test statistic with a p-value less than 10^{-7}, corresponding to a significance level far below 1% according to a Bonferroni adjustment for multiple comparisons, see section 2.1). Then we applied PCA and as we can see this improves the discriminative power of the principal components. On the left example, the variance explained by the first two principal components was about 30%, against the 50% of the right example. This shows that combining different methodologies can improve the results.

Figures **1(c)-1(d)** show the projection of the LEUKEMIA dataset (ALL/AML distinction) onto the two first PLS components. The procedure was the same as in PCA. The left example uses all genes (3571) ; for the right example 50 genes were pre-selected with t-statistics (see section 2.1). The discrimination between classes is evident, and shows the potentiality of this method. We can see that PLS components are more robust and not significantly affected by the inclusion of non-predictive variables (genes).

Nguyen *at al.* [NGU 02] conducted a study where they compared PCA with PLS applied to the LEUKEMIA dataset (ALL/AML distinction). Varying the number of pre-selected genes, they constructed 3 components and used logistic discrimination and quadratic discriminant analysis to classify the samples. As expected, the results revealed that PLS is in general better than PCA. The latter had a competitive behavior when the number of pre-selected genes was small. This is obvious because in this case we are already deriving principal components with highly predictive variables. Their methods had always at least one error (2.9%) in the test set.

We have used canonical variates (CV) to project the LEUKEMIA dataset and then have used these components together with K-NN to classify. To prevent the singularity of S_W (remember that we have more variables than samples) we pre-select about 30 genes according to ANOVA (the training set has 38 samples) for the three class case and with t-statistics for the two class case. Figure **1(e)** show the projection of the LEUKEMIA (ALL-B/ALL-T/AML distinction) dataset onto the CV directions. As we can see the discrimination is perfect. To verify the potential of this method we obtained the discriminant variables from the training set and used 1-nearest neighbor to predict the test set. The predictor made only one error (2.9%) for the two class distinction ALL/AML (an AML sample was predicted as ALL). For the three class case ALL-B/ALL-T/AML, one error (2.9%) was also made (an ALL-B sample was predicted as ALL-T).

We went on to evaluate the capability of CHAVL to group genes with predictive potential for tumor discrimination (ALL/AML distinction). The clustering of the 48 pre-selected genes can be seen in figure **2**. Cutting the tree at the uppest level it is possible to identify three clusters where one of them comprises only one gene (gene 33). Obtaining the X_{C_i} values as above, each sample from the training and test sets are then represented by a 3-dimensional vector. Next, 1-nearest neighbour was applied to predict the test set and the results are promising. Only one error was made (2.9%). We have also considered the next level of the tree where six clusters can be identified (two of them with only one gene). Applying the same methods as above the results were identical ; the same AML sample was incorrectly predicted ALL[1]. Hence the parsimonious model is preferred. Figure **3** show the projection of the training and test sets onto the 3-dimensional space $X_{C_1} \circ X_{C_2} \circ X_{C_3}$. As we can see the discrimination is nearly perfect.

3.2. *SRBCT dataset*

This dataset, which can be obtained in www.nhgri.nih.gov/DIR/Microarray, contains the expression levels of $p = 2308$ genes from children suffering from small round blue cell tumors divided into four classes : neuroblastoma (NB, 12), rhabdomyosarcoma (RMS, 20), Burkitt lymphoma (BL, 8) and Ewing sarcoma (EWS, 23). Note that here we only have the equivalent to the training set reported in the literature. In order to apply canonical variates with the same methodology as in the LEUKEMIA case we divided the available samples in training (48 cases) and test (15 cases) subsets trying to keep the distribution of the training set equal to the original. About 40 genes were pre-selected with ANOVA. Figure **1(f)** above show the projection of SRBCT onto the CV directions. Again, we can see that this method perfectly discriminates the classes. Next, 1-nearest neighbor was used to predict the test set and no errors were made (as expected from the literature).

4. References

[DET 02] DETTLING M., BÜHLMANN P., Supervised Clustering of Genes, *Genome Biology*, vol. 3(12), 2002.

[EIS 98] EISEN M., SPELLMAN P., BROWN P., BOTSTEIN D., Cluster analysis and display of genome-wide expression patterns, *Proc. Natl. Acad. Sci. USA*, vol. 95, 1998, p. 14863-14868.

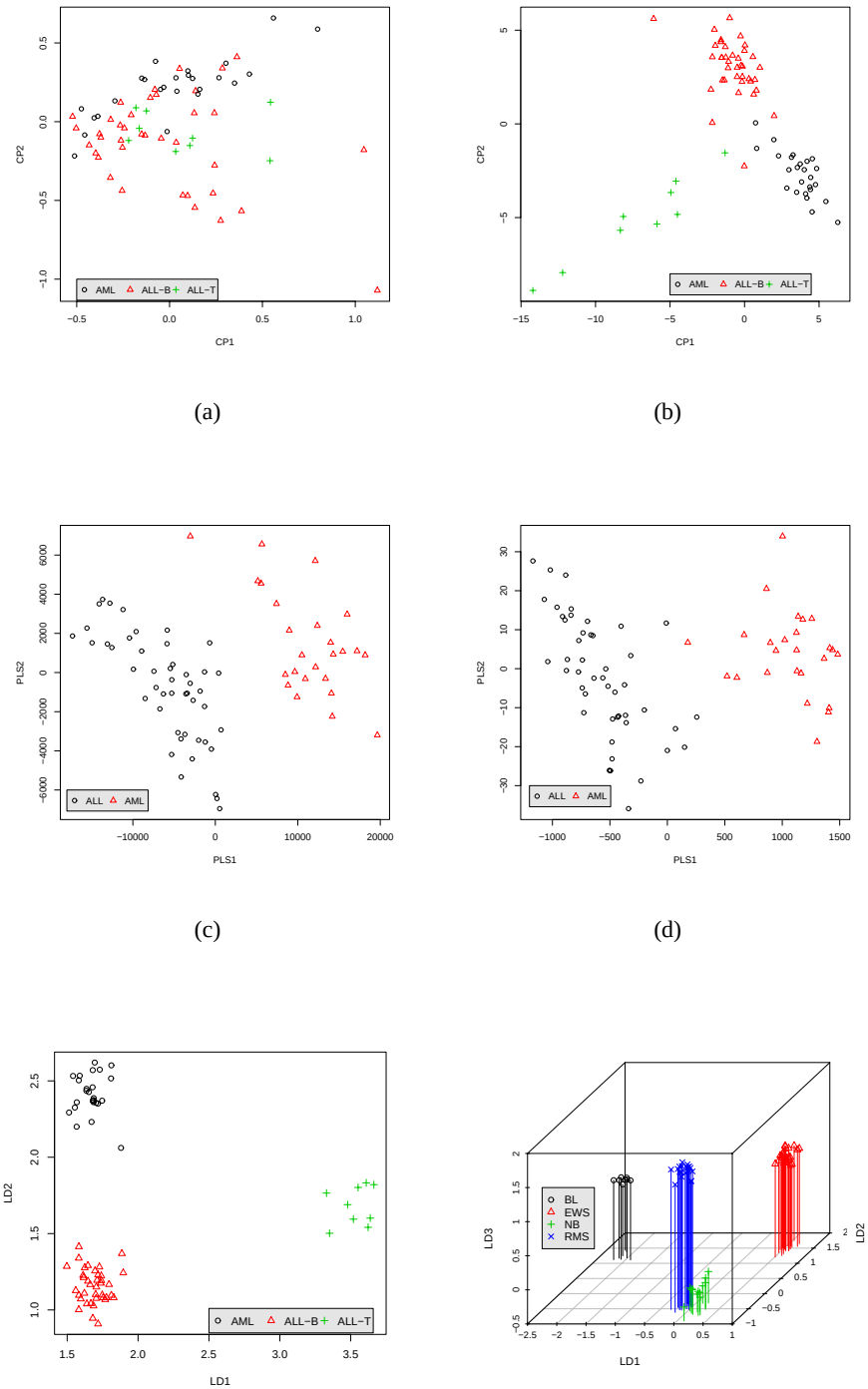
[GOL 99] GOLUB T., SLONIM D., TAMAYO P., HUARD C., GAASENBEEK M., MESIROV J., COLLER H., LOH M., DOWNING J., CAGLIURI M., BLOOMFIELD C., LANDER E., Molecular classification of cancer : class discovery and class prediction by gene expression monitoring, *Science*, vol. 286(5439), 1999, p. 531–537.

[LER 93] LERMAN I., PETER P., H. L., Principes et calculs de la methode implante dans le programme CHAVL I et II, *La Revue de Modulad*, vol. I : 1993, numro 12,pp. 33-70,II : 1994, numro 13, INRIA, 1993.

[NGU 02] NGUYEN D., ROCKE M., Tumor classification by partial least squares using microarray gene expression data, *Bioinformatics*, vol. 18, 2002, p. 39–50.

[STO 90] STONE M., BROOKS R., Continuum regression : cross validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal component regression (with discussion), *J. R. Statist. Soc. B*, vol. 52, 1990, p. 237–269.

1. This sample is suspected to be incorrectly labeled [NGU 02]



(a) (b)

(c) (d)

(e) (f)

FIG. 1. *Upper : projection of the* LEUKEMIA *dataset onto two principal components ; Middle : projection of the* LEUKEMIA *dataset onto two PLS components ; Lower : projection of the* LEUKEMIA *(left) and* SRBCT *(right) datasets onto the canonical variates. More details in the text.*

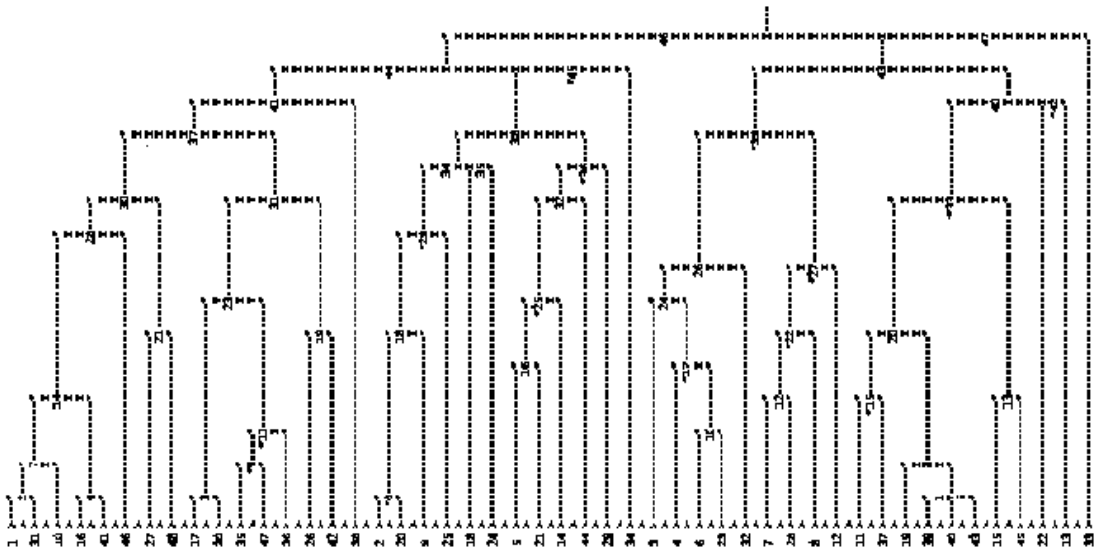


FIG. 2. *Hierarchical tree produced by CHAVL showing the most significant levels*

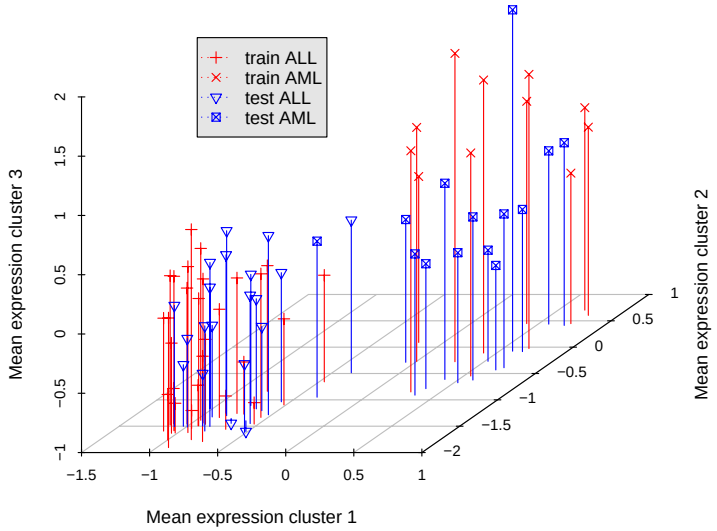


FIG. 3. *Projection of the* LEUKEMIA *dataset onto the 3-dimensional space* $X_{C_1} \circ X_{C_2} \circ X_{C_3}$

Fermés de Galois d'un contexte et classes faibles associées à des mesures de dissimilarité multivoies

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 nous plaçons dans le cadre d'un contexte dit de descriptions ordonnées, où un ensemble fini d'entités est décrit dans un ensemble ordonné. Il s'ensuit une correspondance de Galois entre l'ensemble des parties de l'ensemble des entités et l'espace de description des entités, lorsque celui-ci est un inf-demi-treillis. Cette correspondance de Galois induit alors un opérateur de fermeture dans chacun de ces deux ensembles. Dans cette note, nous nous intéressons particulièrement à l'opérateur de fermeture induit dans l'ensemble des parties de l'ensemble des entités. Nous montrons que les parties non vides fermées au sens de cet opérateur coïncident avec les classes faibles associées à des mesures de dissimilarité multivoies d'un type particulier, qui s'avèrent être quasi-ultramétriques.*

MOTS-CLÉS : *correspondance de Galois, opérateur de fermeture, contexte formel, description ordonnée, dissimilarité multivoies, quasi-ultramétrique, classe faible.*

1. Introduction

Les méthodes de classification fondées sur les mesures de (dis)similarités produisent des classes dont on peut mesurer, en quelque sorte, les degrés d'homogénéité et d'isolation. Toutefois, ces classes apparaissent souvent difficilement interprétables puisqu'elles ne comportent aucune information quant à la description des entités qui les composent. Par ailleurs, les méthodes fondées sur l'analyse formelle de concepts produisent des classes qui, étant fermées au sens d'un opérateur de fermeture, ont une description conceptuelle leur permettant de mieux se prêter à une interprétation. Dans cette note, nous montrons que ces classes fermées au sens d'un opérateur de fermeture peuvent être intégralement produites par une méthode fondée sur les (dis)similarités. En effet, nous nous plaçons dans le cadre d'un contexte dit de descriptions ordonnées, c'est-à-dire, un couple (E, δ) où E est un ensemble fini non vide dont les éléments sont appelés entités, et δ une application définie sur E à valeurs dans un ensemble ordonné $(\Omega, \leq)$. De tels contextes, également rencontrés en analyse de données symboliques [BOC 00], généralisent les contextes formels introduits en analyse formelles de concepts [GAN 99]. Lorsque $(\Omega, \leq)$ est un inf-demi-treillis, l'application δ, appelée descripteur, induit une correspondance de Galois entre les ensembles ordonnés $(\mathcal{P}(E), \subseteq)$ et $(\Omega, \leq)$. Cette correspondance de Galois induit alors des opérateurs de fermeture ϕ_δ et $\phi_\delta{}'$ dans $(\mathcal{P}(E), \subseteq)$ et $(\Omega, \leq)$, respectivement. Nous montrons que l'ensemble des parties ϕ_δ-fermées non vides de E coïncide avec l'ensemble des classes faibles associées à des mesures de dissimilarité multivoies d'un type particulier, qui s'avèrent être quasi-ultramétriques [DIA 98, DIA 97].

2. Mesures de dissimilarité multivoies et contextes de descriptions ordonnées

2.1. *Mesures de dissimilarité multivoies*

Soit E un ensemble fini non vide. Soit k un entier supérieur ou égal à 1. Pour toute partie non vide X de E, $X^*_{\leq k}$ désignera l'ensemble des parties non vides de X d'au plus k éléments. On appellera mesure de *dissimilarité multivoies* sur E toute fonction d à valeurs réelles positives ou nulles définie sur un ensemble de parties non vides de E, telle que $d(X) \leq d(Y)$ lorsque $X \subseteq Y$. Parfois nous considérerons des mesures de dissimilarité dites k-voies, où k est un entier supérieur ou égal à 2, i.e., des mesures de dissimilarité multivoies définies sur $E^*_{\leq k}$. Par ailleurs, nous écrirons simplement $d(x)$ ou $d(x, y)$ pour $x, y \in E$. Les mesures de dissimilarité classiques (2-voies) correspondent au cas $k = 2$. On notera que la condition usuelle $d(x) = 0$ n'est pas requise dans la présente note.

Soit $k \geq 2$ et d une mesure de dissimilarité k-voies sur E. Le *d-diamètre* d'une partie non vide X de E sera $\mathrm{d}iam_d(X) = \max\{d(Y) : Y \in X^*_{\leq k}\}$. Soit $X \in E^*_{\leq k-1}$ et r un nombre réel positif ou nul. La *d-boule* de centre X et de rayon r sera l'ensemble $B^d(X, r)$ (ou simplement $B(X, r)$) défini par $B(X, r) = \{y \in E : d(X + y) \leq r\}$, où $X + y$ désigne $X \cup \{y\}$. Soit maintenant p un entier tel que $1 \leq p \leq k$ et $X \in E^*_{\leq p}$. La (d, p)-*boule* engendrée par X sera l'ensemble B^d_X défini par $B^d_X = B(X, d(X))$ si $|X| \leq k-1$ et $B^d_X = \bigcap_{x \in X} B(X - x, d(X))$ sinon, où $X - x$ désigne $X \setminus \{x\}$.

Une mesure de dissimilarité k-voies d sur E sera dite *quasi-ultramétrique* si elle vérifie à la fois la condition d'inclusion (i.e. pour tout $X \in E^*_{\leq k}$ et tout $Y \in (B^d_X)^*_{\leq k}$, $B^d_Y \subseteq B^d_X$) et la condition du diamètre (i.e. pour tout $X \in E^*_{\leq k}$, $\mathrm{d}iam_d(B^d_X) = d(X)$) [DIA 97]. On notera que toute mesure de dissimilarité 2-voies ultramétrique (i.e. telle que $d(x, y) \leq \max\{d(x, z), d(y, z)\}$ pour tous x, y, z) est quasi-ultramétrique. Par ailleurs, on montre qu'une mesure de dissimilarité k-voies d sur E est quasi-ultramétrique si et seulement si elle satisfait l'implication suivante indépendamment introduite dans [BAN 94] : pour tous $X \in E^*_{\leq k}$, $Y \in E^*_{\leq k-1}$ et $z \in E$, $z \in B^d_X$ implique $\mathrm{d}iam_d(X \cup Y + z) \leq \mathrm{d}iam_d(X \cup Y)$.

2.2. *Contextes de descriptions ordonnées*

En analyse formelle de concepts, un contexte formel est un triplet $\mathbb{K} = (E, A, R)$, où E et A sont des ensembles et R une relation binaire de E vers A [GAN 99]. La relation R induit une correspondance de Galois entre les ensembles ordonnés $(\mathcal{P}(E), \subseteq)$ et $(\mathcal{P}(A), \subseteq)$ par le biais des applications $f : X \mapsto \bigcap_{x \in X} \{a \in A : (x, a) \in R\}$ et $g : I \mapsto \bigcap_{a \in I} \{x \in E : (x, a) \in R\}$, pour $X \subseteq E$ et $I \subseteq A$; ainsi, les applications $g \circ f$ et $f \circ g$ sont des opérateurs de fermeture dans $\mathcal{P}(E)$ et $\mathcal{P}(A)$, respectivement [BIR 67, BAR 70]. Les points fixes de ces opérateurs de fermetures seront appelés *fermés de Galois*. Un *concept formel* de $\mathbb{K}$ est un couple $c = (X, I)$ tel que $f(X) = I$ et $g(I) = X$. Les parties X et I sont respectivement appelées *extension* et *intention* de c. Nous étendons les contextes formels en introduisant les contextes dits de descriptions ordonnées : on appellera *contexte de descriptions ordonnées* un couple (E, δ), où E est un ensemble et δ une application définie sur E à valeurs dans un ensemble ordonné $(\Omega, \leq)$. Les éléments de E seront appelés *entités* et δ *descripteur*. Lorsque E est fini et $(\Omega, \leq)$ un inf-demi-treillis, δ induit une correspondance de Galois entre $(\mathcal{P}(E), \subseteq)$ et $(\Omega, \leq)$ par le biais des applications $f : X \mapsto \inf\{\delta(x) : x \in X\}$ et $g : \omega \mapsto \{x \in E : \omega \leq \delta(x)\}$, pour $X \subseteq E$ et $\omega \in \Omega$. Les opérateurs de fermeture $g \circ f$ et $f \circ g$ seront désignés par ϕ_δ et $\phi_\delta{}'$, respectivement. Leurs points fixes seront respectivement dits ϕ_δ-fermés et $\phi_\delta{}'$-fermés. Pour $X \subseteq E$, $\delta(X)$ désignera l'ensemble des descriptions des entités appartenant à X. Dans tout ce qui suit, E désignera un ensemble fini non vide et (E, δ) un contexte de descriptions ordonnées où δ applique E dans un inf-demi-treillis $(\Omega, \leq)$.

Appelons un *indice* sur un ensemble ordonné $(P, \leq)$ toute application h définie sur P à valeurs réelles positives ou nulles, telle que $h(x) \leq h(y)$ lorsque $x \leq y$. Un indice *strict* sera un indice h tel que $x < y$ implique $h(x) < h(y)$. Une mesure de dissimilarité multivoies d sur E sera dite *inf-compatible* s'il existe un indice h sur $(\Omega, \leq)$ avec lequel elle est inf-compatible, i.e., telle que $d(X) \leq d(Y) \iff h(\inf \delta(X)) \geq h(\inf \delta(Y))$. Si h est strict d sera dit *strictement* inf-compatible. Le lecteur observera que lorsque $(\Omega, \leq)$ est un sup-demi-treillis,

on peut définir une notion de compatibilité duale en inversant l'inégalité à droite de l'équivalence ci-dessus et en remplaçant les bornes inférieures par des bornes supérieures. Rappelons que l'amplitude d'un inf-demi-treillis $(P, \leq)$ est le plus petit entier strictement positif k tel que la borne inférieure de $(k+1)$ éléments de P est toujours la borne inférieure de k éléments parmi ces $k+1$. Cela étant, nous dirons qu'une partie X d'un inf-demi-treillis $(P, \leq)$ est d'amplitude k si k est le plus petit entier strictement positif tel que pour tout $Y \subseteq X$ tel que $|Y| = k+1$, il existe $y \in Y$ tel que $\inf(Y - y) \leq y$.

Théorème 2.1

(i) si $\delta(E)$ est d'amplitude 1, alors toute mesure de dissimilarité 2-voies sur E strictement inf-compatible est ultramétrique.

(ii) si $\delta(E)$ est d'amplitude $k \geq 2$, alors toute mesure de dissimilarité k-voies sur E strictement inf-compatible est quasi-ultramétrique.

3. Caractérisation de classes fermées

Pour tout entier $k \geq 2$, soit d_k une mesure de dissimilarité k-voies sur E. Une partie non vide X de E sera appelée classe *forte* associée à d_2 (ou classe d_2-*forte*) si son *indice d'isolation* d_2-*forte* $\mathbf{i}_1^{d_2}(X) := \min_{\substack{x,y \in X \\ z \notin X}} \{d_2(x,z) - d_2(x,y)\}$ est strictement positif ; elle sera appelée classe *faible* associée à d_k (ou classe d_k-*faible*) si son *indice d'isolation* d_2-*faible* $\mathbf{i}_k^{d_k}(X) := \min_{\substack{Y \in X^*_{\leq k} \\ z \notin X}} \{ \max_{Z \in Y^*_{\leq k-1}} d_k(Z+z) - d_k(Y)\}$ est strictement positif [BAN 89]. On notera que deux classes fortes associées à une mesure de dissimilarité 2-voies sont soit disjointes soit emboîtées, alors que les classes faibles associées à une mesure de dissimilarité k-voies forment une collection hiérarchique k-faible ; une collection hiérarchique k-faible de parties de E est une partie de $(\mathcal{P}(E), \subseteq)$ d'amplitude au plus k [BAN 94, DIA 97, BER 02].

Proposition 3.1 *Pour tout entier k supérieur ou égal 2, soit d_k une mesure de dissimilarité k-voies sur E. Soit X un sous-ensemble non vide de E. Alors, d'une part, les conditions (a1), (a2) ci-dessous sont équivalentes, et chacune d'elles implique (a3). D'autre part, les conditions (b1), (b2) sont équivalentes, et chacune d'elles implique (b3).*

(a1) X est une classe forte associée à d_2.

(a2) $B(x, d_2(x,y)) \subseteq X$ pour tous $x, y \in X$.

(a3) $X = B(x, d_2(x,y))$ pour $x, y \in X$ tels que $d_2(x,y) = \mathrm{diam}_{d_2}(X)$.

(b1) X est une classe faible associée à d_k.

(b2) $B_Y^{d_k} \subseteq X$ pour tout $Y \in X^*_{\leq k}$.

(b3) $X = B_Y^{d_k}$ pour $Y \in X^*_{\leq k}$ tel que $\mathrm{diam}_{d_k}(X) = d_k(Y)$.

Proposition 3.2 *Soit d une mesure de dissimilarité k-voies inf-compatible sur E. Alors, pour tout $X \in E^*_{\leq k}$ et tout $Y \in X^*_{\leq k-1}$, nous avons :*

(i) $\phi_\delta(X) = B(Y, d(X))$ si et seulement si $\inf \delta(B(Y, d(X))) = \inf \delta(X)$.

(ii) $\phi_\delta(X) = B_X^d$ si et seulement si $\inf \delta(B_X^d) = \inf \delta(X)$.

(iii) $\inf \delta(B_X^d) = \inf \delta(X)$ implique $\mathrm{diam}_d(B_X^d) = d(X)$.

(iv) $\inf \delta(B(Y, d(X))) = \inf \delta(X))$ implique $\mathrm{diam}_d(B(Y, d(X))) = d(X)$.

(v) $\phi_\delta(X) = B(Y, d(X))$ implique $B_X^d = B(Y, d(X))$.

Théorème 3.3 *Pour tout entier $k \geq 2$, soit d_k une mesure de dissimilarité k-voies sur E strictement inf-compatible.*

(i) Si $\delta(E)$ est d'amplitude 1, alors l'ensemble $\mathcal{F}_\delta{}^*$ des parties ϕ_δ-fermées non vides de E coïncide avec l'ensemble des $(d_2, 1)$-boules.

(ii) Si $\delta(E)$ est d'amplitude $k \geq 2$, alors $\mathcal{F}_\delta{}^*$ coïncide avec l'ensemble des (d_k, k)-boules.

Ce résultat caractérise, en particulier, aussi bien les extensions que les intentions de concepts formels tels que définis en analyse formelle de concepts [WIL 82, GAN 99]. Il s'ensuit aussi le résultat suivant qui fait clairement le lien entre l'approche de la classification fondée sur les mesures de (dis)similarité et celle fondée sur l'analyse formelle de concepts.

Corollaire 3.4 *Pour tout entier $k \geq 2$, soit d_k une mesure de dissimilarité k-voies sur E strictement inf-compatible.*
(a) Si $\delta(E)$ est d'amplitude 1, alors les conditions suivantes sont équivalentes.
(a1) X est une partie ϕ_δ-fermée non vide de E.
(a2) X est une classe forte associée à d_2.
(b) Si $\delta(E)$ est d'amplitude $k \geq 2$, alors les conditions suivantes sont équivalentes.
(b1) X est une partie ϕ_δ-fermée non vide de E.
(b2) X est une classe faible associée à d_k.

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.

[BAN 94] BANDELT H.-J., DRESS A. W. M., An order theoretic framework for overlapping clustering, *Discrete Mathematics*, vol. 136, 1994, p. 21–37.

[BAR 70] BARBUT M., MONJARDET B., *Ordre et classification*, Hachette, Paris, 1970.

[BER 02] BERTRAND P., JANOWITZ M. F., The k-Weak Hierarchical Representations : an extension of the Indexed Closed Weak Hierarchies, *Discrete Applied Mathematics*, vol. (in press), 2002.

[BIR 67] BIRKHOFF G., *Lattice theory*, 3rd edition, Coll. Publ., XXV, American Mathematical Society, Providence, RI, 1967.

[BOC 00] BOCK H., DIDAY E., Eds., *Analysis of Symbolic Data*, Springer-Verlag, 2000.

[DIA 97] DIATTA J., Dissimilarités multivoies et généralisations d'hypergraphes sans triangles, *Math. Inf. Sci. hum.*, vol. 138, 1997, p. 57–73.

[DIA 98] DIATTA J., FICHET B., Quasi-ultrametrics and their 2-ball hypergraphs, *Discrete Mathematics*, vol. 192, 1998, p. 87–102.

[GAN 99] GANTER B., WILLE R., *Formal concept analysis, Mathematical foundations*, Springer Verlag, Berlin, 1999.

[WIL 82] WILLE R., Restructuring lattice theory : an approach based on hierarchies of concepts, RIVAL I., Ed., *Ordered sets*, p. 445–470, Ridel, Dordrecht-Boston, 1982.

Emboîtements et implications associés aux hiérarchies et systèmes de fermeture

Florent Domenach et Bruno Leclerc

Institute of Policy and Planning Sciences
Tsukuba University
1-1-1 Tenno-Dai
Tsukuba, Ibaraki 305-8573, Japon
domenach@sk.tsukuba.ac.jp

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. *On considère les ensembles de classes particuliers que sont les systèmes de fermeture (ou familles de Moore), et leurs opérateurs de fermeture et relations d'implication et d'emboîtement associés. On rappelle que des résultats forts sont connus dans le cas des hiérarchies. On se base sur ceux-ci pour définir une procédure d'agrégation et d'ajustement des systèmes de fermeture conduisant à une hiérarchie.*

MOTS-CLÉS : *Système de fermeture, famille de Moore, partition, hiérarchie, préordre total, correspondance de Galois, fermeture, implication, emboîtement, classification ascendante.*

1. Introduction

En 1965, Simon Régnier [RÉG 65] proposait d'unifier des variables de types divers par la prise en considération de leurs seules partitions de relèvement, et commençait l'étude des partitions centrales. On trouve la même démarche dans Mirkin [MIR 75], avec une autre approche de la recherche de partitions consensus. Nous proposons ici de généraliser les partitions par les systèmes de fermeture, susceptibles de retenir plus d'information pertinente, et bien adaptés aux approches symboliques. Le problème de l'agrégation des partitions se transmet alors à ces systèmes. L'agrégation des systèmes de fermeture a été abordée par Raderanirina [RAD 01]) et, de façon moins explicite, dans [LEC 03].

On considère ici une démarche adaptée à la recherche d'une hiérarchie. L'étude des systèmes de fermeture particuliers associés aux hiérarchies conduit en effet à définir un *indice d'emboîtement ternaire* permettant de choisir les éléments à regrouper dans une procédure ascendante. Cette procédure, présentée dans [DOM 02], s'apparente aux méthodes agrégatives de scores, comme celles décrites dans [BAG 88] dans le cas des "*X*-arbres" non enracinés.

2. Systèmes de fermeture et hiérarchies

Un cadre habituel en classification est de considérer un ensemble de classes $\mathcal{F} \subseteq \mathcal{P}(S)$ sur S fini, tel que les éléments d'une classe F se ressemblent entre eux ou se distinguent par des propriétés communes. Deux conditions sont alors naturelles :

S est une classe

$F, F' \in \mathcal{F} \Rightarrow F \cap F' \in \mathcal{F}$ (la classe des éléments de S qui sont dans F et dans F')

Avec ces deux propriétés, $\mathcal{F}$ est un **système de fermeture** (ou *famille de Moore*) sur S. De tels systèmes sont naturellement associés à des descriptions d'objets par des variables de divers types.

Exemple 2.1. Soit C un préordre total (ordre total avec ex-aequos) sur S. L'ensemble $\{\{t \in S : (s, t) \in C\} : s \in S\}$ est un système de fermeture sur S. De tels préordres sont induits par des variables ordinales ou numériques.

Exemple 2.2. Soit $\pi = \{S_1,\ldots, S_k\}$ une partition de S, par exemple induite par une variable qualitative. L'ensemble $\pi \cup \{S, \varnothing\}$ est un système de fermeture sur S.

Exemple 2.3. Soit un ensemble fini S décrit par un ensemble fini V d'attributs binaires. On a alors une relation $R \subseteq S \times V$, où $(s, v) \in R$ si l'objet s possède l'attribut v. Pour tous $A \subseteq S$ et $T \subseteq V$, on pose $AR = \{v \in V$: pour tout $s \in S$, $(s, v) \in R\}$ et $RT = \{s \in S$: pour tout $v \in V$, $(s, v) \in R\}$. Soient $\mathcal{R} = \{AR : A \subseteq S\}$ et $C = \{RT : T \subseteq V\}$. Selon les résultats bien connus sur le treillis de Galois de R, les ensembles $\mathcal{R}$ et C sont deux systèmes de fermeture, anti-isomorphes pour l'inclusion par $A \mapsto AR$ et $T \mapsto RT$ pour tous $A \in C$, $T \in \mathcal{R}$.

Exemple 2.4. Soit $\mathcal{H}$ une hiérarchie sur S, c'est-à-dire un ensemble de parties de S vérifiant : (H1) $S \in \mathcal{H}$, (H2) pour tout $s \in S$, $\{s\} \in \mathcal{H}$; et (H3) pour tous $H, H' \in \mathcal{H}$, $H \cap H' \in \{\varnothing, H, H'\}$. L'ensemble $\mathcal{H} \cup \{\varnothing\}$ est un système de fermeture sur S, dit *système hiérarchique*.

Une application φ sur $\mathcal{P}(S)$ est une *fermeture* si elle est *isotone* (pour tous $A, B \subseteq S$, $A \subseteq B \Rightarrow \varphi(A) \subseteq \varphi(B)$), *extensive* (pour tout $A \subseteq S$, $A \subseteq \varphi(A)$), et *idempotente* (pour tout $A \subseteq S$, $\varphi(\varphi(A)) = \varphi(A)$). Alors, l'image $\mathcal{F} = \varphi(\mathcal{P}(S))$ de $\mathcal{P}(S)$ par φ constitue un système de fermeture sur S et, réciproquement, si $\mathcal{F}$ est un système de fermeture, une fermeture φ sur $\mathcal{P}(S)$ est donnée par $\varphi(A) = \bigcap \{F \in \mathcal{F} : A \subseteq F\}$, c'est-à-dire que $\varphi(A)$ est la plus petite classe de $\mathcal{F}$ contenant A, et une telle classe existe toujours.

3. Implications et emboîtements

On considère deux types de relations binaires sur $\mathcal{P}(S)$, respectivement les *systèmes implicatifs complets* (SIC) et les *relation d'emboîtement* sur S. Un système implicatif complet est une relation, notée ici $\rightarrow$, sur $\mathcal{P}(S)$ satisfaisant les conditions :

(I1) $B \subseteq A \rightarrow A \rightarrow B$,

(I2) pour tous $A, B, C \subseteq S$, $A \rightarrow B$ et $B \rightarrow C \rightarrow A \rightarrow C$,

(I3) pour tous $A, B, C, D \subseteq S$, $A \rightarrow B$ et $C \rightarrow D \rightarrow A \cup C \rightarrow B \cup D$.

Une relation d'emboîtement est une relation, notée ici $\sqsubset$, sur $\mathcal{P}(S)$ vérifiant :

(O1) $A \sqsubset B \rightarrow A \subseteq B$;

(O2) $A \sqsubset B \sqsubset C \rightarrow [A \sqsubset C \leftrightarrow A \sqsubset B$ ou $B \sqsubset C]$;

(O3) $A \sqsubset A \cup B \rightarrow A \cup B \sqsubset B$.

Il y a équivalence entre systèmes (et espaces) de fermeture, SICs et relations d'emboîtement ; la première des équivalences ci-dessous est due à Armstrong ([ARM 74] ; cf. l'article de synthèse de Caspard et Monjardet [CAM 03]) ; la seconde est dans [DOL 02] :

$$A \rightarrow B \leftrightarrow B \subseteq \varphi(A)$$

$$A \sqsubset B \leftrightarrow A \subseteq B \text{ et } \varphi(A) \neq \varphi(B)$$

L'interprétation en termes de classification est que A implique B si toute classe contenant A contient B, tandis que A est emboîté dans B signifie que A est inclus dans B et qu'il y a une classe contenant A et ne contenant pas B.

Adams [ADA 86] a donné une première caractérisation des emboîtements particuliers associés aux systèmes hiérarchiques. Une autre caractérisation, pour les emboîtements et pour les implications, est fournie dans [DOL 02]. A côté d'une condition un peu technique (tout sous-ensemble "critique" est de cardinal 2 ; cf. [CAM 03] encore pour lesdits ensembles) vérifiée également par les “hiérarchies faibles”, elle comporte une condition sur les triplets pouvant prendre plusieurs formes équivalentes. Nous retenons ici la suivante :

(T1) pour tous $s, s', t \subseteq S$, $ss' \rightarrow ss't \rightarrow st \rightarrow s'$ et $s't \rightarrow s$.

La condition (T1) exprime un fait bien connu : si la classe minimum de la hiérarchie $\mathcal{H}$ considérée contenant s et s' est strictement incluse dans celle contenant s, s' et t, alors toute classe contenant s et t (s' et t) contient aussi s' (s).

4. Un indice ternaire d'emboîtement et son usage

On définit, pour tout système de fermeture $\mathcal{F}$ sur S, un *indice d'emboîtement ternaire* $\varphi_F(ss')$ par : pour tous $s, s' \subseteq S$ (distincts), $\varphi_F(ss') = |\{t \subseteq S : ss' \rightarrow ss't\}|$.

Cet indice correspond à une similarité entre s et s', mais relative aux autres éléments de S. Dans le cas d'une hiérarchie, on a la propriété suivante :

Proposition 4.1. Soient $\mathcal{H}$ une hiérarchie sur S et deux éléments distincts s, s' de S tels que $\delta_{\mathcal{H}}(ss')$ est maximal. Alors s et s' appartiennent à une classe de $\mathcal{H}$ non triviale minimale pour l'inclusion. Si de plus, la hiérarchie $\mathcal{H}$ est binaire, ss' est une classe de $\mathcal{H}$.

On déduit de cette propriété des procédures agglomératives [Dom02] pour la construction d'une hiérarchie à partir d'un profil $\mathbf{F} = (\mathcal{F}^1, \mathcal{F}^2, \ldots, \mathcal{F}^k)$ de systèmes de fermeture (e.g. un k-uple de hiérarchies), On associe à $\mathbf{F}$ un indice global $\delta^{\mathbf{F}}$ par $\delta^{\mathbf{F}}(ss') = \sum_{1 \le i \le k} \delta_{\mathcal{F}}(ss')$. Ensuite, à chaque étape, on réunit les deux éléments s et s' maximisant l'indice $\delta^{\mathbf{F}}(ss')$, puis on remet à jour les éléments de $\mathbf{F}$ (avec plusieurs façons de le faire). On définit ainsi, en particulier, des procédures de consensus de hiérarchies apparemment inédites (cf. [LEC 98]).

5. Bibliographie

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

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

[BAG 88] BARTHÉLEMY J.-P., GUÉNOCHE A., *Les arbres et les représentations des proximités*, Paris, Masson, 1988.

[CAM 03] CASPARD N., MONJARDET B., The lattices of Moore families and closure operators on a finite set: a survey, *Discrete Applied Math.*, 127 (2), 241–269, 2003.

[DOM 02] DOMENACH F., *Structures latticielles, correspondances de Galois contraintes et classification symbolique*, thèse de l'Université Paris 1, 2002.

[DOL 02] DOMENACH F., LECLERC B., *Closure Systems, Implicational Systems, Overhanging Relations and the case of Hierarchical Classification*, 2002, soumis.

[LEC 98] LECLERC B., Consensus of classifications: the case of trees, in *Advances in Data Science and Classification* (A. Rizzi, M. Vichi, H.-H. Bock, eds), *Studies in Classi-fication, Data Analysis and Knowledge Organization*, Berlin, Springer-Verlag,, 81-90, 1998.

[LEC 03] LECLERC B., The median procedure in the semilattice of orders. *Discrete Applied Math.*, 127 (2), 285–302, 2003.

[MIR 75] MIRKIN B., On the problem of reconciling partitions, In *Quantitative Sociology, International Perspectives on mathematical and Statistical Modelling*, New York, Academic Press, 441-449, 1975.

[RAD 01] RADERANIRINA, V., Treillis et agrégation de familles de Moore et de fonctions de choix, thèse de l'Université Paris 1, 2001.

[RÉG 65] RÉGNIER S., Sur quelques aspects mathématiques des problèmes de classification automatique, *ICC Bulletin* 4, 175-191, 1985 (*Math. Sci. Hum.*, 82, 13-29, 1983).

Utilisabilité d'un environnement de fouille de données

Edwige Fangseu Badjio, François Poulet

38, rue des Docteurs Calmette et Guérin
Parc Universitaire de Laval-Changé
53000 Laval
{edwige.fangseubadjio, poulet}@esiea-ouest.fr

RÉSUMÉ. *Nous présentons dans cet article des travaux visant à améliorer l'utilisabilité des environnements de fouille de données. Partant du constat suivant lequel la prise en main de ces outils n'est pas évidente pour des utilisateurs non spécialistes, nous avons isolé des critères qui nous semblent intéressants pour le guider lors des différents choix qu'il aura à effectuer lors du processus de traitement. Pour ce faire, nous avons mis en place un système à base de connaissances basé sur des tests statistiques pour la prise de décision.Ces mesures ont éte recueillies lors des précédentes exécutions des algorithmse de l'environnement et sont mises à jour au fur et à mesure de son utilisation.*

MOTS-CLÉS : *fouille de données, utilisabilité, système à base de connaissances.*

1. Introduction

La quantité de données stockées est de plus en plus importante, mais ces données n'ont une utilité que si au moins une partie de l'information qu'elles contiennent est utilisée. C'est le but de l'extraction de connaissance dans les données et plus particulièrement de la fouille de données. De très nombreux algorithmes automatiques de fouille ont été développés. De plus, on assiste aujourd'hui à l'avènement de méthodes interactives de fouille de données [Poul 02]. L'utilisateur de ces dernières techniques anthropocentrées est le spécialiste des données et non plus un spécialiste de fouille ou analyse de données. Ce type d'approche présente au moins les avantages suivants :

- on utilise l'expertise du domaine lors de l'ensemble du processus de fouille,
- la compréhensibilité et la confiance dans le modèle construit sont accrues puisque l'utilisateur a participé à sa construction,
- on bénéficie des capacités humaines en reconnaissance de formes.

Par contre, l'utilisateur du système n'étant plus un expert en analyse de données, il faut guider ses choix lors des différentes étapes du processus de fouille, par exemple pour trouver l'algorithme le plus adéquat en fonction de ses données et du problème à traiter.

A cette fin, nos recherches sont orientées vers l'amélioration de l'utilisabilité des environnements de fouille de données. Il s'agit pour nous de trouver des solutions pour que l'utilisateur ait non seulement une prise en main facile de l'outil mais aussi que les résultats obtenus soient le plus compréhensible possible. Actuellement, l'environnement utilisé est un environnement graphique regroupant plusieurs algorithmes automatiques et interactifs de classification (supervisée et non supervisée). Nous nous intéressons plus spécifiquement ici à un choix particulier qui doit être effectué dans le cas de la classification supervisée : le choix d'un algorithme d'induction d'arbres de décision. Pour pouvoir effectuer ce choix, il est nécessaire de trouver des critères de qualité de ces algorithmes et d'évaluer l'adéquation avec le type et la quantité de données à traiter et le problème que l'on désire résoudre. L'idée ici est

de répertorier un ensemble de connaissances relatives aux caractéristiques de plusieurs ensembles de données et aux résultats de l'exécution des algorithmes disponibles dans l'environnement sur chacun de ces ensembles de données. La sélection de l'algorithme de classification à utiliser pour la résolution d'un problème donné sera un processus multicritères permettant de guider l'utilisateur.

La définition des critères de qualité des algorithmes d'induction d'arbre de décision est décrite dans le paragraphe 2, de même que les critères sur les données et la composition de ces différents critères pour obtenir finalement un critère d'évaluation de la pertinence de l'algorithme. Quelques premiers résultats sont présentés ensuite dans le paragraphe 3 avant la conclusion et nos travaux futurs.

2. Critères de choix

Nous abordons dans ce paragraphe le choix des critères à retenir sur les ensembles de données et sur les algorithmes de classification supervisée pour pouvoir ensuite fournir à l'utilisateur un classement des différents algorithmes disponibles, du plus pertinent au moins pertinent en fonction des données qu'il a à traiter. Ces critères sont systématiquement mesurés et enregistrés à chaque utilisation de l'environnement de fouille de données. Ceci nous permet d'alimenter la base de connaissances au fur et à mesure de l'utilisation du système.

2.1. *Critère sur les données*

Les critères retenus sur les données sont aussi bien des critères statistiques, d'analyse discriminante ou exploratoire que des mesures issues de la théorie de l'information. Ces critères sont les suivants :

- le nombre d'individus,
- le nombre de classes,
- le nombre d'attributs (variables),
- le nombre d'attributs symboliques,
- le nombre de valeurs manquantes et leur probabilité relative,
- la moyenne et l'écart type de chaque variable,
- le nombre de fonction discriminante,
- les coefficients de corrélation, d'asymétrie et d'applatissement,
- l'entropie, l'entropie relative et le rapport signal-bruit.

Le détail du calcul de ces mesures peut être trouvé dans [MiST 94].

2.2. *Critère sur les algorithmes*

Pour l'initialisation de notre système, nous avons retenu les critères suivants qui seront calculés et mémorisés à chaque exécution d'un algorithme de l'environnement :

- la compréhensibilité des modèles fournis par l'algorithme,
- le temps d'exécution,
- le taux de précision (avec validation croisée puis un test de Student avec écart-type inconnu).
- la robustesse.

2.3. *Fusion des critères*

Un ensemble non exhaustif de bases de données a été répertorié. Pour chaque base de données de cet ensemble, les caractéristiques des données décrites dans les paragraphes précédents et un classement des algorithmes par efficacité décroissante sont stockés dans la base de connaissances. Nous avions le choix entre soit retourner l'algorithme le plus

approprié à la résolution du problème ou retourner un ensemble d'algorithmes d'efficacité décroissante. Pour permettre au système de prendre aussi en compte les préférences de l'utilisateur, nous avons fait le choix de lui présenter la liste des algorithmes classés dans l'ordre du plus performant au moins performant. Pour un nouveau problème soumis en entrée de l'environnement, les critères sur les données du problème sont calculés, puis on effectue une recherche dans la base de connaissances pour trouver un ensemble de données dont les caractéristiques sont statistiquement équivalentes (ou les plus semblables) aux caractéristiques des données du problème à résoudre. La liste classée des algorithmes est alors proposée à l'utilisateur qui se chargera de lancer l'exécution de l'un des algorithmes.

3. Quelques résultats

Les résultats présentés ci-dessous sont obtenus lors d'exécutions de notre programme. L'idée ici est de répertorier un ensemble de connaissances, connaissances relatives aux données et aux résultats des exécutions des algorithmes.

Le tableau 1 représente un sous-ensemble des résultats obtenus sur des ensembles de données du "ML Repository" de l'Université de Californie- Irvine [BlMe 98].

	Segment	**Satimage**	**Diabetes**	**Australian**	**Shuttle**
Nombre d'observations	2310	6435	768	690	43500
Nombre d'attributs	19	36	8	14	9
Nombre de classes	7	6	2	2	7
Nombre d'attributs continus	19	36	2	6	9
Nombre d'attributs qualitatifs	0	0	0	8	0
Rapport signal-bruit	4.0014	1.2970	1.0377	1.2623	1.6067
Coefficient de corrélation	0.1425	0.5977	0.1439	0.1024	0.3558
Corrélation canonique	0.9760	0.9366	0.5507	0.7713	0.9668
Fract	0.3098	0.3586	1.0000	1.0000	0.6252
Coefficient d'asymétrie	2.9580	0.7316	1.0586	1.9701	4.4371
Coefficient d'aplatissement	24.4813	4.1737	5.8270	12.5538	160.3108
Entropie des classes	2.8072	2.4734	0.9331	0.9912	0.9653
Entropie des attributs	3.0787	5.5759	4.5301	2.3012	3.4271
Information mutuelle	0.6672	0.9443	0.1120	0.1130	0.3348

Tableau 1. *Caractéristiques des ensembles de données utilisés*

Le tableau 2 représente un sous-ensemble des résultats (précision et nombre de feuilles des arbres) obtenus par exécution des algorithmes C4.5 [Quin 93], PBC [AnEK 00], CIAD [Poul 01], OC1 [MKSB 93] et CART [BFOS 84] sur les ensembles de données ci-dessus.

	C4.5		**PBC**		**CIAD**		**CART**		**OC1**	
Australian	84.4	85	82.7	9	86.7	10	96.8	6	85.94	2
Satimage	85.2	563	83.5	33	85	22	84	19	86	16
Segment	96.6	77	94.8	94.8	95.1	19	93.58	19	93.9	10
Diabetes	78.1	20	79	16	79.4	14	78	7	82.16	16
Shuttle	99.9	57	99.9	8.9	99.9	8	99.9	27	99.9	20

Tableau 2. *Précision des algorithmes et nombre de feuilles des arbres de décision*

Le tableau 3 représente des résultats du classement des algorithmes en donnant une importance égale au taux de précision, à la taille des arbres obtenus et au temps d'exécution des algorithmes. Bien sûr, ce mode de calcul ne sert qu'à illustrer notre approche, il est possible d'effectuer des calculs beaucoup plus complexes afin d'affiner les résultats. Le temps

de calcul n'est pas pris en compte pour les deux algorithmes de construction interactive d'arbres de décision utilisés : CIAD et PBC.

Algorithmes	Segment	Satimage	Diabetes	Australian	Shuttle
C4.5	1	2	1	4	2
PBC	2	5	5	5	3
CIAD	3	3	3	2	3
OC1	4	1	4	3	5
CART	5	4	2	1	1

Tableau 3. *Classement des algorithmes*

4. Conclusion

Nous avons présenté dans cet article une approche visant à améliorer l'utilisabilité des environnements de fouille de données en guidant l'utilisateur dans les choix qu'il doit effectuer tout au long du processus de traitement. A titre d'exemple, nous avons illustré notre approche pour permettre de donner une réponse à la question : pour un ensemble de données, quel est l'algorithme ou quels sont les algorithmes de classification supervisée susceptible(s) de fournir les meilleurs résultats ? Pour ce faire, nous sauvegardons un historique de toutes les exécutions réalisées sur les ensembles de données avec les algorithmes disponibles dans l'environnement de fouille développé. Ces expériences sont stockées dans un système à base de connaissances. Le moteur du dit système utilise des tests statistiques pour la prise de décision. Nous travaillons actuellement sur l'extension de notre système afin qu'il puisse prendre en compte l'ajout de nouveaux algorithmes et qu'il puisse traiter de nouvelles données, c'est-à-dire des ensembles de données dont les tests statistiques n'ont pas pu être suffisamment rapprochés de ceux obtenus sur les données précédemment utilisées dans l'environnement.

5. Bibliographie

[BlMe 98] C.Blake, C.Merz, UCI Repository of machine learning databases, [www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, University of California, Department of Information and Computer Science, (1998).

[MiST 94] MICHIE D., SPIEGELHALTER D.J. and TAYLOR C.C. (eds.), *Machine Learning, Neural and Statistical Classification* Ellis Horwood, 1994.

[Poul 02] POULET F., Arbre de décision, clustering et SVM, IXe Rencontres de la Société Francophone de Classification, Toulouse, 297-301, Sep. 2002.

[Poul 01] POULET F., CIAD : Construction Interactive d'Arbres de Décision, *SFC'2001, VIIIe Rencontres de la Société Francophone de Classification*, Pointe-à-Pitre, 275-282, Dec. 2001.

[Quin 93] Quinlan J. R. *Programs for Machine Learning*, Morgan-Kaufman Publishers, 1993.

[BFOS 84] Breiman L., Friedman J. H., Olsen R. A., and Stone C. J., *Classiffication and Regression Trees*,Wadsworth, 1984.

[MKSB 93] Murthy S., Kasif S., Salzberg S., and Beigel R., OC1: Randomized induction of oblique decision trees," in Proc. 11th National Conf. Artificial Intelligence MIT Press, 1993, pp. 322, 327.

[AnEK 00] Ankerst M., Ester M., and Kriegel H.-P., Towards an effective cooperation of the user and the computer for classiffication," in Proc. KDD '2000, Boston, pp. 179, 188.

Cinq points équidistants comme témoins de la difficulté combinatoire de L_1

Bernard Fichet

Laboratoire de Biomathématiques
Université de la Méditerranée, Marseille
13385 Marseille cedex 5.
Bernard.Fichet@medecine.univ-mrs.fr

RÉSUMÉ. Il est bien connu qu'un espace métrique fini (I,d) est de type L_1 si et seulement si d est combinaison conique-convexe de dichotomies, dissimilarités élémentaires associées à des bi-partitions. Parmi ces combinaisons, existent un nombre fini de décompositions dites minimales. Leur connaissance intervient dans la résolution de problèmes de dimensionnalité. Nous exhibons les vingt-deux décompositions dichotomiques minimales de la distance du simplexe régulier sur cinq points, re-validant ou validant ainsi sur cinq points des conjectures liées à la dimension de ce simplexe et à son analyse en composantes principales en norme L_1.

MOTS-CLÉS: Distances de type L_1, dichotomies, dimensionnalité, analyse en composantes principales, simplexe régulier, polyèdre.

1. Introduction

A l'opposé de la géométrie euclidienne, la géométrie L_1 génère des figures isométriques (finies), engendrant des espaces de dimensions distinctes. Cette non-unicité de la dimension est source de questions combinatoires de très haute complexité. Par exemple, étant donné un espace métrique fini (I,d), sur $n=|I|$ points, de type L_1, quelle est la plus petite dimension d'un espace dans lequel peut être plongé isométriquement (I,d)? Même pour une distance aussi simple que celle du simplexe régulier, la question reste toujours ouverte. On ne dispose de résultats que pour de faibles valeurs de n. Semblablement, la représentation du simplexe régulier selon des axes fournis par une analyse en composantes principales (A.C.P.), est aisée en géométrie euclidienne, bien qu'il n'y ait pas unicité de la solution. Beaucoup plus compliquée s'avère être une représentation donnée par une A.C.P. en norme L_1. Nous n'en sommes qu'à des conjectures.

De fait, il existe un lien étroit entre les figures de représentation de (I,d) et ce que l'on nomme les décompositions dichotomiques de d. Parmi toutes celles-ci, il en existe certaines,

dites minimales, en nombre fini. Exhiber toutes les décompositions dichotomiques minimales (elles peuvent être en nombre exponentiel !) est un moyen d'aborder (mais non de résoudre !) les problèmes sus-évoqués.

Pour la distance d_1 du simplexe régulier, il existe une décomposition minimale pour $n = 3$, il en existe deux pour $n = 4$. Nous montrons ici qu'elles sont au nombre de vingt-deux, appartenant à cinq familles, pour $n = 5$. Une analyse précise de ces cinq familles valide, pour $n = 5$, une conjecture sur l'A.C.P. en norme L_1, et permet de retrouver un résultat sur la dimensionnalité pour $n = 5$ et 6, lui-même source d'une conjecture pour n quelconque.

2. Résultats

Rappelons qu'un couple (I, d) constitué d'un ensemble fini I et d'une dissimilarité d, est dit de type L_1 si (I, d)est isométriquement plongeable dans un certain R^N muni de la norme L_1. Plus formellement, il existe N et pour tout i de I, des x_{ik}, $k = 1,...N$, tels que :

$$\forall i, j \in I, \quad d(i, j) = \sum_k \left| x_{ik} - x_{jk} \right| \qquad (1)$$

Nécessairement, d doit être une semi-distance. La caractérisation des semi-distances de type L_1 a été obtenue par différents auteurs, dans différents domaines, et s'exprime en termes de dichotomies. Rappelons qu'une dichotomie δ_J est associée à une bi-partition non trivale (J, J^c) de I, ou coupe. C'est une dissimilarité telle que $\delta_J(i, j)$ égale 1 si $i \in J$, $j \notin J$ (ou vice-versa), et est nulle ailleurs. Alors, voir par exemple les deux articles de Le Calvé et Fichet dans Dodge (1987) , ou Critchley et al. (1994), Deza et al. (1997), d est de type L_1 si et seulement si il existe des $\alpha_J \geq 0$ tels que : $d = \sum_J \alpha_J \delta_J$ (décomposition dichotomique). En d'autres termes, dans l'espace des fonctions numériques sur $I \times I$, symétriques et à diagonale nulle, l'ensemble D_1 des semi-distances de type L_1, forme un cône polyédrique. En outre, comme montré par Critchley et al. (1994), les dichotomies engendrent les rayons extrêmes de ce cône. Il s'en suit une caractérisation de D_1 en termes de programmation linéaire, mais de très haute complexité.

La L_1-dimensionnalité de d, i.e. le plus petit N tel que (1) soit satisfaite, est liée aux décompositions dichotomiques. Ces dernières forment un polytope de l'espace dual, les sommets étant caractéristiques des décompositions minimales, voir Benayade et al. (1994, 2002). A tout plongement dans R^N, correspond une décomposition dichotomique donnée, où chaque axe définit une suite de dichotomies emboîtées, i.e. telles que les ensembles caractéristiques soient, à leur complémentaire près, emboîtés. Ainsi, la L_1-dimensionnalité peut être recherchée sur les décompositions minimales seules. Pour plus amples développements sur ce sujet, on peut consulter Fichet (1994). Dans le cas général, i.e. si non-unicité de la décomposition dichotomique, la recherche de la plus petite dimension est très fortement combinatoire. Même pour la distance d_1 du simplexe régulier normalisé, cette

dernière reste inconnue. Il est conjecturé que cette dimension est $\lceil (n+1)/2 \rceil$ et qu'elle conduit à l'unique figure donnée par l'étoile. Une telle borne a été reconnue comme supérieure pour toutes les distances arborées, Hadlock (1978), Fichet (1994). Cette conjecture est vérifiée jusqu'à $n = 8$, Bandelt et al. (1998), Koolen et al. (2000).

Nous exhibons ici les 22 décompositions dichotomiques minimales de d_1 pour $n = 5$, et retrouvons ainsi comme corollaire le résultat de Bandelt et al. (1998), pour $n \leq 6$.

L'analyse en composantes principales(A.C.P.) en norme L_1, telle que définie par Benayade et al. (1994), relève de la même complexité combinatoire. Etant donné $d \in D_1$, l' A.C.P. en norme L_1 est définie comme la recherche d'une composante d' de d, i.e. $d' \in D_1$, $d' \leq d$, $(d - d') \in D_1$, unidimensionnelle et maximisant un critère de type L_1. On itère ensuite axe par axe. Les deux critères (linéaires) usuels sont celui de la médiane où l'on maximise $\sum_i d'(m,i)$ (où m est l'extension médiane de (I,d), voir Benayade et al. (2002)) et celui, dit global, où l'on maximise $\sum_{i,j} d'(i,j)$. Dans les deux cas , il suffit de travailler sur les décompositions minimales de d. Mais même pour la distance aussi triviale que d_1 du simplexe régulier, le résultat reste inconnu pour le critère global (le critère de la médiane conduit à l'étoile). Nous conjecturons que l'étoile demeure la solution pour le critère global. Nos résultats ci-dessus valident cette conjecture pour $n = 5$.

3. Bibliographie

[BAN 98] BANDELT H-J., CHEPOI V., LAURENT M., Embedding into rectilinear spaces, *Discrete and Computational Geometry*, 19, 595-604, 1998.

[BEN 94] BENAYADE M., FICHET B., Algorithms for a geometrical P.C.A. with the L_1-norm. In Diday E. et al. eds., , *New Approaches in Classification and Data Analysis*, Springer-Verlag, Berlin, 75-84, 1994.

[BEN 02] BENAYADE M., FICHET B., The Median Extension of Data Analysis Metric Structures, In Dodge Y., ed., *Statistical Data Analysis Based on the L_1-Norm and Related Methods*, Birkhäuser, Basel, 367-377, 2002.

[CRI 94] CRITCHLEY F., FICHET B., The partial order by inclusion of the principal classes of dissimilarities on a finite set, and some of their basic properties. In Van Cutsem B., ed., *Classification and Dissimilarity Analysis, Lecture Notes in Statistics*, Springer, 1994.

[DEZ 97] DEZA M., LAURENT M., *Geometry of Cuts and Metrics*, Springer-Verlag, 1997.

[DOD 87] DODGE Y., *Statistical Data Analysis Based on the L_1-Norm* and *Related Methods*, North-Holland, Amsterdam, 1987.

[FIC 94] FICHET B., *Dimensionality problems in L_1-norm representations.* In Van Cutsem B., ed., *Classification and Dissimilarity Analysis, Lecture Notes in Statistics*, Springer, 1994.

[HAD 78] HADLOCK F., HOFFMAN F., Manhattan trees, *Utilitas Mathematica*, 13, 55-67, 1978.

[KOO 00] KOOLEN J., LAURENT M., SCHRIJVER A., Equilateral dimension of the rectilinear space, *Designs, Codes and Cryptography*, 21, 149-164, 2000.

Un algorithme de détection de maxima de densité basé sur la distance distributionnelle : *application à la classification optimale fine d'un corpus documentaire.*

Alain Lelu* , Claire François**

**Laboratoire de Mathématique, Informatique et Génome (MIG) - INRA*
Jouy-en-Josas cedex - 78352 – alain.lelu@jouy.inra.fr
et Université de Franche-Comté / LASELDI
***INIST / Unité de Recherche et Innovation*
2 allée de Parc de Brabois – 54514 Vandoeuvre-lès-Nancy cedex – francois@inist.fr

RÉSUMÉ. Contrairement aux algorithmes EM ou à centres mobiles qui convergent vers un optimum local d'un critère global, et ce faisant empêchent toute analyse comparative (ajout/ suppression d'observations ou de descripteurs), notre méthode d'Analyse en Composantes Locales converge vers un optimum absolu : l'ensemble de tous les maxima locaux d'un critère de densité adaptative. Dans chaque classe, les observations et descripteurs sont caractérisés par un indice de centralité (ou facteur local) basé sur la distance distributionnelle. Une application à un corpus documentaire en géologie est présentée.

MOTS-CLÉS : classification automatique, décomposition factorielle oblique, densité adaptative, modes de densité, voisins réciproques, distance distributionnelle.

1. Introduction

La méthode de classification automatique dont nous présentons ici une application en grandeur réelle au domaine documentaire a été conçue au confluent de trois ordres de préoccupations, pour viser des résultats sûrs, stables et nuancés : 1) opérer dans un espace de données pourvu de la même propriété d'équivalence distributionnelle que celui de l'Analyse Factorielle des Correspondances [BEN 81], garante d'une bonne stabilité de l'analyse au regard des fusions / éclatements de descripteurs dont les significations sont voisines, 2) converger, à finesse d'analyse donnée, vers un optimum absolu d'un indicateur de qualité, comme le font les méthodes factorielles utilisant la décomposition en éléments propres d'une matrice, dont les vecteurs propres balisent les extréma et les points-selle d'un « paysage d'inertie », 3) ne pas classer en tout ou rien, mais doter chaque individu d'un indicateur de centralité dans sa classe propre - tout autant que dans les autres classes, voisines ou non.

2. Distance et cosinus distributionnels

Toute méthode d'analyse des données est caractérisée à la base par trois choix : 1) une transformation opérée sur les vecteurs-données bruts, 2) une métrique, ou pondération des dimensions dans lesquels ces vecteurs sont définis, 3) une pondération de ces vecteurs. Sur le nuage de points ainsi défini, de nombreuses techniques de synthèse d'information et réduction des dimensions peuvent être appliquées : classification ascendante ou descendante hiérarchique, classification à centres mobiles, décomposition aux valeurs singulières, incluant toutes

les méthodes d'analyse factorielle qui en sont les variantes. Notre méthode des K-means axiales n'échappe pas à la règle ; voyons tout d'abord la transformation des données et la métrique choisies, qui définissent notre distance :

Une lignée ancienne de travaux [MAT 55] [ESC 78] [DOM 79] [FIC 85] se sont intéressés à ce que ces derniers auteurs appellent distance distributionnelle , qu'on explicitera ici dans notre cadre applicatif d'analyse d'un ensemble de textes décrits par les fréquences de leurs mots : la distance distributionnelle entre 2 textes est la distance euclidienne, classique (dimensions équipondérées), entre les 2 points t1 et t2, de coordonnées les vecteurs $\mathbf{z}_{t1}$ et $\mathbf{z}_{t2}$, situés sur l'hypersphère unité dans l'espace des I mots et représentant chacun une unité textuelle, définis par la transformation suivante sur les données : $\mathbf{z}_{t1} : \{\{(x_{it1}/ x_{.t1})^{1/2}\}$; $\mathbf{z}_{t2} : \{(x_{it2}/ x_{.t2})^{1/2}\}$

(les vecteurs sont en caractères gras ; x_{it} désigne la fréquence du mot i dans le document t, et $x_{.t}$ le nombre total de mots du document t ; les accolades désignent un ensemble : de valeurs de ses composantes pour un vecteur, de vecteurs pour une matrice)

Cette distance peut prendre en compte des vecteurs ayant des composantes négatives ; elle est liée à la mesure du gain d'information de Renyi d'ordre ½ [REN 66] ; Escofier et Volle ont montré qu'elle satisfaisait à la propriété d'équivalence distributionnelle caractéristique de la distance du khi-deux utilisée en AFC.

3. Décomposition aux valeurs singulières dans l'espace distributionnel

Si l'on transforme et pondère le tableau (mots × textes) des fréquences de mots comme suit :

. Coordonnées : des vecteurs-colonnes $\mathbf{x}_t : \{ x_{it} \} \rightarrow \mathbf{y}_t : \{\sqrt{x_{it}}\}$

des vecteurs-lignes $\mathbf{x}_i : \{ x_{it} \} \rightarrow \mathbf{y}_i : \{ \sqrt{x_{it}} \}$

les cosinus entre vecteurs-colonnes $\mathbf{y}_t$ (resp. entre vecteurs-ligne $\mathbf{y}_i$) sont liés à la notion de distance distributionnelle D par la relation : $D(t_1,t_2)^2 = 2(1-\cos(t_1,t_2))$

(resp. $D(i_1,i_2)^2 = 2(1-\cos(i_1,i_2))$)

Si on calcule les directions propres $\mathbf{U} = \{ \mathbf{u}^{(k)} \}$ (resp $\mathbf{W} = \{ \mathbf{w}^{(k)} \}$) du nuage des points-colonnes $\mathbf{y}_t$ (resp, des points-lignes $\mathbf{y}_i$), au moyen de la décomposition aux valeurs singulières du tableau des racines carrées (pour des données négatives : racine des valeurs absolues, avec le signe des valeurs), on démontre que les cosinus ci-après se déduisent des directions propres :

$$\mathrm{Cos}(\mathbf{y}_t, \mathbf{u}^{(k)}) = w_t^{(k)} (\lambda^{(k)} / x_{.t})^{1/2} \quad (1)$$

$$\mathrm{Cos}(\mathbf{y}_i, \mathbf{w}^{(k)}) = u_i^{(k)} (\lambda^{(k)} / x_{i.})^{1/2} \quad (2)$$

Ces cosinus peuvent être considérés comme les facteurs d'un cas particulier et simple d'analyse factorielle sphérique, pour reprendre la terminologie de M. Volle, dite centrée sur le « tableau nul ». Ils sont liés entre eux et avec les $w_t^{(k)}$ et $u_i^{(k)}$ par des formules de transition, en particulier :

$$\mathrm{Cos}(\mathbf{y}_t, \mathbf{u}^{(k)}) = \Sigma_i\, u_i^{(k)} \sqrt{(x_{it} / \lambda^{(k)})}$$

$$\mathrm{Cos}(\mathbf{y}_i, \mathbf{w}^{(k)}) = \Sigma_t\, w_t^{(k)} \sqrt{(x_{it} / \lambda^{(k)})}$$

Le principe de notre algorithme d'Analyse en Composantes Locales est de réaliser la partition en classes des unités textuelles dans l'espace distributionnel, et d'extraire pour chaque classe les facteurs mots et documents définis ci-dessus. Chaque facteur est alors un indicateur de *centralité* (ou « typicité ») du texte (« document ») ou du mot dans sa classe (cf. [PAG 98] pour l'application d'une notion de centralité voisine, basée sur les liens et non sur les mots, en tant que mesure de notoriété globale d'une page Web, et utilisée dans le moteur de recherche Google). En complément, une représentation cartographique des thèmes est créée à partir d'une analyse en composantes principales du nuage des vecteurs-thèmes dans l'espace des mots.

5. Détections des modes d'une fonction densité

La plupart des méthodes de classification non supervisées convergent vers un maximum local d'un critère global de qualité de reconstitution des données. Ceci est bien connu en ce qui concerne les algorithmes à centres mobiles [DID 79], mais reste vrai pour des méthodes plus récentes de classification floue et recouvrante, qui utilisent l'algorithme EM, comme [BUN 02]. Ce qui est gênant pour des utilisateurs qui, au delà de quelques « formes fortes » systématiques et souvent triviales s'intéressent à un découpage plus fin de leur sujet d'étude :

selon la valeur de la graine d'initialisation au hasard, pour un même jeu de données, un même algorithme et un même nombre de classes demandées, des classes peuvent apparaître ou disparaître, fusionner ou éclater, et les frontières se ré-ajuster. D'où la difficulté pour l'utilisateur qui supprime, ajoute ou fusionne des descripteurs, ou introduit des observations recueillies dans une autre tranche de temps, de savoir si les effets visibles après classification sont dûs à ses manipulations ou à la variabilité intrinsèque de l'algorithme. Ainsi sur le jeu de données documentaires issues de la base Pascal de l'INIST utilisé ici (1400 références bibliographiques médicales provenant du programme de coopération scientifique franco-tunisienne et indexées manuellement), notre algorithme à centres mobiles K-Means Axiales [LEL 94] a montré qu'un tiers environ des 30 classes demandées était sujet à de telles instabilités ; toutefois, toutes les classes créées restent interprétables : tout se passe comme si l'algorithme, en plus des « formes fortes » stables, tirait au hasard au sein d'une population de « vraies classes » d'effectifs moyens ou faibles, en nombre bien supérieur à quelques dizaines.

De là l'idée de mettre au jour exhaustivement cette population au moyen d'un algorithme repérant l'ensemble des maxima locaux d'un paysage de densité défini sur l'espace des données. En effet, étant donné une fonction locale traduisant la densité des vecteurs-données, une information très riche sur la structure des données peut en découler, plus riche que des appartenances en tout ou rien à des classes. Par exemple, la recherche des modes de cette fonction permet d'identifier des classes, ou plutôt des pôles ; les points-selle traduisent la présence d'individus d'appartenance ambivalente, etc. Qui plus est, une fois fixé le paramètre de « granularité », de finesse de l'analyse, le « paysage de densité » est complètement déterminé, et la recherche exhaustive de tous ses modes (ou autres caractéristiques), c'est-à-dire l'énumération complète de ses optima locaux, équivaut à la découverte d'un optimum global unique pour un critère local. Mais cette fois la configuration de cet optimum n'est pas bouleversée si on enlève ou ajoute quelques vecteurs-données, elle ne connaît que des changements localisés à une partie seulement du paysage de densité, répondant en cela aux exigences de stabilité que nous nous sommes fixés plus haut.

Pour ce faire, deux principes algorithmiques ont été utilisés à ce jour : 1) des techniques de montée en gradient lorsque la fonction densité est partout ou presque partout continue (cf. [ASS 89] pour une revue des travaux pionniers en ce domaine), 2) des techniques énumératives, par ex. [TRE 79], quand cette fonction est discrète. Dans ces textes, les exemples donnés, souvent à petite échelle, sont séduisants, et le passage à une échelle supérieure semble sans problème.

Nos travaux nous ont conduits à adopter une mesure de densité basée sur un voisinage adaptatif, et non de rayon fixe : d'où le choix de la notion de voisinage des K plus proches voisins, adaptative par définition. Nous avons défini la densité d'un tel voisinage comme la moyenne des cosinus entre tout point et ses voisins.

Comme cette fonction est discrète, il n'est pas possible d'utiliser un algorithme de montée en gradient pour la maximiser. C'est pourquoi nous nous sommes tournés vers un algorithme de classification 1) utilisant la notion de densité, 2) compatible avec des fonctions densité discrètes, à savoir l'algorithme de percolation de Trémolières [TRE 79]. Contrairement au cas des algorithmes de montée en gradient sur un paysage de densité continu, où les maxima n'ont pas de raison de correspondre à des points du nuage à analyser, l'algorithme de percolation caractérise le mode de chaque classe par un point du nuage dont la densité est supérieure à celle de ses voisins (ou plusieurs points, si leur densité est égale).

Cet algorithme se moule d'autant plus dans nos préoccupations qu'il nuance la notion d'appartenance à une classe : il dégage des « points-frontière » appartenant à plusieurs classes (que nous avons rebaptisés : points ambivalents), et la densité constitue un indice de centralité d'un individu dans sa classe. Dans la publication [TRE 94] cet auteur a encore distingué d'autres catégories de nuances d'appartenance.

Le voisinage défini par les K plus proches voisins n'est pas une notion symétrique : un point peut appartenir au voisinage d'un autre sans que la réciproque soit nécessairement vraie. Après avoir réalisé une adaptation de l'algorithme de percolation aux voisinages non symétriques [LEL 98], nous avons constaté empiriquement que se limiter aux voisinages *réciproques*, symétriques, améliorait la robustesse et la qualité de l'analyse.

Le travail de l'INIST / URI a donné le jour à une interface Web de visualisation et navigation dans les résultats sans laquelle il était exclu de valider la méthode sur des données réelles, compte tenu du nombre important de classes obtenues.

6. Résultats et validation par les experts du domaine

Nos 1397 notices bibliographiques sont décrites par un vocabulaire « brut » de 2629 mots-clés ; après élimination des hapax, inutiles par nature dans une analyse multidimensionnelle, et de certains mots-clés trop

génériques, la taille du vocabulaire « validé » s'établit à 935 mots-clés. Le meilleur paramétrage de notre analyse en composantes locales s'est avéré avec 3 comme nombre de plus proches voisins réciproques, dans l'espace distributionnel décrit plus haut : les valeurs 1 à 5 ne provoquaient pas de grosses variations du nombre de thèmes, mais leur contenu était moins clair. Nous avons ainsi obtenu 703 documents isolés et 232 thèmes caractérisés chacun par 2 à 5 documents-noyaux, reliés parfois par 1 à 3 documents « passerelles » ambivalents. Le 1er facteur d'analyse factorielle sphérique de chaque thème nous a donné d'une part la caractérisation de ce thème par ses mots-clés les plus « saillants », d'autre part la projection de tous les documents, isolés ou non, noyaux ou non, sur l'axe du thème, permettant d'enrichir l'interprétation de ce thème. Une carte globale obtenue par analyse en composantes principales des profils de mots des 232 thèmes, cliquable et re-dimensionnable à volonté sur l'interface VISA de l'INIST, rend accessible ces résultats sur Internet.

Les deux médecins documentalistes qui ont expertisé les résultats ont estimé que les 232 classes détectées étaient homogènes et faisaient sens à leurs yeux ; les nuances entre certaines classes leur échappant parfois, elles en ont regroupé certaines, aboutissant à 175 thèmes de libellés distincts. La plupart des spécialités médicales sont représentées et elles ont réparti les thèmes en une vingtaine de grandes catégories médicales à partir de leur connaissance du domaine (santé publique, bactérioses, cardiologie, neurologie, …). Elles estiment que leur disposition sur une carte interactive qui rapproche les thèmes voisins et donne accès à leur contenu permet d'examiner, de re-nommer et catégoriser tous les thèmes en une à deux demi-journées de travail. Une carte de deuxième niveau peut être obtenue par ACP des profils des regroupements de thèmes, facilitant l'intelligibilité et la consultation ultérieure des résultats.

7. Bibliographie

[ASS 89] ASSELIN DE BEAUVILLE J.P., "Panorama sur l'utilisation du mode en classification automatique », *Automatique, Productique, Informatique Industrielle*, vol. 23, 1989, 113-13.

[BEN 81] BENZECRI J.P. ET COLL., *Pratique de l'Analyse des Données : Linguistique et Lexicologie,* Dunod, Paris, 1981.

[BUN 02] BUNTINE W., "Variational extensions to EM and multinomial PCA", in *proc. ECML 2002*, Elomaa T., Mannila H., Toivonen H. (Eds.), Springer, 2002

[DID 79] DIDAY E. ET COLL., *Optimisation en classification automatique*, INRIA, Rocquencourt, 1979

[DOM 79] DOMENGES D., VOLLE M., "Analyse factorielle sphérique : une exploration", *Annales de l'INSEE*, vol.35, 1979, p. 3-84, INSEE, Paris

[ESC 78] ESCOFIER B., "Analyses factorielles et distances répondant au principe d'équivalence distributionnelle", *Revue de Stat. Appliquée,* vol. 26, n°4, 1978, p. 29-37, Paris

[FIC 85] FICHET B. ET GBEGAN A., "Analyse factorielle des correspondances sur signes de présence-absence", in Diday et al. eds., *4e Journées Analyse des Données et Informatique*, 1985. INRIA, Rocquencourt.

[MAT 55] MATUSITA K., "Decision rules, based on the distance for problems of fit, two examples, and estimation", *Annals of Statistical Mathematics*, 1955, vol. 26, p. 631-640, Tokyo.

[LEB 77] LEBART L., MORINEAU A., TABARD N., *Techniques de la description statistique*, Dunod, Paris, 1977.

[LEL 94] LELU A., "Clusters *and* factors: neural algorithms for a novel representation of huge and highly multidimensional data sets", in E. Diday, Y. Lechevallier & al. eds., *New Approaches in Classification and Data Analysis*, p. 241-248, 1994, Springer-Verlag, Berlin,

[LEL 98] LELU A., FERHAN S., "Clustering a textual dataflow by incremental density-modes seeking", Actes *IFCS'98* (International Federation of Classification Societies), pp. 206-209, Université La Sapienza, Rome,1998

[PAG 98] PAGE L., BRIN S., MOTWANI R., WINOGRAD T., "The PageRank Citation Ranking: Bringing Order to the Web", *Stanford Digital Library Technologies Project*, 1998 <http://citeseer.nj.nec.com/page98pagerank.html>

[REN 66] RENYI A., *Calcul des probabilités*. Dunod, Paris,1966.

[TRE 79] Trémolières R., "The percolation method for an efficient grouping of data", *Pattern Recognition*, vol.11, n°4, 1979

[TRE 94] Trémolières R., "Percolation and multimodal data structuring", *New Approaches in Classification and Data Analysis*, Diday E. et al. (eds.), p. 263-268, Springer Verlag, Berlin, 1994

Algorithme MCMC à sauts réversibles pour les mélanges gaussiens multivariés.
Application à la détection du nombre de composants.

Guillaume Saint Pierre
Bernard Garel

Université Paul Sabatier Toulouse III,
118 route de Narbonne 31062 Toulouse, France
saint@cict.fr, garel@cict.fr

RÉSUMÉ. *Ce travail présente une généralisation de Green et Richardson (1997) au cas des mélanges gaussiens multivariés. Dans un contexte bayésien, nous présentons l'algorithme MCMC à sauts réversibles en explicitant notamment le jacobien du mouvement de séparation/combinaison. De multiples simulations sont effectuées afin d'évaluer les performances de l'algorithme, avec une attention toute particulière au nombre de composants détectés. En outre, nous comparons deux techniques permettant d'éliminer le "label switching".*

MOTS-CLÉS : *statistique bayésienne, mélanges gaussiens multivariés, algorithme MCMC à sauts réversibles, label switching, jacobien*

1. Introduction

Les modèles de mélanges gaussiens sont de plus en plus utilisés en statistique, et notamment en classification. Une littérature importante est maintenant disponible, dans laquelle on retiendra (MCL 00) contenant une étude exhaustive des problèmes intervenants dans les mélanges. Depuis l'avènement de l'algorithme EM, l'estimation des paramètres d'un mélange gaussien est devenue classique. Un problème crucial en classification n'est cependant toujours pas résolu de manière satisfaisante : la détection du nombre de composants du mélange (autrement dit le nombre de classes), nécessaire à une bonne estimation des paramètres.

Certains auteurs, comme (GAR 03) ou (LIU 03), ont développé des approches basées sur le test du rapport des maximums de vraisemblance afin de pallier aux faiblesses des critères de type AIC ou BIC. Dans un contexte bayésien, une réelle avancée vers une analyse rigoureuse et complète, a été obtenue en 1997 avec l'article de Green et Richardson (RIC 97). Se basant sur l'algorithme à sauts réversibles détaillé dans (GRE 95), ces auteurs l'appliquent avec succès au cas des mélanges gaussiens univariés. Cependant, peu de travaux considèrent le cas des mélanges gaussiens multivariés, pourtant très utiles pour modéliser des images ou des bases de données.

Nous présentons dans cet exposé l'une des premières généralisations de l'algorithme MCMC à sauts réversibles au cas des mélanges gaussiens multivariés. Notre algorithme est capable d'estimer les paramètres d'un mélange gaussien multivarié, en même temps que le nombre de composants de celui-ci.

2. Modélisation

On considère ici l'étude des modèles où les données observées $y^{(n)} = (y_1, ..., y_n)$ sont considérées indépendantes et identiquement distribuées selon un mélange gaussien multivarié fini à k composants de densité

$$p(y \mid \pi, \mu, \Sigma) = \pi_1 \mathcal{N}_r(y; \mu_1, \Sigma_1) + ... + \pi_k \mathcal{N}_r(y; \mu_k, \Sigma_k),$$

où $\pi = (\pi_1, ..., \pi_k)$ sont les proportions du mélange de somme 1, $\mu = (\mu_1, ..., \mu_k)$ et $\Sigma = (\Sigma_1, ..., \Sigma_k)$ les moyennes et les matrices de covariances, et r la dimension des données. On utilise de plus un modèle à classes latentes en introduisant des variables z_i telles que $\mathbb{P}[z_i = j] = \pi_j$ et on note $z^{(n)} = (z_1, ..., z_n)$. On peut considérer qu'à chaque $k \in \{1, 2, ..., k_{\max}\} = \mathcal{K}$, correspond un modèle noté $\mathcal{M}_k$ associé au vecteur des paramètres $\theta^{(k)} = \left(\pi, z^{(n)}, \mu, \Sigma\right) \in \mathbb{R}^{n_k}$ $(n_k \in \mathbb{N})$. Le modèle bayésien naturel est alors donné par

$$p\left(k, \theta^{(k)}, y\right) = p(k)\, p\left(\theta^{(k)} \mid k\right) p\left(y \mid k, \theta^{(k)}\right), \tag{1}$$

où $p(k)$ est la loi a priori sur la "dimension" du modèle, $p\left(\theta^{(k)} \mid k\right)$ est la loi a priori sur le vecteur des paramètres conditionnellement à k, et $p\left(y \mid k, \theta^{(k)}\right)$ est la densité de l'observation y. L'inférence sur les paramètres d'intérêts k et $\theta^{(k)}$ est alors effectuée en étudiant la loi a posteriori

$$p\left(k, \theta^{(k)} \mid y\right) \propto p(k)\, p\left(\theta^{(k)} \mid k\right) p\left(y \mid k, \theta^{(k)}\right). \tag{2}$$

L'idée générale de l'algorithme MCMC à sauts réversibles est de simuler la loi a posteriori $p\left(k, \theta^{(k)} \mid y\right)$ en utilisant une chaîne de Markov se déplacant sur $\mathcal{C} = \cup_{k \in \mathcal{K}} \{k\} \times \mathbb{R}^{n_k}$. Une condition de réversibilité nous permettra d'assurer la convergence de la chaîne vers la densité ciblée, à partir de laquelle nous déduirons l'estimation des divers paramètres en utilisant le maximum a posteriori.

3. L'algorithme à sauts réversibles

Les éléments de la chaîne de Markov générés par l'algorithme sont de la forme $\left(k, \theta^{(k)}\right)$ où k représente la dimension du modèle courant. Dans le cas des mélanges gaussiens, on envisagera 2 types de mouvements : ceux ne changeant pas la dimension du modèle mais uniquement le paramètre $\theta^{(k)}$, et ceux changeant la dimension ($k \to k+1$ ou $k \to k-1$). Pour le premier type de mouvements, on pourra se référer par exemple à (ROB 96) décrivant les algorithmes de Gibbs et de Metropolis-Hastings, permettant de générer une chaîne de Markov de distribution stationnaire p sur un ensemble quelconque E. La construction d'un noyau de transition d'une telle chaîne est remarquablement simple et peut s'obtenir de la manière suivante :

1. On choisit un noyau auxiliaire de transition quelconque Q sur E ;
2. A partir du point courant ξ, on génère un point ξ^* selon un noyau $Q(\xi^*, \xi)$;
3. On accepte le point ξ^* avec une probabilité $\rho(\xi, \xi^*) = \min\left\{1, \frac{p(\xi^*) Q(\xi, \xi^*)}{p(\xi) Q(\xi^*, \xi)}\right\}$.

On montre alors que le noyau K résultant de ces deux opérations (proposition et acceptation/rejet) vérifie les équations de réversibilité

$$K(A, B) = \int\int_{A \times B} p(\xi)\, Q(\xi^*, \xi)\, d\xi d\xi^* = \int\int_{B \times A} p(\xi^*)\, Q(\xi, \xi^*)\, d\xi^* d\xi = K(B, A).$$

Après un temps de chauffage, on peut considérer que les sorties de l'algorithme sont distribuées selon π.

Lorsque l'ensemble E est constitué de plusieurs sous espaces de dimensions différentes, $E = \mathbb{R}^{n_1} \cup \mathbb{R}^{n_2}$ par exemple, avec $n_1 < n_2$, il faut définir un mécanisme de saut de $\mathbb{R}^{n_1}$ vers $\mathbb{R}^{n_2}$, puis de $\mathbb{R}^{n_2}$ vers $\mathbb{R}^{n_1}$, qui permette à la chaîne d'être irréductible. Le mécanisme est le suivant : à partir du point ξ de $\mathbb{R}^{n_1}$, on génère de façon aléatoire $n_2 - n_1$ composantes qui permettent d'obtenir un point de $\mathbb{R}^{n_2}$. On définit alors une transformation déterministe

de $\mathbb{R}^{n_2}$ dans lui même qui donne ξ^*, et une transformation déterministe de $\mathbb{R}^{n_2}$ dans $\mathbb{R}^{n_1}$ qui rend le saut de ξ^* vers ξ possible (irréductibilité). La procédure d'acceptation rejet est alors identique au cas précédent. Le calcul de la probabilité d'acceptation nécessite celui du jacobien de la transformation qui fait passer de ξ à ξ^*.

$$\rho\left(\xi,\xi^*\right)=\min\left\{1,\frac{p\left(\xi^*\right)Q\left(\xi,\xi^*\right)}{p\left(\xi\right)Q\left(\xi^*,\xi\right)}\left|J\left(\xi,\xi^*\right)\right|\right\}.$$

La difficulté essentielle de cet algorithme réside justement dans la définition du difféomorphisme déterminant les sauts de ξ^* vers ξ, et de ξ vers ξ^*. Le jacobien associé peut parfois s'avérer très difficile à calculer, en particulier dans le cas multivarié considéré dans cet exposé.

4. Description de l'algorithme

Suivant (RIC 97), on rajoute une couche d'hyperparamètres afin d'améliorer la flexibilité du modèle. On autorise les lois a priori sur les paramètres μ, Σ, k, π, à dépendre respectivement des hyperparamètres $(\xi, \kappa, \alpha, \beta, \lambda, \delta)$ avec $\eta = (\xi, \kappa, \alpha, \beta)$. Le schéma (1) permet de visualiser les dépendances entre les paramètres et les hyperparamètres.

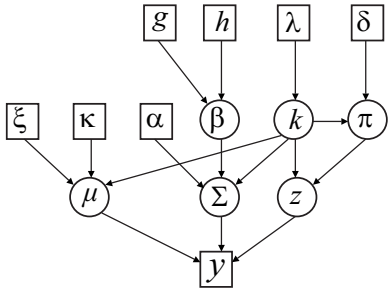


FIG. 1. *Graphe acyclique ordonné correspondant au modèle bayésien pour les lois de mélanges multivariées*

L'état courant de la chaîne est toujours $\left(k, \pi, z^{(n)}, \beta, \mu, \Sigma\right)$. On met successivement à jour les différents composants selon l'ordre suivant : (a) séparer ou combiner deux composants, (b) mise à jour de π, (c) mise à jour de (μ, Σ), (d) mise à jour des allocations $z^{(n)}$, (e) mise à jour de l'hyperparamètre β, (f) naissance ou mort d'un composant.

Les mouvements b,c,d et e nécessitent une mise à jour automatique selon les lois conditionnelles car il n'y a pas de changement de modèle. Les mouvements a et f nécessitent quant à eux l'emploi des sauts réversibles. L'ensemble de ces mouvements (de a à f) est répété 20 000 fois. A chaque itération on garde en mémoire les paramètres courants. Lorsque l'on stoppe l'algorithme, les 5000 premières itérations sont enlevées et on effectue une analyse a posteriori sur les 15 000 éléments de la chaîne restants.

Le mouvement de naissance/mort est presque identique a celui proposé par (RIC 97). L'apport essentiel de notre travail se situe au niveau du mouvement de séparation/combinaison qui est le coeur de l'algorithme. Nous devons notamment nous assurer que les matrices de covariances générées seront toujours définies positives, et ensuite calculer un jacobien de dimension très élevé. Le mouvement de combinaison sera défini comme l'inverse du mouvement de séparation, c'est pourquoi nous ne détaillerons que ce dernier.

On choisit de manière aléatoire de combiner deux composants (c'est-à-dire réduire la taille du modèle) ou de séparer deux composants (c'est-à-dire augmenter la taille du modèle). On notera j_1 et j_2 les indices des composants à combiner ou obtenus par séparation, et j^* l'indice du composant à séparer ou obtenu par combinaison. Afin de maximiser la probabilité d'acceptation du nouvel élément proposé, le difféomorphisme doit vérifier une condition

de conservation des trois premiers moments, appellée "moment matching condition" par Green. Comme on l'a vu précédemment, pour définir le difféomorphisme il devient nécessaire d'introduire de nouvelles variables, générées aléatoirement, correspondants au nombre de degrés de liberté manquants lorsque l'on passe d'un espace à k composants à un espace à $k+1$ composants. En dimension r, cela signifie que nous devons générer $1+r+\frac{r(r+1)}{2}$ nouvelles variables :

$$u1 \sim \beta(2,2) \quad u2_i \sim \beta(2,1) \quad u3_{ij} \sim \beta(2,1)$$
$$\text{pour } i = 1, ..., r \text{ et } i \leqslant j \leqslant r.$$

Pour que les matrices générées soient définies positives, nous avons eu l'idée d'utiliser la décompostion de Cholesky, consistant à décomposer Σ_j sous la forme $V_j V_j'$ où V_j est une matrice triangulaire inférieure. Les paramètres des nouveaux composants générés lors du mouvement de séparation sont donc obtenus de la manière suivante :

$$\begin{array}{lllllllll} \pi_{j_1} & = & u_1 \pi_{j^*} & \mu_{j_1} & = & \mu_{j^*} + \sqrt{\frac{1-u_1}{u1}} V_{j^*} \frac{u_2}{\sqrt{2r}} & \Sigma_{j_1} & = & \frac{1}{2u1} V_{j^*} \left(I_r - \frac{u_2 u_2'}{2r} + \frac{u_3}{2r} \right) V_{j^*}' \\ \pi_{j_2} & = & (1-u_1)\pi_{j^*} & \mu_{j_2} & = & \mu_{j^*} - \sqrt{\frac{u_1}{1-u1}} V_{j^*} \frac{u_2}{\sqrt{2r}} & \Sigma_{j_2} & = & \frac{1}{2(1-u1)} V_{j^*} \left(I_r - \frac{u_2 u_2'}{2r} - \frac{u_3}{2r} \right) V_{j^*}', \end{array} \quad (3)$$

avec $\Sigma_{j^*} = V_{j^*}' V_{j^*}$. Lorsque l'on effectue un mouvement de séparation on doit :

1. Choisir le composant j^* à séparer ;
2. Effectuer la décomposition de Cholesky $\Sigma_{j^*} = V_{j^*}' V_{j^*}$;
3. Effectuer le difféomorphisme (3) ;
4. Calculer la probabilité d'acceptation à l'aide du jacobien, et accepter ou non la nouvelle configuration.

Le jacobien d'un tel changement de variable ne peut se calculer qu'en utilisant des logiciels de calcul formel comme Maple ou Mathematica. Une fois celui-ci obtenu, nous sommes en mesure d'étudier le comportement de l'algorithme dans un contexte de détection du nombre de composants, et notamment sa sensibilité à la loi a priori. Comme de nombreux auteurs, nous avons utilisé une loi de Poisson tronquée à $k_{\max}$. Nos premières simulations apportent de nombreux résultats encourageants.

Bibliographie

[GAR 03] GAREL B., Asymptotic theory of the likelihood ratio test for the identification of a mixture, soumis, 2003.

[GRE 95] GREEN P., Reversible Jump MCMC Computation and Bayesian Model Determination, *Biometrika*, vol. 82, 1995, p. 711–732.

[LIU 03] LIU X., SHAO Y., Asymptotics for the likelihood ratio test in a two-component normal mixture model, *Annals of Statistics*, vol. 31, 2003, à paraitre.

[MCL 00] MCLACHLAN G., PEEL D., *Finite mixture models*, Wiley-Interscience, New York, 2000.

[RIC 97] RICHARDSON S., GREEN P., On Bayesian analysis of mixtures with an unknown number of components (with discussion), *Journal of the Royal Statistical Society, B*, vol. 59, 1997, p. 731–792.

[ROB 96] ROBERT C. P., *Méthodes de Monte Carlo par chaînes de Markov*, Éditions Économica, Paris, 1996.

Trois approches pour la modélisation des données symboliques de type intervalle

André Hardy et Pascale Lallemand

Unité de Statistique
Département de Mathématique
Facultés Universitaires Notre-Dame de la Paix
8 Rempart de la Vierge
B-5000 NAMUR Belgique

andre.hardy@fundp.ac.be

pascale.lallemand@fundp.ac.be

RÉSUMÉ. *Nous nous intéressons dans ce travail à la modélisation des données symboliques de type intervalle et au problème de la détermination du nombre de classes en classification automatique pour ce type de données symboliques. Nous proposons trois modélisations différentes pour les variables intervalles. Chacune d'entre elles va nous permettre de nous ramener à des données quantitatives classiques. Nous appliquons alors des méthodes de détermination du nombre de classes aux hiérarchies de partitions fournies par quatre méthodes de classification hiérarchiques, et aux ensembles de partitions produites par la méthode de classification symbolique SCLUST. Des ensembles de données artificielles et réelles sont utilisés de manière à tester et à comparer ces trois approches.*

MOTS-CLÉS : *classification, validation, détermination du nombre de classes, données symboliques, variable intervalle.*

1. Les données de type intervalle

Le problème de classification auquel nous nous intéressons est le suivant.

$E = \{x_1,\ x_2,\ ...,\ x_n\}$ est un ensemble de n objets sur lesquels on mesure la valeur de p variables $Y_1,\ Y_2,\ ...,\ Y_p$. Nous recherchons une partition $P = \{C_1,\ C_2,\ ...,\ C_k\}$ de l'ensemble E des objets en k classes.

Nous nous intéressons, dans cet article, à la détermination du nombre de classes pour des données symboliques de type intervalle [BOC 00].

La variable intervalle Y_j, dont l'espace d'observation Y est $\mathcal{R}$, est définie par

$$\begin{array}{rccll} Y_j: & \mathrm{E} & \rightarrow & \mathcal{B}_j & \\ & x_k & \mapsto & \xi_{kj} & := Y_j(x_k) = [\alpha, \beta] \subset \mathcal{R} \end{array}$$

où $\mathcal{B}_j$ est l'ensemble de tous les intervalles fermés et bornés de $\mathcal{R}$.

2. Les méthodes de classification

Avant d'appliquer une méthode de détermination du nombre de classes, il faut disposer, suivant les cas, d'une hiérarchie de partitions, ou tout simplement d'un ensemble de partitions de l'ensemble des données. Il faut alors en extraire la "meilleure" classification. Nous utiliserons tout d'abord une méthode de classification symbolique, SCLUST [BOC 01], [VER 00]. Cette procédure de classification est une généralisation symbolique de la méthode des nuées dynamiques classique [CEL 89]. Les prototypes des classes sont ici les hyperrectangles de gravité de chacune des classes. SCLUST permet de traiter des données symboliques, en utilisant des variables intervalles, multivaluées et modales. Nous nous intéressons également à quatre méthodes de classification hiérarchiques classiques : les méthodes du lien simple, du lien complet, du centroïde et de Ward.

3. Les méthodes de détermination du nombre de classes

Les méthodes choisies sont les meilleures règles d'arrêt analysées dans l'étude de Milligan et Cooper [MIL 85], à savoir la méthode de Caliński et Harabasz, le test de Duda et Hart, l'indice C, l'indice Γ et le test de Beale. La méthode de Caliński et Harabasz, l'indice C et l'indice Γ sont définis à partir des matrices de dispersion intraclasse et interclasse de l'ensemble des données. Les deux autres méthodes sont des tests statistiques. Nous utilisons également le test statistique des Hypervolumes développé à partir de la méthode de classification des Hypervolumes [HAR 96].

4. Modélisation pour des données symboliques de type intervalle

Dans ce travail, nous utilisons et nous comparons trois modélisations pour des objets symboliques de type intervalle. Si nous mesurons sur chaque objet la valeur de p variables intervalles, chacun d'entre eux pourra être représenté par un hyperrectangle dans un espace Euclidien p-dimensionnel.

La première approche [HAR 02] consiste à simuler un processus de Poisson homogène dans chacun des hyperrectangles représentant les objets symboliques.

Dans la deuxième modélisation, nous représentons un intervalle par son centre et sa longueur. Chaque objet symbolique peut alors être représenté par un point dans chacun des espaces bi-dimensionnels associé à chaque variable intervalle.

Dans la troisième modélisation, un intervalle est synthétisé par ses bornes inférieure et supérieure.

Nous nous ramenons ainsi chaque fois à des données quantitatives classiques.

Les quatre méthodes de classification hiérarchiques (lien simple, lien complet, centroïde, Ward) nécessitent le calcul d'une matrice de distances, ou de dissimilarités, entre les objets, à partir de laquelle les critères agglomératifs associés à chaque méthode effectuent les regroupements.

Pour la première approche on pourra utiliser la distance euclidienne classique entre les points simulés.

Pour la deuxième approche, si x et y sont deux objets symboliques décrits par p variables intervalles, notons par

$$x_j = (M_j(x), L_j(x)), \qquad x_j \in R^2 \qquad (j = 1 \ldots, p)$$
$$y_j = (M_j(y), L_j(y)), \qquad y_j \in R^2 \qquad (j = 1 \ldots, p)$$

les représentations Milieu-Longueur associées respectivement aux objets x et y pour chacune des p variables intervalles.

Pour la troisième approche, nous aurons

$$x_j = (Min_j(x), Max_j(x)), \qquad x_j \in R^2 \qquad (j = 1 \ldots, p)$$
$$y_j = (Min_j(y), Max_j(y)), \qquad y_j \in R^2 \qquad (j = 1 \ldots, p)$$

les représentations Minimum-Maximum associées respectivement aux objets x et y pour chacune des p variables intervalles.

Dans les deux derniers cas, la distance entre les objets symboliques x et y, définie par

$$D(x, y) := \sum_{j=1}^{p} d(x_j, y_j)$$

est la somme des p distances entre les représentations Milieu-Longueur (ou Minimum-Maximum) x_j et y_j $(j = 1, \ldots, p)$ des objets x et y. Dans cette expression, d est une distance euclidienne quelconque.

Les hiérarchies de partitions liées à chaque méthode hiérarchique peuvent alors être construites suivant les différents critères agglomératifs.

Les cinq meilleures méthodes de détermination du nombre de classes de Milligan et Cooper peuvent donc être appliquées aux hiérarchies ainsi produites.

Le test des Hypervolumes est basé sur le calcul d'enveloppes convexes de points ; il ne requiert pas la connaissance d'une matrice de distances, mais seulement de la position des points. Ces positions sont connues dans chacune des p représentations Milieu-Longueur (Minimum-Maximum). Nous prenons parmi les p variables celle qui contribue le plus à l'inertie de l'ensemble des objets symboliques. Nous retenons le nombre de classes donné par le test des Hypervolumes associé à cette variable.

La méthode de classification symbolique SCLUST n'est pas une méthode hiérarchique. Parmi les cinq règles d'arrêt de Milligan et Cooper, seuls la méthode de Caliński et Harabasz, l'index C et l'index Γ sont directement applicables aux résultats donnés par des méthodes de classification non hiérarchiques.

5. Exemples

Plusieurs exemples de données artificielles et réelles sont présentés, permettant de comparer les trois approches, et de mettre en évidence les spécificités de chacune d'entre elles.

6. Bibliographie

[BOC 00] BOCK, H.-H., DIDAY, E. (eds) (2000) : *Analysis of Symbolic Data, Exploratory methods for extracting statistical information from complex data*. Studies in Classification, Data Analysis, and Knowledge Organisation, Springer Verlag.

[BOC 01] BOCK, H.-H., DE CARVALHO, F., LECHEVALLIER, Y. , VERDE, R. (2001) : Report of the Meeting ASSO - W6.2 Classification group (Munchen).

[CEL 89] CELEUX. G., DIDAY, E., GOVAERT, G., LECHEVALLIER, Y., RALAMBONDRAINY, H. (1989) : Classification automatique des données, Dunod, Paris.

[DEL 02] DELOGNE, S. (2002) : Méthodes de détermination du nombre de classes pour des données symboliques de type intervalle, Mémoire, FUNDP, Namur.

[HAR 96] HARDY, A. (1996) : On the number of clusters. *Computational Statistics and Data Analysis* 23, 83-96.

[HAR 02] HARDY, A., LALLEMAND, P. (2002) : Determination of the number of clusters for symbolic objects described by interval variables. Studies in classification, data analysis, and knowledge organization, Springer Verlag, 311-318.

[MIL 85] MILLIGAN, G.W., COOPER, M.C. (1985) : An examination of procedures for determining the number of clusters in a data set. *Psychometrika* 50, 159-179.

[NYI 03] NYIRANSENGIMANA M.-R. (2003) : Une nouvelle modélisation pour des données symboliques de type intervalle, Mémoire, FUNDP, Namur.

[VER 00] VERDE, R., de A. T. de CARVALHO, F., LECHEVALLIER, Y. (2000) : A dynamical clustering algorithm for multi-nominal data, Studies in classification, data analysis, and knowledge organization, Springer Verlag, 387-393.

L'algorithme IEM pour données binaires

F.-X. Jollois et M. Nadif

LITA, UFR MIM, Université de Metz
Ile du Saulcy, 57045 Metz, France
jollois@sciences.univ-metz.fr, nadif@iut.univ-metz.fr

RÉSUMÉ. *L'algorithme EM est très populaire pour l'estimation de paramètres d'un modèle de mélange. Un inconvénient majeur de cet algorithme est la lenteur de sa convergence. Afin de pouvoir travailler sur des tableaux binaires de grande taille, nous allons utilisé et étudié le comportement d'une variante de EM appelée "Incremental EM" (IEM) qui permet d'accélérer la convergence.*

MOTS-CLÉS : *Classification, algorithme EM, algorithme IEM*

1. Introduction

L'utilisation des modèles de mélange dans la classification est devenue une approche classique et très puissante (voir par exemple [BAN 93, CEL 95]). En traitant la classification sous cette approche, l'algorithme EM [DEM 77] composé de deux étapes : Estimation et Maximisation est devenue quasiment incontournable. Celui-ci est très populaire pour l'estimation de paramètres. Ainsi, de nombreux logiciels sont basés sur cette approche, comme Mclust-EMclust [FRA 99] ou EMmix [MCL 98]. Ce succès tient à sa simplicité, à ses propriétés théoriques et à son bon comportement pratique. De plus, un intérêt grandissant se fait ressentir actuellement pour les données qualitatives. On peut citer comme exemple le logiciel AutoClass [CHE 96].

Malheureusement, son principal inconvénient réside dans sa lenteur due au nombre élevé d'itérations parfois nécessaire pour la convergence ce qui rend son utilisation inapproprié pour les données de grande taille. Plusieurs versions ont été faites pour accélérer cet algorithme et beaucoup d'entres elles agissent sur l'étape maximisation. Ici nous avons choisi d'étudier une version particulièrement adaptée aux données de grande taille et qui utilise une étape partielle d'estimation au lieu d'une étape Estimation complète. Cette version semble très efficace pour des mélanges Gaussiens, nous proposons ici de l'appliquer sur un modèle de mélange de Bernoulli et de discuter son comportement.

2. Modèle de mélange et algorithme EM

Dans l'approche modèle de mélange, les individus $\mathbf{x}_i, \ldots, \mathbf{x}_n$ à classifier sont supposés provenir d'un mélange de k densités dans des proportions inconnus $p_1, \ldots, p_k$. Ainsi, chaque objet $\mathbf{x}_i$ est une réalisation d'une densité de probabilité (p.d.f.), décrite par :

$$f(\mathbf{x}_i, \theta) = \sum_{k=1}^{s} p_k \phi_k(\mathbf{x}_i; \alpha_k)$$

où $\phi_k(\mathbf{x}_i;\alpha_k)$ représente la densité de $\mathbf{x}_i$ de paramètre α_k. Le vecteur des paramètres à estimer θ est composé de $\mathbf{p} = (p_1,\ldots,p_k)$ et $\alpha = (\alpha_1,\ldots,\alpha_k)$. De ceci, nous en déduisons la log-vraisemblance du vecteur $\mathbf{x} = (\mathbf{x}_i,\ldots,\mathbf{x}_n)$, donnée par :

$$L(\mathbf{x},\theta) = \sum_{i=1}^{n} \log\left(\sum_{k=1}^{s} p_k \phi_k(\mathbf{x}_i;\alpha_k)\right). \quad [1]$$

Dans la suite, nous allons aborder le problème de la classification sous l'approche estimation : les paramètres sont d'abord estimés, puis la partition en est déduite par la méthode du maximum a posteriori (MAP). L'estimation des paramètres du modèle passe par la maximisation de $L(\mathbf{x},\theta)$. Une solution itérative pour la résolution de ce problème est l'algorithme EM [DEM 77]. Le principe de cet algorithme est de maximiser de manière itérative l'espérance de la log-vraisemblance complétée conditionnellement aux données $\mathbf{x}$ et la valeur du paramètre courant $\theta^{(n)}$:

$$Q(\theta|\theta^{(n)}) = \sum_{i=1}^{n}\sum_{k=1}^{s} t_{ik}^{(n)} \left(\log(p_k) + \log\phi_k(\mathbf{x}_i,\alpha_k)\right)$$

où $t_{ik}^{(n)} \propto p_k^{(n)}\phi_k(\mathbf{x}_i,\alpha_k^{(n)})$ est la probabilité conditionnelle a posteriori. Chaque itération de EM a deux étapes :

- **Estimation** : On calcule $Q(\theta|\theta^{(n)})$, notons que dans le contexte modèle de mélange, cette étape se réduit aux calculs des $t_{ik}^{(n)}$.
- **Maximisation** : On cherche la paramètre $\theta^{(n+1)}$ qui maximise $Q(\theta|\theta^{(n)})$.

3. Accélération de EM : IEM

L'algorithme Incremental EM (IEM) [NEA 98] qui est une variante de EM est destiné à réduire le temps de calcul en réalisant des étapes Estimation partielles. Soit $y = y_1,\ldots,y_B$ une partition des données en B blocs disjoints (B est proposé par l'utilisateur). Chaque bloc y_b contiendra $r = n/B$ observations si n est un multiple de B et plus sinon. L'algorithme IEM parcourt tous les blocs de façon cyclique. A chaque itération, on met à jour les probabilités a posteriori d'un bloc dans l'étape Estimation. Ci-dessous, on décrit plus en détail la $n$ième itération :

- **Estimation** : Dans cette étape, on retient un bloc y_b et on met à jour les probabilités a posteriori $t_{ik}^{(n)}$ pour toutes les observations appartenant aux bloc y_b, quant aux autres observations (appartenant aux autres blocs) nous avons $t_{ik}^{(n)} = t_{ik}^{(n-1)}$. L'espérance conditionnelle associée au bloc b notée Q_b est mise à jour, quant à celles associées aux autres blocs elles restent inchangées. Autrement dit la quantité globale qu'on cherchera à maximiser dans l'étape maximisation peut s'écrire :

$$Q(\theta|\theta^{(n)}) = Q(\theta|\theta^{(n-1)}) - Q_b(\theta|\theta^{(n-1)}) + Q_b(\theta|\theta^{(n)}).$$

- **Maximisation** : On cherche comme dans l'algorithme EM classique, le paramètre $\theta^{(n+1)}$ qui maximise $Q(\theta|\theta^{(n)})$.

De cette façon, chaque observation $\mathbf{x}_i$ est visitée après les B étapes d'estimation partielles. Une approximation de la log-vraisemblance ne décroît pas à chaque itération. La justification théorique de cet algorithme a été faite par Neal et Hinton [NEA 98]. Notons que lorsque $B = 1$, IEM se réduit à EM.

4. Illustration sur des données binaires et discussion

Ici, comme nous nous intéressons aux données de type binaire, nous utilisons le modèle de mélange de Bernoulli [CEL 91, GOV 90]. Dans ce cas et sous l'hypothèse d'indépendance conditionnelle, la densité de probabilité d'une observation $\mathbf{x}_i$ peut s'écrire :

TAB. 1. *Temps mis (en secondes) par IEM sur des données simulées, selon le nombre de blocs demandé (1 seul bloc correspond à EM standard).*

	5000x5		10000x10		10000x50	
B	temps	L	temps	L	temps	L
1	307	-16164	706	-62356	989	-291943
2	227	-16163	619	-62356	1005	-291943
4	172	-16163	564	-62356	1122	-291943
8	50	-16163	672	-62356	1388	-291943
16	29	-16163	599	-62356	2086	-291943
32	13	-16473	187	-62357	3358	-291943
64	17	-16473	52	-65591	5216	-291943
128	23	-16472	69	-65593	10291	-291943
256	40	-16474	97	-65587	12525	-291943

$$f(\mathbf{x}_i, \theta) = \sum_{k=1}^{s} p_k \phi_k(\mathbf{x}_i; \alpha_k) = \sum_{k=1}^{s} p_k \prod_{j=1}^{d} (1 - \alpha_k^j)^{1-x_i^j} (\alpha_k^j)^{x_i^j}.$$

En appliquant l'algorithme IEM sur des données simulées, nous avons étudié plusieurs aspects dont le choix du nombre de blocs et la taille des données. Pour illustrer ces différents aspects, nous avons choisi de présenter quelques résultats obtenus à partir de 3 tableaux simulés de taille différente $(5000 \times 5, 10000 \times 10, 10000 \times 50)$.

Dans nos expériences numériques, où l'on a lancé 20 fois chaque algorithme, on initialise les algorithmes EM et IEM à partir des mêmes paramètres choisis au hasard $\theta^{(0)} = (p_1^{(0)}, \ldots, p_s^{(0)}, \alpha_1^{(0)}, \ldots, \alpha_s^{(0)})$. A partir des résultats présentés dans Tab. 1 (temps total pour les 20 essais), nous constatons pour les deux premiers tableaux de données, qu'un trop petit nombre n'apportera pas forcément une accélération de l'algorithme très importante, alors qu'un trop grand nombre demandera plus de calculs intermédiaires, et finalement n'accélérera pas l'algorithme. Aussi nous pouvons relever que IEM accélère nettement l'algorithme, au détriment parfois de la qualité des estimations obtenues (voir la log-vraisemblance L arrondie). D'où l'intérêt de proposer un nombre approprié de blocs en fonction de la taille de l'échantillon. Pour le dernier cas, avec $n = 10000$ et $d = 50$, où le rapport entre n et d est égal à 200 alors que dans les cas précédents, le rapport est égal à 1000, IEM apparaît toujours moins rapide que EM. Ce qui laisse à penser que l'utilisation de IEM n'est profitable que dans certaines situations.

Dans ce travail, à partir des simulations de Monte-Carlo, nous allons étudier ces différents aspects, aussi nous comparerons cet algorithme avec d'autres variantes de IEM : Lazy EM [THI 01] et Sparse EM [NEA 98]. Ces deux algorithmes cherchent à minimiser le nombre de calcul des probabilités a posteriori t_{ik}. Le premier algorithme noté LEM, nous semble le plus performant. Il consiste à identifier les individus présentant une probabilité t_{ik} supérieure à un seuil. Ces individus sélectionnés n'interviendront plus dans les calculs durant un certain nombre d'itérations. Les deux aspects : choix du seuil et le nombre d'itérations seront également discutés.

5. Bibliographie

[BAN 93] BANFIELD J. D., RAFTERY A. E., Model-based Gaussian and non-Gaussian Clustering, *Biometrics*, vol. 49, 1993, p. 803–821.

[CEL 91] CELEUX G., GOVAERT G., Clustering Criteria for Discrete Data and Latent Class Models, *Journal of Classification*, vol. 8, 1991, p. 157–176.

[CEL 95] CELEUX G., GOVAERT G., Gaussian Parcimonious Clustering Methods, *Pattern Recognition*, vol. 28, 1995, p. 781–793.

[CHE 96] CHEESEMAN P., STUTZ J., Bayesian Classification (AutoClass) : Theory and Results, FAYYAD U., PIATETSKY-SHAPIRO G., UTHURUSAMY R., Eds., *Advances in Knowledge Discovery and Data Mining*, AAAI Press, 1996, p. 61–83.

[DEM 77] DEMPSTER A., LAIRD N., RUBIN D., Mixture Densities, Maximum Likelihood from Incomplete Data via the EM Algorithm, *Journal of the Royal Statitical Society*, vol. 39, n^o 1, 1977, p. 1–38.

[FRA 99] FRALEY C., RAFTERY A. E., MCLUST : Software for Model-Based Cluster and Discriminant Analysis, rapport n^{o}342, 1999, University of Washington.

[GOV 90] GOVAERT G., Classification Binaires et Modèles, *Revue de Statistique Appliquée*, vol. 37, 1990, p. 67–81.

[MCL 98] MCLACHLAN G. J., PEEL D., User's guide to EMMIX-Version 1.0, rapport, 1998, University of Queensland.

[NEA 98] NEAL R., HINTON G., A View of the EM Algorithm that Justifies Incremental, Sparse, and Other Variants, JORDAN M., Ed., *Learning in Graphical Models*, 1998, p. 355–371.

[THI 01] THIESSON B., MEEK C., HECKERMAN D., Accelerating EM for Large Databases, rapport n^{o}MSR-TR-99-31, 2001, Microsoft Research.

Chaos Game Representation et Traitement des Séries Temporelles

Gaëlle Legrand, Pierre-Emmanuel Jouve, Nicolas Nicoloyannis

Laboratoire ERIC
Université Lumière Lyon 2
Bâtiment L
5 av. Pierre Mendès-France
69 676 BRON cedex FRANCE
{glegrand ;pjouve ;nicolas.nicoloyannis}@dionysos.univ-lyon2.fr

RÉSUMÉ. *Cet article introduit l'utilisation de la méthodologie Chaos Game Representation (CGR) pour l'étude des séries temporelles. Nous énonçons le principe de construction des CGR et nous présentons une étude illustrative de séries temporelles.*

MOTS-CLÉS : *Séries Temporelles, Chaos Game Representation, Classification, ACP*

1. Introduction

La méthodologie du Chaos Game Representation (CGR) est un puissant outil de visualisation de séries discrètes. En effet, le CGR est une technique qui permet de transformer une séquence unidimensionnelle en une forme graphique bidimensionnelle tout en conservant la structure des sous-séquences présentes dans la série. Ainsi, afin d'obtenir une meilleure représentation et une classification d'un ensemble de séries temporelles, nous avons décidé d'appliquer la méthodologie du CGR aux séries temporelles.

Dans une première partie, nous allons présenter la méthodologie du CGR. Ensuite, nous en décrirons les avantages. Et, dans la dernière partie, nous étudierons la manière dont nous allons appliquer le CGR aux séries temporelles.

2. La méthodologie du CGR

Jeffrey, [JEF 90], [JEF 92], fut l'un des premiers à utiliser le CGR. Il s'en est servi pour étudier les séquences d'ADN. Aussi, pour présenter la méthodologie du CGR, nous allons l'appliquer à la génomique, [FER 01], [EDW 02]. La méthode du CGR permet la représentation et la recherche de sous-séquences au sein de longues séries de lettres. Les sous-séquences locales et globales de la série sont mises en évidence graphiquement et les structures inconnues de la série sont visuellement révélées. La série est visualisée dans une image carré. Dans cette image, chaque pixel est associé à une seule et unique sous-séquence. Pour utiliser cette méthode, il est nécessaire de considérer un alphabet de N lettres, $N \in \aleph^*$. Dans le cas des séries d'ADN, l'alphabet est de taille $N = 4$ et est composé des 4 nucléotides : C, G, T, A. La série est envisagée comme une suite de lettres et est lue lettres par lettres. Ainsi, toutes les sous-séquences présentes dans cette série peuvent être examinées.

L'image est divisée en 4 quadrants délimités par les médianes du carré. Chaque lettre de l'alphabet est associée à un quadrant du carré dans lesquels les sous-sequences se terminant par la lettre correspondante sont situées. Le segment $[GA]$ correspond à l'axe des ordonnées et le segment $[GT]$, à l'axe des abscisses. La longueur du côté du

carré représente l'unité. Le centre du carré a pour coordonnées $(1/2, 1/2)$. Pour chaque lettre de la séquence, un déplacement dans la direction de l'angle associé à cette lettre et de longueur égale à la moitié de la distance entre le point précédent et l'angle considéré est effectué.

Pour obtenir les sous-séquences de deux lettres, il suffit de subdiviser chaque quadrant du carré. Pour obtenir les sous-séquences de trois lettres ou plus il suffit de subdiviser à nouveau les sous-quadrants du carré et ainsi de suite. Chaque sous-quadrant du carré correspond à un sous-séquence particulière comme nous le montre la figure 1. Dans un sous-quadrant, il y a tous les sous-séquences finissant par les mêmes lettres. L'absence d'une sous-séquence dans une série se traduit par le fait que le sous-quadrant lui correspondant est blanc. Les fréquences des mots sont représentées par l'intensité de chaque pixel : plus le pixel est foncé, plus la fréquence est grande.

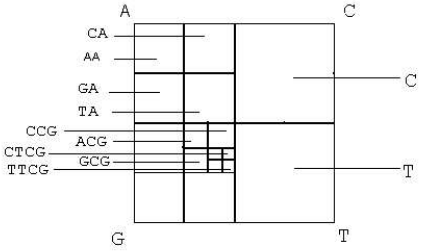


Fig. 1. Visualisation par CGR

Il est, bien sûr, possible d'utiliser la méthodologie du CGR pour des domaines autres que celui de l'analyse des séquences d'ADN. Pour cela, il suffit que la série étudiée soit ou puisse être transformée en une suite de lettres.

3. Application aux séries temporelles

En premier lieu, il convient de coder les séries temporelles : c'est à dire de transformer une série d'événements temporels en une suite d'événements discrets, représentés par des lettres. Pour notre étude, nous procédons de la manière suivante : après avoir défini l'amplitude de la série, le nombre de lettres composant l'alphabet est déterminé. Ainsi, pour un alphabet de N lettres, l'amplitude de la série sera divisée en N morceaux. L'alphabet sera construit par l'affectation d'une lettre à chaque intervalle de l'amplitude de la série. Ensuite, chaque observation de la série initiale correspond à une lettre. Pour illustrer l'application des CGR, nous utilisons un alphabet de taille $N = 4$ et des sous-séquences de longueur $P = 3$. Ce sont des exemples naïfs mais qui nous permettent de représenter efficacement notre propos. Nous avons expérimenté l'application des CGR sur Synthetic Control Chart Time Series, [ALC 99]. Nous avons utilisé trois types de séries : séries normales (oscillation autour d'un point fixe), séries cycliques et séries avec une tendance positive (voir figure 2.). Les séries numérotées de 1 à 9 sont les séries normales, de 10 à 18 les séries cycliques et de 19 à 27 les séries avec une tendance positive. Les figures 3, 4, 5 nous montrent les résultats obtenus par l'application du CGR.

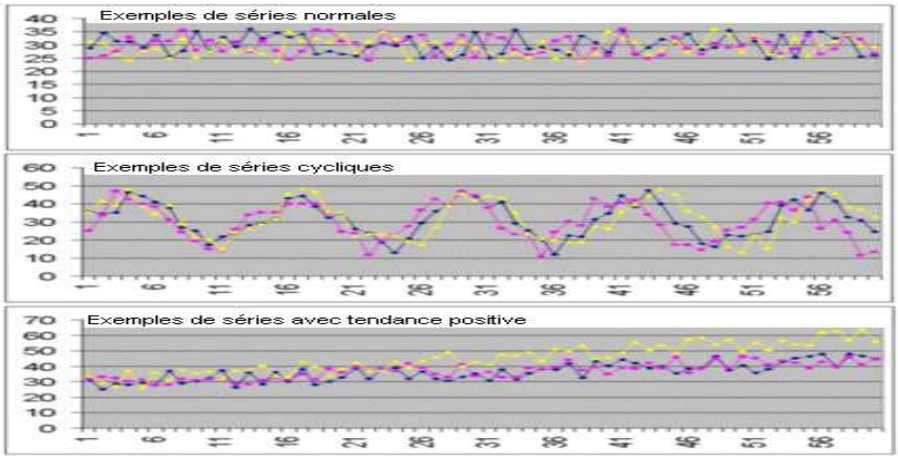


Fig. 2. Exemples de séries utilisées pour l'expérimentation

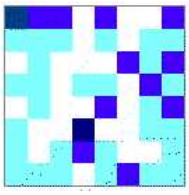


Fig. 3. Visualisation CGR type pour séries normales

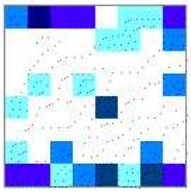


Fig. 4. Visualisation CGR type pour séries cycles

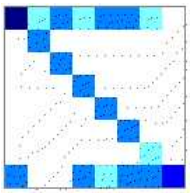


Fig. 5. Visualisation CGR type pour séries avec tendance positive

4. Apports liés à la méthodologie du CGR

Grâce à sa méthode de construction, le CGR possède certaines propriétés très intéressantes lors de son utilisation avec les séries temporelles. Les séries temporelles sont des données séquentielles qui sont naturellement ordonnées par le temps. Or, le CGR permet de conserver l'ordre temporel : à partir des coordonnées d'un point, il est possible de retrouver l'évolution complète de la sous-séquence précédent ce point. Cette propriété permet de ne perdre aucune information sur la structure sous-jacente de la série.

Les études liées à l'extraction et à la localisation des motifs séquentiels présents dans les séries chronologiques est au centre des préoccupations actuelles. Grâce au CGR, il est possible, visuellement, d'appréhender la structure sous-jacente de la série ainsi que la fréquence d'apparition des sous-séquences la composant. Le CGR permet également de classer un ensemble de séries en fonction de leur comportement au cours du temps. En effet, l'étude visuelle de nos CGR nous permet de retrouver les trois catégories de séries. Les séries possédant une tendance positive sont caractérisées ici par la présence d'un Z dans le CGR. Les séries normales ne sont caractérisées par aucune structure particulière. Et, les séries cycliques se caractérisent par deux barres horizontales : une dans la partie haute et une dans la partie basse du CGR.

5. Classification supervisée et non supervisée des types de séries temporelles

Afin de classifier les séries temporelles, nous utilisons une distance Euclidienne qui nous permet d'évaluer la similarité entre deux CGR. Chaque CGR est considéré comme un individu et est associé à un point dans un espace à N^P dimensions (pour notre exemple, $P = 3$ et $N = 4$ donc la dimension de l'espace est 256),[DES 99]. Les variables sont les 256 sous-séquences de 3 lettres possibles. La valeur d'une variable est la fréquence d'apparition de la sous-séquence associée à cette variable. Nous obtenons pour notre exemple des données composées de 27 individus et de 256 variables. Nous utilisons comme méthode de classification une ACP ainsi qu'une classification ascendante hiérarchique. Ces deux méthodes nous permettent de retrouver les trois groupes de séries (normales, cycliques et présence d'une tendance positive). Pour l'ACP, les groupes sont bien dissociés. L'utilisation de cette méthodologie est une alternative à l'utilisation de l'ACP fonctionnelle. La classification ascendante hiérarchique entraîne une partition en trois groupes discriminant parfaitement les types de séries. Ceci n'aurait pas été le cas si nous avions appliqué la classification ascendante hiérarchique sur les données brutes.

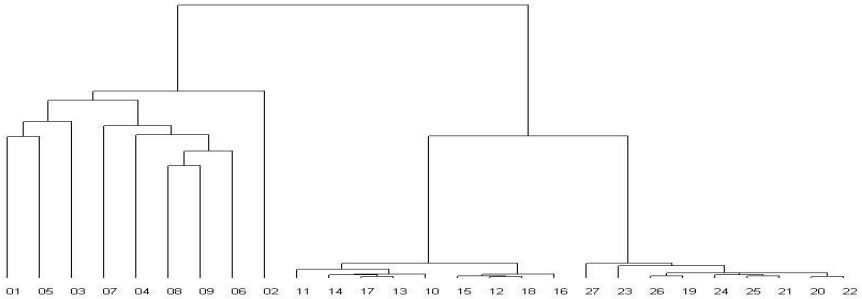


Fig. 6. Classification ascendante hiérarchique

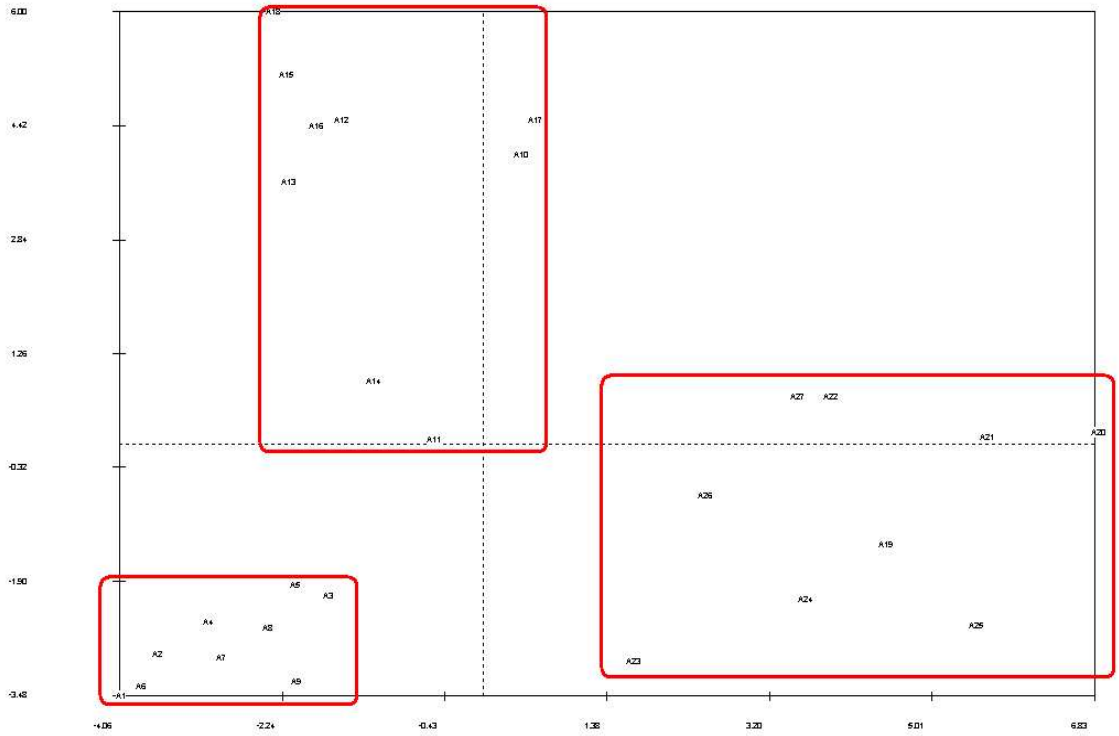


Fig. 7.ACP

Notons qu'il est également possible d'utiliser d'autres méthodes de classification (arbre de décision, réseaux de neurones...).

6. Conclusion

Cette approche est intéressante du point de vue de sa simplicité, de l'information qu'elle nous apporte au sujet des motifs séquentiels et des résultats liés à la classification. Nous envisageons de poursuivre cette étude et de mettre au point des codages de séries plus performants afin de rendre les résultats obtenus plus pertinents. Dans ce but, nous comptons travailler sur la taille de l'alphabet et des sous-séquences étudiées.

7. Bibliographie

[ALC 99] ALCOCK R., MANOLOPOULOS Y., Time-Series Similarity Queries Employing a Feature-Based Approach., *7th Hellenic Conference on Informatics.*, August 27-29. Ioannina,Greece, 1999, p. 167-176.

[DES 99] DESCHAVANNE P. J., GIRON A., VILAIN J., FAGOT G., FERTIL B., Genomic Signature : characterization and Classification of Species Assessed by Chaos Game Representation of Sequences, *Molecular Biology and Evolution*, vol. 16, 1999, page 13911399.

[EDW 02] EDWARDS S., FERTIL B., GIRON A., DESCHAVANNE P., DNA language and the phylogenic position of flightless birds, *Systematic Biology*, , 2002.

[FER 01] FERTIL B., DUFRAINE C., CHAPUS C., GIRON A., DESCHAVANNE P. J., Visualisation et analyse du style du genome, *Proc. XIII congrès de la Soc. Francophone de Classification*, 2001, p. 139-146.

[JEF 90] JEFFREY H. J., Chaos game representation of gene structure, *Nucleic Acids Research*, vol. 18, 1990, p. 2163-2170.

[JEF 92] JEFFREY H. J., Chaos game visualization of sequences, *Comput. and Graphics*, vol. 16, 1992, p. 25-33.

Liens entre critère métrique et critère probabiliste en classification croisée dans le cas continu

Y. Khemal Bencheikh

Département de mathématiques
Faculté des sciences
Université Ferhat Abbas
Sétif 19000, Algérie
bencheikh_00@yahoo.fr

RÉSUMÉ. *Ce travail porte sur les liens qui existent entre les méthodes de classification automatique croisée et les modèles probabilistes lorsque les données sont quantitatives. Pour ceci, nous définissons la notion de critère métrique et de critère probabiliste , nous montrons ensuite qu'un critère probabiliste peut toujours être considéré comme un critère métrique et établissons enfin les conditions pour que la réciproque soit vraie. Ces résultats sont alors appliqués à deux familles de critères métriques : Les premiers sont définis à partir des distances quadratiques, les seconds à partir de la distance L_1 . Cette approche permet de préciser en particulier les différences entre la méthode des distances adaptatives et la méthode de reconnaissance de mélange dans le cas gaussien et de montrer que les critères utilisant la distance en valeur absolue correspond à un mélange de lois exponentielles bilatérales, de proposer de nouveaux critères pouvant améliorer la qualité des résultats.*

MOTS-CLÉS : *Classification automatique croisée, mélange de lois de probabilité, distance quadratique, distance L_1 .*

1. Introduction

L'une des principales difficultés pour les méthodes de classification automatique est le choix du critère et de la métrique utilisée. Lorsqu'il est possible de trouver un modèle de mélange de lois de probabilités tel que l'estimation des paramètres du modèle par l'approche classification ([SCO 71], [SCH 74], [CEL 88] , [GOV 90] , [BEN 92] conduisent à l'optimisation d'un critère numérique de classification, on obtient un éclairage nouveau de ce critère et de la métrique sous jaçente permettant de les justifier ou éventuellement de les rejeter. [GOV 89] s'est intéressé aux liens qui existent entre la classification automatique et les modèles probabilistes lorsque les données mettent en jeu un seul ensemble, nous proposons de le faire ici lorsque les données mettent en jeu deux ensembles ; c'est le cas de la classification croisée. Dans les deux premiers paragraphes, nous définissons deux types de critères et nous étudions dans quelles conditions ces critères peuvent être équivalents. Dans le troisième paragraphe, on fait une application des résultats obtenus aux paragraphes précédent à deux types de métriques, on étudie alors les liens qui existent entre les lois de Gauss et les distances adaptatives, cette approche permet de préciser en particulier les différences entre la méthode des distances adaptatives et la méthode de reconnaissance de mélange dans le cas gaussien [GOV 75] .

2. Définition des deux types de critères

On suppose dans tout ce qui suit que les données initiales sont fournies sous la forme d'un tableau X de n lignes et p colonnes contenant les valeurs prises par n individus pour p variables quantitatives. Ces valeurs seront notées x_i^j,$i = 1, ..., n$ et $j = 1, ..., p$.

2.1. *Critères métriques*

Il s'agit de trouver une partition $(P_1, ..., P_K)$ de l'ensemble I des individus en K classes, une partition $(Q_1, ..., Q_M)$ de l'ensemble J des variables en M classes et un $K.M$-uple (λ_k^m) ;$k = 1, ..., K$ et $m = 1, ..., M$ (un par classe) minimisant le critère suivant :

$$W(P\text{x}Q, L) = \sum_{k=1}^{K} \sum_{m=1}^{M} \sum_{i \epsilon P_k} \sum_{j \epsilon Q^m} D(x_i^j, \lambda_k^m)$$

Ce critère qui dépend de la mesure de dissimilarité D sera appelé critère métrique et noté $CM(R, L, D)$.
$L = \{\lambda_k^m, k = 1, ..., K$ et $m = 1, ..., M\}$.

2.2. *Critères métriques équivalents*

On dira que deux critères métriques sont équivalents si et seulement s'ils sont définis sur les mêmes ensembles R et L et s'il existe une bijection ϕ de R strictement croissante vérifiant :

$$CM(R, L, D_1) = \phi \circ CM(R, L, D_2)$$

où D_1 et D_2 sont les mesures de dissimilarité associées aux deux critères.

Proposition 1 :
$\forall \alpha \epsilon R_+$ et $\beta \epsilon R$ les critères $CM(R, L, D)$ et $CM(R, L, \alpha D + \beta)$ sont équivalents.

2.3. *Critère probabiliste*

On reprend ici la représentation de [BEN 99]. Il s'agit de rechercher une partition PxQ=$\{P_k$xQ^m,$k = 1, ..., K$ et $m = 1, ..., M\}$, K et M étant supposés connus, telle que chaque classe P_kxQ^m soit assimilable à un sous-échantillon qui suit une loi $f(., \lambda_k^m)$. Il s'agit alors de maximiser le critère de vraisemblance classifiante suivant :

$$VC(P\text{x}Q, L) = \sum_{k=1}^{K} \sum_{m=1}^{M} \ln R(P_k\text{x}Q^m, \lambda_k^m)$$

où L est le $K.M$-uple $(\lambda_k^m, k = 1, ..., K$ et $m = 1, ..., M)$ et $R(P_k$x$Q^m, \lambda_k^m)$ est la vraisemblance du sous-échantillon P_kxQ^m qui suit la loi $f(., \lambda_k^m)$.
Ce critère qui dépend de la famille F de fonctions de densités définies sur R sera appelé critère probabiliste et noté $CP(R, F)$.

3. Lien entre les deux types de critères

3.1. *Critères métriques associés à un critère probabiliste*

Proposition 2 :

$$CP(R, F) = -CM(R, L, D)$$

où L est l'ensemble de définition des paramètres de la famille F et D est définie par :

$$\forall x \epsilon R, \lambda \epsilon L \quad D(x, \lambda) = -\ln f(x, \lambda).$$

Le critère métrique ainsi défini est appelé critère métrique associé.

3.2. *Conditions pour qu'un critère métrique soit associé à un critère probabiliste*

Proposition 3 :

Un critère métrique $CM(R, L, D)$ est associé à un critère probabiliste si et seulement si $\forall \lambda \epsilon L$ la fonction $x \rightarrowtail \exp(-D(x, \lambda))$ est continue et vérifie $\int_R \exp(-D(x, \lambda))dx = 1$

3.3. *Critères probabiliste équivalent à un critère métrique*

En utilisant la proposition 1, on peut obtenir une condition plus faible permettant de montrer qu'un critère métrique est équivalent (et non associé) à un critère probabiliste.

Proposition 4 :

Etant donné le critère métrique $CM(R, L, D)$, s'il existe un réel r tel que la quantité :
$s = \int_R r^{(-D(x,\lambda))}dx$ soit indépendante de λ , alors le critère probabiliste $CP(R, F)$ où F est défini par les fonctions de densité f : $f(x, \lambda) = \frac{1}{s} r^{-D(x,\lambda)}$ est un critère équivalent.

4. Application à deux types de métriques

En utilisant les liens proposés dans les deux paragraphes précédents , nous montrons qu'en général les métriques quadratiques sont liées aux lois gaussiennes et que les métriques de type L_1 sont liées aux lois exponentielles bi-latérales. De plus nous avons obtenus de nouveaux critères utilisant des distances adaptatives qu'il serait intéressant de tester. Les résultats obtenus sont résumés dans le tableaux ci dessous.

Métrique utilisée	Lois de probabilités équivalentes
I-Métriques quadratiques	
1- $D(x, \lambda_k^m) = \alpha(x - \lambda_k^m)^2$	$r = e, s = \sqrt{\frac{\pi}{\alpha}}$ et $\sigma^2 = \frac{1}{2\alpha}$ $f(x, \lambda_k^m) = \sqrt{\frac{\pi}{\alpha}} \exp(-D(x, \lambda_k^m)) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{(x-\lambda_k^m)^2}{2\sigma^2})$
2- $D(x, (a_k^m, \alpha_k^m)) = \alpha_k^m(x - a_k^m)^2$ $\lambda_k^m = (a_k^m, \alpha_k^m)$	$s = \sqrt{\frac{\pi}{\alpha_k^m}}$ quantité dépendante de λ_k^m. Il n'existe pas de critère probabiliste équivalent.
3- $D(x, (a_k^m, \alpha_k^m, \beta_k^m)) = \alpha_k^m(x - a_k^m)^2 + \beta_k^m$ $\lambda_k^m = (a_k^m, \alpha_k^m, \beta_k^m)$	Pour $s = 1$ $f(x, (a_k^m, \alpha_k^m)) = \sqrt{\frac{\alpha_k^m}{\pi}} \exp(-\alpha_k^m(x - a_k^m)^2)$
II-Métriques de type L_1	
1- $D(x, \lambda_k^m) = \alpha\|x - \lambda_k^m\|$	$f(x, \lambda_k^m) = \frac{\alpha}{2} \exp(-\alpha\|x - \lambda_k^m\|)$
2- $D(x, (\alpha_k^m, \beta_k^m)) = \alpha_k^m\|x - \beta_k^m\|$	$s = \frac{2}{\alpha_k^m}$ dépend de λ_k^m ; il n'existe pas de critère probabiliste équivalent
3- $D(x, (\alpha_k^m, \beta_k^m, \gamma_k^m)) = \alpha_k^m\|x - \beta_k^m\| + \gamma_k^m$	$si\ \ s = 1\ \ f(x, (\alpha_k^m, \beta_k^m)) = \frac{\alpha_k^m}{2} \exp(-\alpha_k^m\|x - \beta_k^m\|)$

5. Conclusion

Nous venons de voir que la comparaison des critères métriques et probabilistes permet d'apporter un éclairage nouveau de certaines méthodes de classification, de justifier à posteriori certaines contraintes imposées souvent pour des raisons techniques d'optimisation et de proposer de nouveaux critères justifiés utilisant des distances adaptatives, ces dernières permettent d'améliorer la qualité des résultats obtenus, comme dans le cas de données binaires [BEN 02]. Il resterait à considérer le cas où l'ensemble L des noyaux est différent de l'ensemble à classifier d'une part, d'autre part développer et tester les algorithmes de classification croisée correspondant aux nouveaux critères proposés dans ce papier et de les comparer avec les algorithmes classiques utilisant des distances non adaptatives.

6. Bibliographie

[BEN 92] BENCHEIKH Y., Classification Automatique et Modèles, Thèse de Doctorat, Université de Metz, 1992.

[BEN 99] BENCHEIKH Y., Classification Croisée et Modèles, *Rairo Operations Research*, vol. 33, 1999, p. 525-541.

[BEN 02] BENCHEIKH Y., Classification Croisée et Distance L_1 Adaptative, *Revue de Statistique Appliquée*, vol. 03, 2002, p. 53-72.

[CEL 88] CELEUX G., Classification et Modèles, *Revue de Statistique Appliquée*, vol. 04, 1988, p. 43-58.

[GOV 75] GOVAERT G., classification Avec Distance Adaptative, Thèse de Doctorat 3ème Cycle, Université Paris 6, 1975.

[GOV 89] GOVAERT G., modèle de classification et distance dans le cas continue, rapport n^{o}988, 1989, rapport de recherche, INRIA de Paris.

[GOV 90] GOVAERT G., Classification Binaire et Modèles, *Revue de Statistique Appliquée*, vol. 38, 1990, p. 67-81.

[SCH 74] SCHROEDER A., Reconnaissance des Composants d'un Mélange, Thèse de Doctorat 3ème Cycle, Université Paris 6, 1974.

[SCO 71] SCOTT A. E. S., Clustering Methods Based on Likelihood Ratio Criteria, *Biometrics*, vol. 27, 1971, p. 387-397.

Visualisation automatique de trajectoires factorielles pour données évolutives

Christian Koul à Ndjang'ha, Georges Sturbois

Facultés Universitaires Catholiques de Mons (FUCAM)
151, Chaussée de Binche
B – 7000 Mons (Belgique)
e-mail christian.koul@fucam.ac.be

RESUME. Les tableaux de données évolutives (n sujets, p variables, T étapes) mènent à la considération des trajectoires des n sujets. Le logiciel LVT présenté ici, programmé en Visual Basic, exploite un logiciel de base (SPAD, SAS, SPSS) pour obtenir les graphiques individuels de ces trajectoires et s'intéresse à la classification automatique de ces trajectoires. En fonction de la partition adoptée, le logiciel présente la trajectoire de chaque sujet avec celle de la moyenne de la classe à laquelle il appartient. Un affichage ordonné des trajectoires des sujets, regroupés selon les classes de trajectoires, peut également être obtenu.

MOTS-CLES : Données évolutives, trajectoires factorielles, classification des trajectoires.

Le logiciel LVT (Logiciel de Visualisation des Trajectoires) dont on présente ici les différentes étapes de fonctionnement, s'intéresse aux données évolutives décrites sous forme de tableaux $\boldsymbol{X_t}$ successifs *(t=1, 2, …,T)*. Ces tableaux (Fig. 1) de *n* lignes (sujets) et *p* colonnes (variables) peuvent être de composition se prêtant à une analyse en composantes principales (ACP), analyse factorielle des correspondances (AFC) ou analyse des correspondances multiples (ACM).

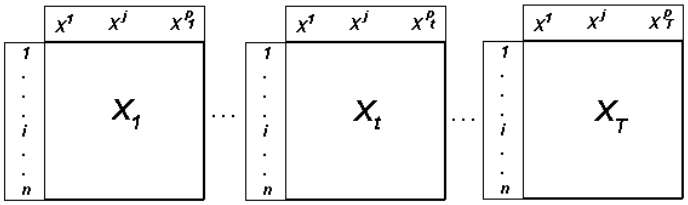


Fig. 1 *Tableaux X_t successifs*

Sur les plans des facteurs F_1-F_2, F_1-F_3, F_2-F_3,… issus de l'analyse factorielle au temps *t=1*, sont projetés en éléments supplémentaires les *n* sujets aux temps $t=2, t=3, \ldots, t=T$. La même opération peut être réalisée à partir de l'analyse factorielle effectuée sur le tableau moyen où sont projetés en éléments supplémentaires les *n* sujets au temps $t=1, t=2, \ldots, t=T$.

A partir du fichier des coordonnées $F_{\alpha,i}^{(k)}$ des *n* sujets sur les q axes factoriels retenus (α = 1, 2,…,q), obtenus via SAS, SPSS ou SPAD, le logiciel LVT saisit automatiquement les données suivantes (Fig. 2) :

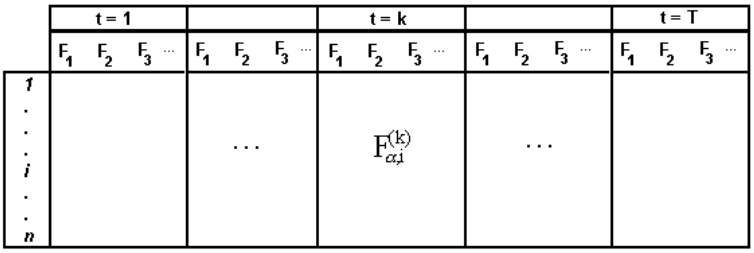
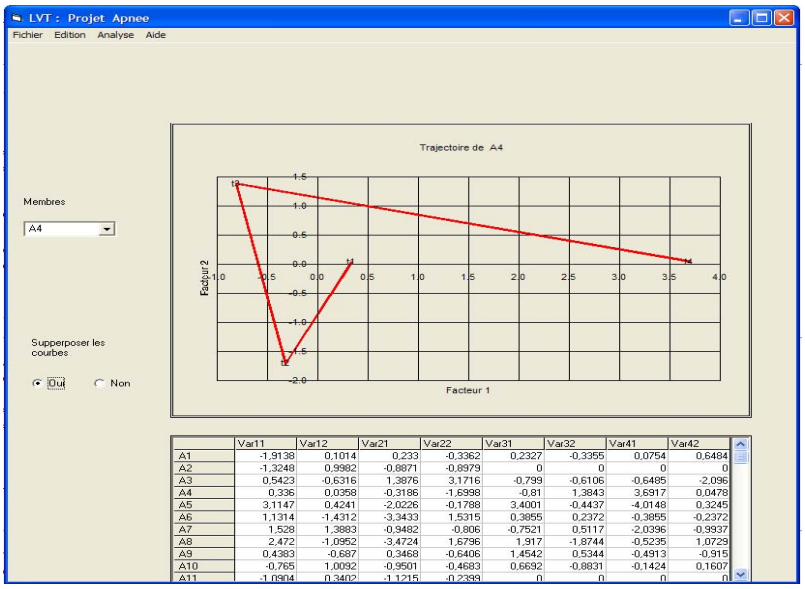


Fig. 2 *Tableaux des coordonnées factorielles pour t=1, 2, …, T*

Sur base de ce fichier, le logiciel fournit pour chaque sujet ou pour une sélection d'entre eux, un graphique individualisé (Fig. 3) de sa trajectoire sur les plans F_1-F_2, F_1-F_3, F_2-F_3,… Ces trajectoires peuvent être regroupées, selon une même échelle, à la mesure de 12 sur un folio A4.



	Var11	Var12	Var21	Var22	Var31	Var32	Var41	Var42
A1	-1,9138	0,1014	0,233	-0,3362	0,2327	-0,3355	0,0754	0,6484
A2	-1,3248	0,9982	-0,8871	-0,8979	0	0	0	0
A3	0,5423	-0,6316	1,3876	3,1716	-0,799	-0,6106	-0,6485	-2,096
A4	0,336	0,0358	-0,3186	-1,6998	-0,81	1,3843	3,6917	0,0478
A5	3,1147	0,4241	-2,0226	-0,1788	3,4001	-0,4437	-4,0148	0,3245
A6	1,1314	-1,4312	-3,3433	1,5315	0,3855	0,2372	-0,3855	-0,2372
A7	1,528	1,3883	-0,9482	-0,806	-0,7521	0,5117	-2,0396	-0,9937
A8	2,472	-1,0952	-3,4724	1,6796	1,917	-1,8744	-0,5235	1,0729
A9	0,4383	-0,687	0,3468	-0,6406	1,4542	0,5344	-0,4913	-0,915
A10	-0,765	1,0092	-0,9501	-0,4683	0,6692	-0,8831	-0,1424	0,1607
A11	-1,0904	0,3402	-1,1215	-0,2399	0	0	0	0

Fig. 3 : *Trajectoire du sujet 4 sur le plan (F1 – F2)*

On s'intéresse ensuite particulièrement à une classification des *n* trajectoires obtenues. La méthode adoptée ici a été proposée par Carlier [CAR 90] et a bénéficié des travaux conduits au GERI[1] par Dazi et Le Barzic [DAZ 96].

Le tableau des données qui sert de base à la classification est établi à partir des coordonnées-positions $F_{\alpha,i}^{(t)}$ et des coordonnées-évolutions $E_{\alpha,i}^{(t)}=F_{\alpha,i}^{(t)}-F_{\alpha,i}^{(t-1)}$ pour $t=2, t=3,...,t=T$. Il s'agit de rechercher une classification des trajectoires des sujets prenant en compte les positions et les évolutions de manière équilibrée.

L'approche utilisée consiste à prendre comme premières coordonnées celles du point de départ de la trajectoire, c'est à dire $F_{\alpha,i}^{(1)}$. Les coordonnées à chaque étape suivante ($t=2, t=3,...,t=T$) sont soit les coordonnées-positions soit les coordonnées-évolutions. On note $a_t = 1$ si l'on considère les coordonnées–positions à l'étape t et $a_t = 0$ si l'on considère les coordonnées–évolutions à l'étape t; on a toujours $a_1 = 1$. Ainsi $2^{(T-1)}$ configurations de données sont envisageables.

Pour une configuration donnée, l'inertie I du nuage des n sujets décrits par ces coordonnées et dotés des poids p_i se décompose en deux termes I_1 et I_2 où I_1 est la part de l'inertie I due aux coordonnées–positions et I_2 la part due aux coordonnées–évolutions.

$$I_1 = \sum_{i=1}^{n}\left[\sum_{\alpha=1}^{q}\sum_{\{t/a_t=1\}}^{T} p_i\left(F_{\alpha,i}^{(t)} - \overline{F_\alpha}^{(t)}\right)^2\right] \qquad I_2 = \sum_{i=1}^{n}\left[\sum_{\alpha=1}^{q}\sum_{\{t/a_t=0\}}^{T} p_i\left(E_{\alpha,i}^{(t)} - \overline{E_\alpha}^{(t)}\right)^2\right]$$

Si l'on souhaite que l'influence des coordonnées–positions soit du même ordre d'importance que l'influence des coordonnées–évolutions, la configuration recherchée doit être telle qu'elle minimise la valeur absolue de la différence entre I_1 et I_2.

Sur base du tableau des données correspondant à ce compromis, une classification des trajectoires est opérée via la méthode de Ward. Elle conduit via SPAD, SAS ou SPSS, à obtenir la représentation arborescente qui décrit les classes des trajectoires proposées.

La partition jugée la meilleure est choisie par l'utilisateur ou automatiquement (par SPAD p. ex.). L'homogénéité des classes obtenues est ensuite optimisée par réaffectations (opération de consolidation). Le logiciel LVT dispose à ce stade via les résultats de SPAD, SAS ou SPSS, d'un nouveau fichier où, à chaque sujet i, est associé le numéro de la classe à laquelle il appartient.

[1] Groupe d'Etude et de Réflexion Interrégional – Rue Pasquier, 31 – 75008 Paris.

A partir de ce fichier, le logiciel établit les coordonnées de la trajectoire moyenne de chaque classe et en présente les graphiques (Fig. 4)

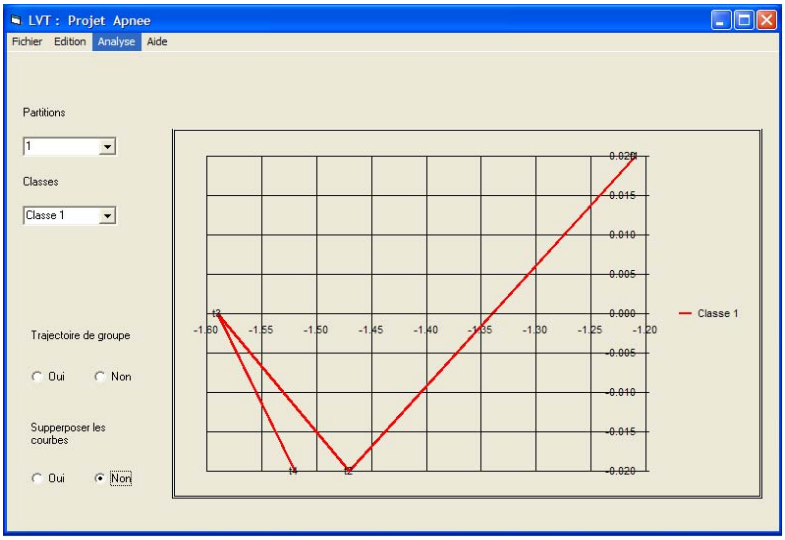


Fig. 4: *Trajectoire moyenne, sur le plan (F1 – F2), pour la classe 2 (partition en 3 groupes)*

Le logiciel LVT permet finalement d'obtenir les trajectoires de chaque sujet avec superposition de la trajectoire moyenne de la classe de trajectoires à laquelle le sujet appartient (Fig. 5). Il permet d'afficher automatiquement la série ordonnée des graphiques relatifs aux sujets de chaque classe, ce qui facilite la comparaison et la perception des trajectoires.

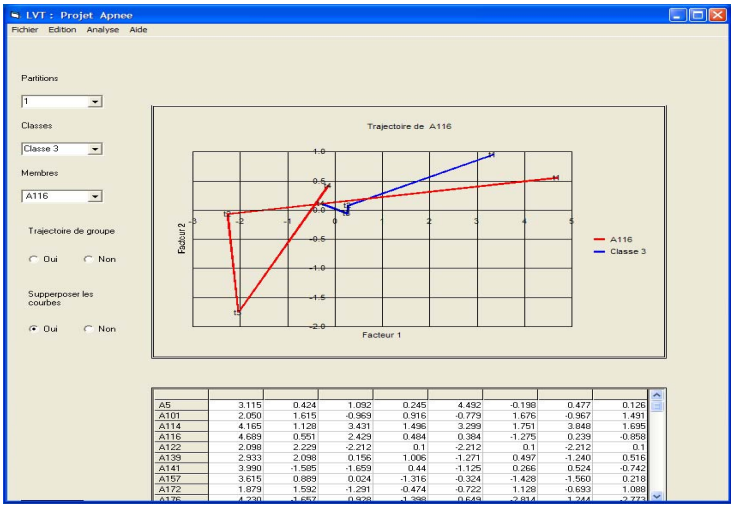
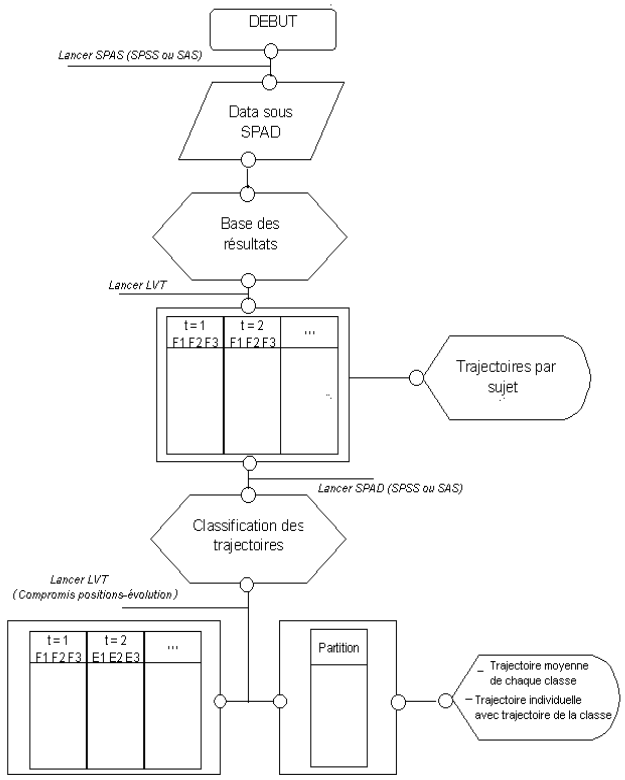


Fig. 5: *Trajectoire du sujet 116, sur le plan (F1 – F2), avec la trajectoire moyenne de la classe 3*

La succession des opérations effectuées par le logiciel LVT – qui a l'originalité par rapport aux autres logiciels de présenter les trajectoires des sujets, non pas superposées sur un même graphique, mais séparément et de s'intéresser à la représentation de la classification des trajectoires – est résumée sur le schéma disposé en annexe.

Le logiciel LVT dont on suit ici une version initiale, fera l'objet de dévelopements au niveau de l'interface graphique et l'interface de programmation pour obtenir une interaction totalement automatique avec les logiciels SAS et SPSS.

Annexe La succession des opérations effectuées par le logiciel LVT.



Références

[CAR 90], CARLIER A., *Factor analysis of evolution and cluster methods on trajectories* - Paul Sabatier University, Toulouse.

[DAZ 96], DAZI F., LE BARZIC J.F., *L'analyse des données évolutives: méthodes et applications*, Technip, 1996.

[JAM 99], JAMBU M.,*Méthodes de base de l'analyse des données*, Eyrolles,1999.

[LAV 88], LAVIT C., *Analyse conjointe de tableaux quantitatifs*, Masson, 1988.

[LEB 95], LEBART L., MORINEAU A., PIRON M., *Statistique exploratoire multidimensionnelle* , Dunod, 1995.

[RUS 99], RUSSO M., ECHOLS M., *Automating Science and Engineering Laboratories with Visual Basic* , Wiley Inter-Science, 1999.

Classification automatique : Applications au Web Mining

Yves Lechevallier[1], Doru Tanasa[2], Brigitte Trousse[2], Rosanna Verde[3]

[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]INRIA-Institut National de Recherche en Informatique et en Automatique, Sophia Antipolis- B.P.93, 2004 Route des Lucioles, 06902 Sophia Antipolis Cedex, France, {Doru.Tanasa, Brigitte.Trousse}@inria.fr
[3]Dip. Strategie Aziendali e Metodologie Quantitative - SUN – Seconda Università di Napoli, Piazza Umberto I, 81043 Capua, Italie rosanna.verde@unina2.it

RÉSUMÉ Dans ce travail nous présentons une approche classificatoire appliquée aux données du Web Usage Mining..
MOTS-CLÉS :Classification automatique, classe, Web Usage Mining.

1. Introduction

Le développement du Web a entraîné au cours de ces dernières années une explosion des données liées à son activité. Pour analyser ce nouveau type de données, sont apparues de nouvelles méthodes d'analyse regroupées sous le terme Web Mining dont les trois axes de développement actuels sont les suivants :

- *Web Content Mining* : analyse textuelle avancée, intégrant les particularités du Web telles que les liens hypertextes et la structure sémantique des pages,
- *Web Structure Mining* : analyse de la structure de liens hypertextes de pages Web en vue d'une catégorisation des pages et sites Web et/ou une classification de sites Web,
- *Web Usage Mining* : analyse des comportements de navigation

La création de sites de grande taille nécessite de prendre en compte la navigabilité du site du point de vue de l'utilisateur. L'étude de parcours à partir des logs issus de fichiers serveurs ou de traces propriétaires peut aider le responsable du site Web à repenser la structure et l'ergonomie du site pour repérer les problèmes des utilisateurs et améliorer la navigabilité.

2. À nouveau champ, de nouvelles structures de données

Les données analysées par le Web Usage Mining proviennent aujourd'hui principalement des fichiers « log http ». La structure du site (graphe des liens hypertexte) et l'information sur les utilisateurs du site (leurs profils) peuvent constituer d'autres sources supplémentaires d'information.

2.1 Présentation des fichiers log HTTP

Suivant le protocole client-seveur http, le poste client qui souhaite accéder à une ressource va émettre une requête adressée au serveur et contenant l'adresse d'allocation de la ressource :

```
GET http://www.inria.fr/accueil.html
```

A l'autre bout, le serveur de l'INRIA interprète la requête http, accède à la ressource demandée et la retourne au client. Comme dans la plupart des programmes informatiques, l'ensemble des opérations effectuées par le serveur sont enregistrées dans des fichiers « log » qui permettent de disposer d'une trace détaillée de l'activité du serveur. Nous utilisons le format de logs HTTP le plus répandu, l'ECLF (*Extended Common Log Format*) [LUO 95].

2.2 Pré-traitement des fichiers log HTTP : Nettoyage des données

Le nettoyage des données pour les fichiers log Web consiste à supprimer les requêtes inutiles de fichiers « log ». Ces requêtes concernent souvent les images et les fichiers multimédia. L'identification de robots Web et la suppression des requêtes provenant de ces robots sont les autres tâches de cette étape. Pour plus de détails concernant l'étape de prétraitement de fichier log HTTP le lecteur intéressé peut se rapporter à [TAN 03].

2.3. Difficultés de construction des sessions

L'unité d'analyse étant la séquence de pages visualisées et non la simple requête, il est préalablement nécessaire de regrouper les requêtes contenues dans les fichiers log pour reconstituer les sessions de visites. Bien que cette tâche paraisse à première vue assez aisée, l'analyste est confrontée à un certain nombre de problèmes techniques.

- *Identification des utilisateurs :*

Pour regrouper les requêtes, il est nécessaire de savoir quels utilisateurs les ont émises. Si l'utilisateur a accepté de s'enregistrer et s'identifie avec un login, alors le repérage est immédiat, mais cela ne concerne qu'une très faible minorité des visites. Une autre méthode répandue mais nécessitant l'acceptation de l'utilisateur, consiste à écrire dans la mémoire du navigateur, c'est-à-dire sur le poste client, un fichier d'identification nommée *cookie* qui sera réutilisé dans chacune des requêtes et permettra au serveur d'en identifier la provenance. En fait on ne dispose que de l'adresse IP qui est identique pour tous les utilisateurs partageant un même router ou pour ceux accédant à l'Internet via le même serveur proxy. Dans ce cas il est difficile de parler d'identification d'utilisateur.

- *Identification de sessions :*

Dans le cas où l'utilisateur aurait été identifié par une des deux premières méthodes décrites plus haut toutes les requêtes qui proviennent de cet utilisateur constitueront sa session. Dans les autres cas nous considérons le couple ip/agent pour construire les différentes sessions.
Le début de session est défini par le fait que la provenance de l'utilisateur (URL enregistrée dans le referrer) est extérieure au site. Par contre, aucun signal n'indique la déconnexion du site, ce qui pose un problème pour déterminer la fin des sessions. Les critères proposés sont en fait des seuils temporels d'inactivité allant de 25-30 minutes à 24 heures.

- *Reconstitution des parcours :*

Après avoir déterminé le début et la fin de la session et avoir filtré les requêtes auxiliaires, reste à reconstituer l'ordre chronologique de visualisation des pages sur le site, c'est-à-dire le *parcours* du visiteur. En effet, si on veut étudier des séquences de pages vues, et non simplement leurs associations il faut tenir compte du caractère séquentiel des sessions étudiées. Tant que la page de provenance (referrer) correspond à la page précédemment visualisée, le tracé du parcours sur le site est aisé. Une confusion peut cependant intervenir du fait que les dernières pages visualisées sont stockées dans la mémoire du navigateur et que par conséquent, lorsque le visiteur repasse par des pages précédemment visualisées, le poste client n'adresse aucune requête au serveur. Dans ce cas fréquent, il est nécessaire de *lire entre les lignes* du fichier « log » pour reconstituer le segment de parcours non enregistré.

3. Classification

Après les phases de nettoyage et de transformation de données qui permettent de construire un tableau de description des sessions nous abordons la phase d'analyse. L'objectif de cette phase est de découvrir différents comportements d'utilisateurs ou des catégories de comportement de navigation par diverses approches [SÄU 01] : Analyse des séquences fréquences, segmentation, modèle prédictifs, réseau neuronal et classification automatique.

3.1. Travaux existants

Il existe des nombreuses méthodes de classification utilisées dans la fouille de données. Cependant, peu de méthodes ont été appliquées aux données du Web : BIRCH dans [FU 99], CLIQUE dans [PER 98], EM dans [CAD 00] car il est difficile, voire impossible, d'adapter certaines méthodes aux particularités des données Web compte tenu de la taille de ces tableaux tant pour les sessions que pour les pages différentes.
Dans [MOB 02] les auteurs considèrent deux méthodes de classification, mais qui ne prennent pas en compte l'ordre des *pages* dans les sessions. Dans [FU 99] les sessions des utilisateurs sont généralisées en utilisant une induction, basée sur les attributs, qui réduit la dimension des données. Les pages sont organisées par une structure hiérarchique liée à l'adresse physique de la page Web. Les données ainsi généralisées sont classées en utilisant un algorithme efficace BIRCH de classification hiérarchique, introduit par [ZHA 96]. Une classification non-supervisé basée sur un réseau de neurones est utilisée dans [BEN 03] pour grouper les sessions similaires (issues d'un site annuaire thématique) en classes.

3.2. Notre approche

Nous aborderons uniquement l'aspect classification automatique et conceptuel, dans ce cas il s'agit de structurer l'ensemble des sessions ou l'ensemble des pages en typologies afin de dégager des comportements similaires et de les identifier. Cependant dans le cas où les objectifs sont définis par des groupes de pages (*rubriques*) nous proposons de modéliser chaque objectif par un objet symbolique. Les algorithmes utilisées sont de type Nuées Dynamiques [DID 71]. Ils recherchent simultanément une partition P de E en k classes et un vecteur L de k prototypes minimisant :

$$\Delta(P^*, L^*) = Min\left\{\Delta(P, L) \,\middle|\, P \in P_k, L \in D^k\right\}$$

avec P_k l'ensemble des partitions de E en k classes non vides. Ce critère Δ exprime l'adéquation entre la partition P et le vecteur des k prototypes. Il est souvent défini comme la somme des distances entre tous les objets s de E et le prototype g_i de la classe C_i la plus proche. L'algorithme procède, alternativement par une étape de représentation suivie d'une étape d'allocation.

Classification des sessions

La phase de classification de l'ensemble des sessions sera suivie par la description des classes et leur positionnement sur un plan factoriel. A partir des ces descriptions nous montrerons comment les modéliser sous la forme de concepts ou bien mettre en œuvre des arbres de décision afin d'obtenir pour chaque classe des règles.

Classification symbolique

A partir de la variable « referer » on peut définir par des requêtes des objets symboliques qui représentent la description de cette classe qui est identifiée à partir d'une connaissance experte. Dans ce cas nous devons utiliser des méthodes de classification, développées dans le cadre de l'analyse symbolique qui s'applique sur des variables multivaluées. Le concept de *prototype* est ici un modèle de représentation d'une classe et il servira à la construction d'indicateurs d'interprétation de ces classes.

4. Perspectives

Une des limites de toutes ces techniques de Web Usage Mining est qu'elles sont difficilement interprétables du fait qu'elles ne décrivent les parcours qu'en termes de noms de documents HTML principalement connus par les concepteurs du site. Un marquage sémantique des pages faciliterait donc la lecture des résultats obtenus, et pourrait intervenir dans la conception même de ces outils de Data Mining.

5. Références

[BEN 03] BENEDEK A., TROUSSE B., « Visualization Adaptation of Self-Organizing Maps for Case Indexing », In *27th Annual Conference of the Gesellschaft fur Klassifikation*, Cottbus, Germany, 12-14 mars 2003.

[CAD 00] CADEZ I. V., HECKERMAN D., MEEK C., SMYTH P., AND WHITE S., « Visualization of navigation patterns on a web site using model-based clustering », In *Proceedings of the sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 280 – 284, Boston, Massachusetts, 2000.

[DID 71] DIDAY, E « La méthode des Nuées dynamiques ».Rev. *Stat. Appliquée*, Vol XIX, p19-34, 1971.

[FU 00] FU Y., SANDHU K., SHIH M., « A generalization-based approach to clustering of web usage sessions », In *Proceedings of the 1999 KDD Workshop on Web Mining*, San Diego, CA, vol. 1836 of LNAI, pag 21 – 38, Springer, 2000.

[LOU 95] LUOTONEN A., « The Common Logfile Format », http://www.w3.org/Daemon/User/Config/Logging.html, 1995.

[MOB 02] MOBASHER B., DAI H., LUO T., AND NAKAGAWA M., « Discovery and evaluation of aggregate usage profiles for web personalization », *Data Mining and Knowledge Discovery*, 6(1):61 – 82, janvier 2002.

[PER 98] PERKOWITZ M., ETZIONI O., « Adaptive web sites: Automatically synthesizing web pages », In *AAAI/IAAI*, pages 727 – 732, 1998.

[SÄU 01] SÄUBERLICH F., HUBER K.-P., « A Framework for Web Usage Mining on Anonymous Logfile Data », SAS Institue GmbH, 2001.

[TAN 03] TANASA D., TROUSSE B., « Le prétraitement des fichiers log Web dans le Web Usage Mining Multi-sites », In *Journées Francophones de la Toile*, juin – juillet 2003.

[ZHA 96] ZHANG T., RAMAKRISHNAN R., LIVNY M., « Birch: An efficient data clustering method for very large databases », In H. V. Jagadish and Inderpal Singh Mumick, editors, *Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data*, Montreal, Quebec, Canada, juin 4-6, 1996, pages 103 – 114, ACM Press, 1996.

Clusters d'ensembles de données larges dans le Web Log Mining

Gabriella Schoier et Giuseppe Melfi

Dipartimento di Scienze economiche e statistiche
Università di Trieste
Piazzale Europa, 1
I-34127, Trieste, Italie
Gabriella.Schoier@econ.units.it

Groupe de Statistique
Université de Neuchâtel
Espace de l'Europe, 4
CH-2002, Neuchâtel, Suisse
Giuseppe.Melfi@unine.ch

RÉSUMÉ. *Nous présentons une solution du problème de l'identification de clusters denses dans l'analyse de données concernant les accès à l'internet d'un ensemble d'utilisateurs. L'algorithme utilisé ici est une modification d'un algorithme proposé pour un problème de même nature concernant les reseaux sociaux.*

MOTS-CLÉS : *Data Mining, Web Log Mining, Clusters*

1. Introduction

Il y a une compétition intense entre les sociétés basés sur l'Internet pour acquérir de nouveaux clients et retenir les clients existants ; pour cette raison la personnalisation dans le web est devenue une partie indispensable de l'e-commerce. En particulier la personnalisation basée sur le Web Usage Mining ou le Web Log Mining, développée pour extraire des modèles intéressants dans les accès sur le Web a plusieurs avantages sur les techniques plus traditionnelles [SRI 00, MOB 02]. Considérons une série finie d'unités (les adresses I.P. des ordinateurs des utilisateurs) sur lesquelles deux variables relationelles ont été mesurés (ayant visité au moins M pages en commun ; étant resté le même intervalle de temps sur la même page) ; ceci forme un réseau N (la série d'unités et de relations associées) [WAS 94]. Dans le but d'analyser un tel réseau on peut considérer les résultats dérivant de deux théories de réseaux sociaux classiques : la théorie *small-world* [KOC 89] et la théorie *peer influence* [FRI 98]. La première a montré qu'il y a un haut degré de clustering local dans les réseaux, donc une approche pour étudier la structure de grands réseaux impliquerait l'identification de clusters locaux et l'analyse des relations à l'intérieur et entre les clusters. La seconde a prouvé que, en se basant sur un procédé d'influence endogène, les unités proches ont une tendance à converger sur des attitudes similaires et ainsi les clusters dans un réseau small-world doivent être similaires le long de multiples dimensions.

Dans ce papier nous présentons une solution au problème d'identification de clusters denses dans l'analyse des enregistrement d'accès au web, en considérant une modification d'un algorithme connu de l'analyse de réseaux sociaux [MOO 01]. L'avantage de cette approche est une structure réduite et plus flexible sur laquelle des techniques différentes telles que le blockmodelling [SCH 02] peuvent être utilisées. Nous comparons aussi les résultats de l'algorithme avec ceux de Batagelj et Mrvar [BAT 02a, BAT 02b] basée sur la méthode k-core.

2. Sur l'identification de clusters denses dans les données de Web Usage Mining

Le point de départ de l'analyse sont les fichiers d'enregitrement d'accès (web access logs) d'utilisateurs du site web réel, *www.girotondo.com*, un portail pour les enfants. Dans ce site il y a sept rubriques différentes : *Bacheca (Lanterne)*, *Corso (Cours)*, *Favolando (Fables)*, *Giochi (Jeux)*, *Links (Liens)*, *News (Nouvelles)*, *Percome (Comment)*, et il y a 362 pages de jhtml. La période d'observation est du 29/11/2000 au 18/01/2001.

Un tel fichier presente les données dans une forme brute. Dans la Table 1, un extrait est presenté.

Table 1 - Fichiers d'enregistrement d'accès

130.93.25.19	20/DEC/2000 :10 :19 :44+0100	"GET/mappa/01.jhtml HTTP/1.0"	200	2472	Mozilla/4.0
235.58.54.78	20/DEC/2000 :10 :19 :41+0100	"GET/news/archivio.jhtml HTTP/1.0"	200	115	Mozilla/4.0
267.12.83.56	20/DEC/2000 :10 :19 :40+0100	"GET/news/01/01/01.jhtml HTTP/1.0"	200	793	Mozilla/4.0
241.27.83.61	20/DEC/2000 :10 :19 :37+0100	"GET/favolando/01.jhtml HTTP/1.0"	200	88	Mozilla/4.0

Les fichiers d'enregistrement d'accès de serveur contiennent : le nom du domaine (ou l'adresse I.P.) de la demande ; la date et le temps de la demande ; la méthode de la demande (GET ou POST) ; l'URL de la page demandée ; le résultat de la demande (le succès, l'échec, l'erreur, etc.) ; la taille des données du fichier ; l'identification de l'agent client.

Une entrée dans le fichier des accès est automatiquement ajoutée à chaque fois qu'une demande pour une ressource atteint un serveur. Les enregistrements de fichiers contenant de l'information de n'importe quel objet (avec extension .gif, .jpeg, etc.) qui n'est pas une adresse internet est annulée pour obtenir ainsi un nouveau fichier. De cette façon nous avons un fichier indiquant l'adresse d'Internet pour chaque page visitée. Nous avons ensuite éliminé les pages visitées par moins de cinq adresses I.P. et ainsi 117 pages ont été considérées. Après le prétraitement un fichier de 1000 enregistrements a été utilisé. Les données consistent en une série d'adresses I.P. sur lesquelles deux variables relationelles (ayant visité au moins $M = 35$ pages en commun, étant resté le même intervalle de temps sur la même page pour des intervalles de temps fixés à l'avance) ont été mesurées ; pour chacune de ces deux variables les données sont représententées dans une matrice à deux modes (les adresses I.P. $\times$ pages). Les deux matrices sont changées en une matrice à un mode (I.P. $\times$ I.P.) en utilisant le programme UCINET [BOR 99] ; la matrice qui en résulte, appelée matrice des adiacences, est composée de zéros et de uns. Les coefficients de la matrice sont 1 si les utilisateurs (assimilés aux adresses I.P.) ont visité au moins 35 pages en commun et sont resté au moins 30 minutes dans les pages visitées en commun, 0 autrement (voir Table 2).

Table 2 - Matrice des adiacences

	138.222.202.11	151.15.169.130	151.2.15.154	.....
138.222.202.11	-	0	1	...
151.15.169.130	0	-	0	...
151.2.15.154	1	0	-	...
.....	...	...	...	

Maintenant nous introduisons une matrice $N \times m$ des influences, Y, où pour chacune des N adresses I.P. (lignes) corresponds un vecteur à m composantes qui décrit les influences. La matrice Y a autant de lignes que les adresses I.P. d'utilisateurs et un nombre de colonnes correspondant au nombres d'influences directes auxquelles chaque individu est sujet. Pour l'exemple qui suit nous avons décidé de utiliser une matrice Y à trois colonnes ($m = 3$). Ceci corresponds à assumer que chaque utilisateur peut avoir jusqu'à trois influences, en négligéant d'ultérieures éventuelles influences.

Pour construire cette matrice nous utilisons une version modifiée de l'algorithme de la moyenne du voisinage récursif (Recursive Neighbourhood Mean algorithm, RNM) proposé par Moody [MOO 01] et écrit en SAS. La modification, (Modified Recursive Neighbourhood Mean algorithm, MRNM) consiste dans le calcul pondéré de la moyenne après un certain nombres d'itérations et generalize l'algorithme RNM. L'algorithme peut être décrit comme suit :

1) Assigner à chaque adresse I.P. dans le réseau un nombre aléatoire uniforme entre 0 et 1 pour chacune des m variables. On obtient ainsi une matrice $Y^{(0)}$ $(N \times m)$ de nombres aléatoires.

2) La matrice $Y^{(t+1)}$ est définie par la formule

$$Y_{ik}^{(t+1)} = \frac{\sum_{j \in L_i} Y_{jk}^{(t)} N_{ij}}{\sum_{j \in L_i} N_{ij}} \qquad k = 1, \ldots, m, \qquad i = 1, \ldots, N,$$

où L_i est le sous-ensemble de $1, \ldots, N$ correspondant aux adresses I.P. qui sont en relation avec i, et N_{ij} est le nombre de pages en commun visitées par i et j.

3) Répéter le pas 2) n fois.

REMARQUE. — Pour $N_{ij} = 1$ pour tout $i, j = 1, \ldots, N$, l'algorithme ci-dessus corresponds à l'algorithme RNM classique [MOO 01].

Cette procédure demande dans l'input, la liste des adjacences, c'est-à-dire les paires de points entre lesquels une relation existe. A ce point l'algorithme RNM modifié est appliqué. Le résultat est la matrice Y.

Dans une situation idéale, $Y = \lim_{n \to \infty} Y^{(n)}$. Toutefois dans notre cas $n = 7$ a suffit pour obtenir des resultats tout à fait satisfaisant.

Sur les trois variables de position une "Ward's minimum variance cluster analysis" est exécuté. De telle façon nous obtenons un clustering clair qui révèle une structure de trois clusters entre les unités appartenant au réseau.

Table 3 - Table des résultats

I.P.	cluster	var1	var2	var3
1	1	0.48816	0.42557	0.53592
2	3	0.48822	0.42593	0.53589
3	1	0.48816	0.42557	0.53592
...	.....	.....	.....	.....

Le premier cluster, le plus nombreux est formé par les adresses I.P. qui ont une haute fréquence de relations ; le second est identifié par les adresses I.P. qui n'ont pas beaucoup de relations tandis que le tiers par les adresses I.P. qui ont peu de relations. Deux adresses I.P. ne sont classés nulle part, ce qui correspond à ce que l'on s'attendait car il s'agit des addresses I.P. du webmaster du site.

Pour la visualisation du réseau le programme PAJEK [BAT 02a] a été appliqué

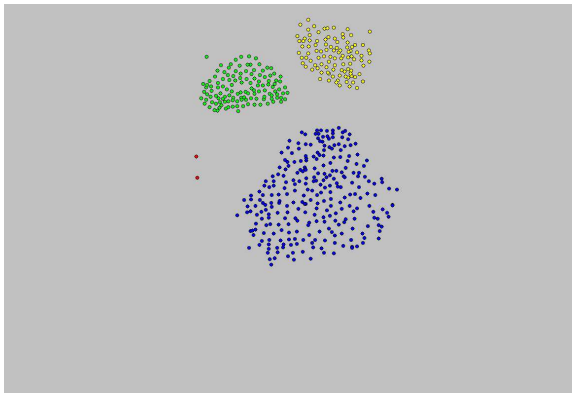


Fig. 1. Partition selon la méthode MRNM.

Si nous comparons les résultats de l'application de la procédure RNM modifiée avec la liste des adjacences, les adresses I.P. sont classés correctement.

Les résultats ont été comparés avec la méthode k-core [BAT 02b].

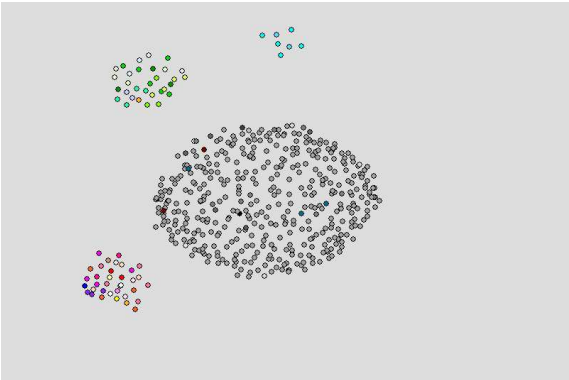


Fig. 2. Partition selon la méthode k-core.

Dans ce cas certains éléments ne sont pas correctement classés, car les deux éléments n'ayant pas de rélations avec les autres ne sont pas individués.

3. Conclusions

Dans cet article nous avons présenté une solution au problème d'identification de clusters denses dans l'analyse de fichiers d'enregistrement d'accès, en considérant une modification d'un algorithme connu de l'analyse de réseaux sociaux. Ainsi nous avons obtenu un outil pour étudier les clients dans les termes de leur comportement et leur information personnelle. Ceci permet l'accumulation d'éléments utiles pour l'amélioration de sites web et le développement de systèmes quand les séries de données sont grandes ou même énormes.

4. Bibliographie

[BAT 02a] BATAGELJ V., MRVAR A., *PAJEK : Program for large Network Analysis*, http ://www.vlado.fmf.uni-lj.si/pub/networks/pajek/, 2002.

[BAT 02b] BATAGELJ V., MRVAR A., *Partitioning approach to visualization of large graphs*, http ://www.vlado.fmf.uni-lj.si/pub/, 2002.

[BOR 99] BORGATTI S., EVERETT M., FREEMAN L., *Ucinet for Windows Software for Social Network Analysis*, Harvard : Analytic Technologies, http ://www.analytictech.com, 1999.

[FRI 98] FRIEDKIN N., E.C. J., Social position in influence networks, *Social Networks*, vol. 19, 1998, p. 122-143.

[KOC 89] KOCHEN M., *The small World*, Ablex Publishing Corporation, Norwood, NJ, 1989.

[MOB 02] MOBASHER B., DAI H., LUO T., SUNG Y., ZHU J., Integrating Web Usage and Content Mining for more Effective Personalization, *http ://www.maya.cs.depaul.edu/˜mobasher/personalization*, , 2002.

[MOO 01] MOODY J., Peer influence groups : identifying dense clusters in large networks, *Social Networks*, vol. 23, 2001, p. 261-283.

[SCH 02] SCHOIER G., Blockmodeling Techniques for Web Mining, W. H., B. R., Eds., *Proceedings of Compstat 2002*, Berlin, 2002, Springer and Verlag.

[SRI 00] SRIVASTAVA J., COLLEY J., DESHPANDE M., TON P., *Web Usage Mining : Discovery and Applications of Usage Patterns from Web Data*, http ://www.maya.cs.depaul.edu/˜mobasher/personalization, 2000.

[WAS 94] WASSERMAN S., FAUST K., *Social Network Analysis : Methods and Applications*, Cambridge University Press, New York, 1994.

Dissimilarités circulaires et hypercycles

Christophe OSSWALD

Département IASC, ENST Bretagne, 29285 Brest Cedex
Christophe.Osswald@enst-bretagne.fr

RÉSUMÉ. *Nous présentons une extension à la notion de pyramide (ou pseudo-hiérarchie), en donnant à tous les éléments des rôles semblables pour l'ordre induit, et en garantissant que la classification peut se représenter de façon planaire. Nous établissons une bijection entre ce modèle et une famille de dissimilarités. Finalement, nous proposons des algorithmes de reconnaissance et d'approximation.*

MOTS-CLÉS : *Classification, dissimilarités, pyramides, hypercycles, rigidité*

1. Définitions et notations

Soit X un ensemble fini à n éléments, notés indifféremment $x, y, z \ldots$ ou $x_1, x_2, \ldots, x_n$. Une **dissimilarité** (propre) d sur X est une application de $X \times X$ dans $\mathbb{R}$ vérifiant $d(x,y) = 0 \iff x = y$; $d(x,y) = d(y,x)$; $d(x,y) \geqslant 0$. Si $A \subsetneq X$, son **diamètre** est $\mathrm{diam}_d(A) = \max_{x,y \in A} d(x,y)$. Le **graphe-seuil** G_λ de d pour un seuil λ admet X comme ensemble de sommets et la paire xy est une arête si et seulement si $d(x,y) \leqslant \lambda$. Nous appelons **classe** de d un sous-graphe complet maximal pour l'inclusion d'un de ses graphes-seuil, c'est-à-dire une **clique maximale** de ce graphe-seuil. La **2-boule** $B(x,y)$ est constituée de tous les sommets z de X vérifiant $d(x,z) \leqslant d(x,y)$ et $d(y,z) \leqslant d(x,y)$.

Les classes d'une dissimilarité d forment l'ensemble $\mathcal{H}_d$. L'ensemble des 2-boules de d est noté $2\mathcal{B}_d$ et l'ensemble de ses boules $\mathcal{B}_d$.

Soit $G = (X, E)$ un graphe d'ensemble de sommets X et d'arêtes E. $A \subseteq X$ est **rigide** sur G si A est un sous-ensemble connexe de G. Plus généralement, $\mathcal{H}$ une famille de parties de X est rigide sur G lorsque chacune de ses parties est rigide sur G. Un **hypercycle** est un telle famille, rigide sur un cycle $C = (X, E)$. Notons qu'un hypercycle a au plus $n(n-1)$ parties.

Soit $A \subseteq X$. On désigne par $[A]_G$ l'ensemble des parties connexes du sous-graphe de G induit par A. Soit $\mathcal{H}$ une famille de parties rigide sur G. $\overline{[\mathcal{H}]}_G$ est la famille de partie de X dont les parties sont les parties connexes de G de toutes les intersections de toutes les parties de $\mathcal{H}$.

Soit f une application de $\mathcal{E} \subseteq 2^X$ dans $\mathbb{R}^+$. Une famille de parties $\mathcal{H} = (X, \mathcal{E})$ munie d'une application f est dit **indicée** si $A \subsetneq B \Rightarrow f(A) < f(B)$. Il est dit **pré-indicée** si $A \subsetneq B \Rightarrow f(A) \leqslant f(B)$. Il est dit **faiblement indicée** si $A \subsetneq B$, $f(A) = f(B)$ entraîne que $A = \bigcap_{A \subsetneq B} B$. Il est dit G**-faiblement indicée** si $A \subsetneq B$, $f(A) = f(B)$ entraîne que $A \in [\bigcap_{A \subsetneq B} B]_G$.

2. Dissimilarités circulaires et hypercycles

Soit θ un ordre total sur X et d une dissimilarité sur X. θ est circulairement compatible avec d si et seulement si les quatre conditions suivantes sont vérifiées pour tous $x\theta y\theta z\theta t$:

1. $d(x,z) \geq \min\{\max\{d(x,y), d(z,y)\}, \max\{d(x,t), d(z,t)\}\}$

2. $d(y,t) \geq \min\{\max\{d(y,x), d(t,x)\}, \max\{d(y,z), d(t,z)\}\}$

3. Si $d(x,z) > \max\{d(x,y),\ d(x,t),\ d(y,z),\ d(y,t),\ d(z,t)\}$ alors pour tous $u \in X$, $\max\{d(x,u), d(z,u)\} \leq \max\{d(x,y), d(x,t), d(y,z), d(y,t), d(z,t)\}$

4. Si $d(y,t) > \max\{d(x,y), d(x,z), d(x,t), d(y,z), d(z,t)\}$ alors pour tous $u \in X$, $\max\{d(y,u), d(t,u)\} \leq \max\{d(x,y), d(x,z), d(x,t), d(y,z), d(z,t)\}$

Nous dirons qu'une dissimilarité est circulaire si elle admet un ordre circulairement compatible. Cette définition diffère quelque peu de celle indiquée par Hubert *et al* (1998). Nous l'avons choisie car elle assure une bijection entre les dissimilarités circulaires et les hypercycles C-faiblement indicés.

Notons que la notion de compatibilité ne dépend que de la position des x_i sur un cycle. Ainsi, si θ est circulairement compatible, son inverse θ^{-1} l'est aussi, de même que tous les θ_i (avec $x_i\theta_i x_{i+1}\theta_i \ldots \theta_i x_n \theta_i x_1 \theta_i \ldots x_{i-1}$). Le résultat suivant rend compte de cette observation.

Proposition 1 *Soit d une dissimilarité sur X. S'il existe un cycle $C = (X, E)$ tel que chaque classe de d soit rigide sur C, alors il existe un ordre θ sur X circulairement compatible avec d.*

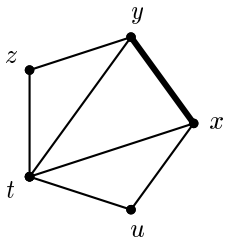
Proposition 2 *Soit d une dissimilarité sur X. Si $\mathcal{H}_d$ est un hypercycle, alors $2\mathcal{B}_d$ en est un aussi. Si $2\mathcal{B}_d$ est un hypercycle, $\mathcal{B}_d$ l'est également.*

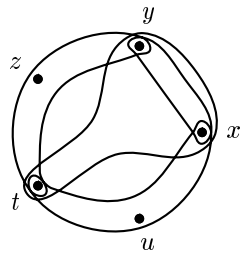
Démonstration Soit d une dissimilarité telle que ses cliques maximales forment un hypercycle sur un cycle C. Soient x et y dans X. Pour tout z de $B(x,y)$ il existe une classe A de d telle que $\{x,y,z\} \subseteq A$ et $\text{diam}(A) = \text{diam}(\{x,y,z\})$. A étant rigide sur C, il existe un chemin dans G de z à x, et de z à y, dans $B(x,y)$. Ainsi $B(x,y)$ est rigide sur C et les 2-boules de d forment un hypercycle sur C. Soit d une dissimilarité telle que ses 2-boules forment un hypercycle sur un cycle C. Pour tout y dans $B(x,\alpha)$, $B(x,y) \subseteq B(x,\alpha)$ est rigide sur C. Ainsi $B(x,\alpha)$ est rigide sur C et les boules de d forment un hypercycle sur C. □

Notons que les implications réciproques seraient fausses : les 2-boules de d_1 forment un hypercycle, pas ses cliques maximales. Les boules de d_2 forment un hypercycle, pas ses cliques maximales ni ses 2-boules :

d_1 :

	x	y	z	t	u
x	0				
y	1	0			
z	3	2	0		
t	2	2	2	0	
u	2	3	3	2	0

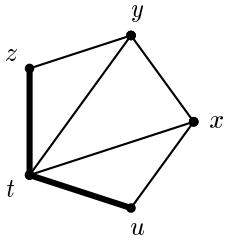


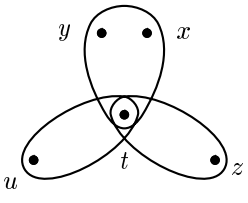


des cliques maximales de d_1

d_2 :

	x	y	z	t	u
x	0				
y	2	0			
z	3	2	0		
t	2	2	1	0	
u	2	3	3	1	0





des 2-boules de d_2

La proposition 3 ci-dessous constitue une réciproque à la proposition 1.

Proposition 3 *Soit d une dissimilarité sur X. La famille des classes de d admet un cycle sous-jacent si et seulement s'il existe un ordre linéaire sur X circulairement compatible avec d.*

Nous utilisons la famille de parties $[\overline{\mathcal{H}}]_G$ pour étendre la notion de fermeture par intersection qui permet de construire une C-faiblement indicée, rigide sur un cycle C, à partir d'un hypercycle indicé : l'intersection de deux classes peut ajouter deux classes disjointes au système. En étendant le théorème de bijection générale (Bertrand, 2000), complétant le système de classe rigide sur C par $[\overline{\mathcal{H}}]_C$ plutôt que par $\overline{\mathcal{H}}$, nous obtenons :

Proposition 4 *Soit $\mathcal{H} = (X, \mathcal{E}, f)$ une famille de parties indicée et rigide sur un cycle C. Il existe une unique dissimilarité circulaire d telle que $([\overline{\mathcal{H}}]_C, diam_d) = (\mathcal{H}, f)$, et pour chaque dissimilarité circulaire, il existe un unique hypergraphe C-faiblement indicé rigide sur ce cycle C, induit par $([\overline{\mathcal{H}_d}]_C, diam_d)$.*

3. Algorithmes

Reconnaissance d'un hypercycle

Algorithme 1: Trouver un cycle sous-jacent

$S \leftarrow (A, \neg A)$; $Z \leftarrow (1, 2)$;
répéter
  $\mathcal{A} \leftarrow \{A \in \mathcal{E} \mid A$ intersecte proprement deux zones $\}$;
  si $A \neq \phi$ **alors** Choisir $A \in \mathcal{A}$; Aller en 1;
  $\mathcal{B} \leftarrow \{A \in \mathcal{E} \mid A$ intersecte proprement une zone, et ne contient une mais pas toutes les autres zones $\}$;
  si $\mathcal{B} \neq \phi$ **alors** Choisir $A \in \mathcal{B}$; Aller en 1;
  $\mathcal{D} \leftarrow \{A \in \mathcal{E} \mid A$ intersecte proprement une zone, et contient toutes ou aucune des autres zones $\}$;
  si $\mathcal{D} \neq \phi$ **alors** Choisir $A \in \mathcal{D}$ tel que $|A|$ maximal;
  sinon S ne peut être raffiné : construire le cycle C; Aller en 2;
1 **Mettre à jour** S et Z en ajoutant la classe A;
jusqu'à $\mathcal{A} \cup \mathcal{B} \cup \mathcal{D} = \phi$;
2 Verifier que les classes non utilisées sont compatibles avec C;

L'algorithme 1 permet de trouver un cycle sur lequel toutes les classes d'un hypercycle sont rigides. Nous construisons, à chaque étape de l'algorithme, une partition de X en **zones**, S. Cette partition est en relation avec le cycle de la succession des **quartiers**, chaque zone correspondant à un ou deux quartiers. Le quartier est identifié par sa zone, non par les éléments qu'il contient.

Nous appelons **île** une classe de $\mathcal{D}$ qui soit est incluse dans une zone, soit son complémentaire est inclus dans une zone. Placer une île sur le cycle en construction est le seul moyen de séparer une zone en plusieurs quartiers. Maximiser la taille de l'île (cardinal de la classe considérée) évite de devoir remettre en cause cette séparation. Notons qu'il est équivalent de dire que A est rigide sur un cycle C, ou que le complémentaire de A l'est.

La procédure de mise à jour consiste en l'identification des limites de la classe ajoutée, en une ou deux créations de zones, et en une renumérotation des quartiers. La complexité globale de l'algorithme est en $\mathcal{O}(n^4)$.

Approximation

Etant NP-difficile (Barthélemy et Brucker, 2000) en norme finie, le problème de l'approximation d'une dissimilarité quelconque par une dissimilarité circulaire ne peut être résolu de façon satisfaisante. De plus, on peut constater que les dissimilarités circulaires se comportent mal vis-à-vis des inférieures maximales.

En se restreignant aux dissimilarités circulaires qui sont des quasi-ultramétriques (Bandelt et Dress, 1989 ; Diatta et Fichet, 1994 ; Diatta, 1996) et en utilisant les résultats de François Brucker (2001) sur les quasi-ultramétriques faiblement sous-dominantes, on peut mettre au point un algorithme permettant d'approcher en temps polynomial une dissimilarité quelconque par une dissimilarité circulaire à cycle fixé. Poue ce faire, nous proposons un algorithme qui, à cycle fixé, approche une dissimilarité d quelconque par une quasi-ultramétrique circulaire inférieure.

Nous considérons les $d(x, y)$ dans l'ordre croissant, et étendons un arc de cycle entre les éléments x et y de façon à construire une quasi-hiérarchie. Le processus s'arête lorsque deux classes couvrent X en ayant une intersection non connexe sur le cycle (violation de l'inégalité des 4 points pour les quasi-ultramétriques) (Diatta et Fichet, 1994) ou lorsque trois classes couvrent le cycle en ayant deux à deux des intersections non-vides (violation de la condition dite de Gilmore que $A \cap B \cap C \neq \phi$ entraîne qu'il existe $D \supseteq A \cap B \cap C$ et $f(D) \leqslant \max\{f(A), f(B), f(C)\}$) (Bertrand, 2000).

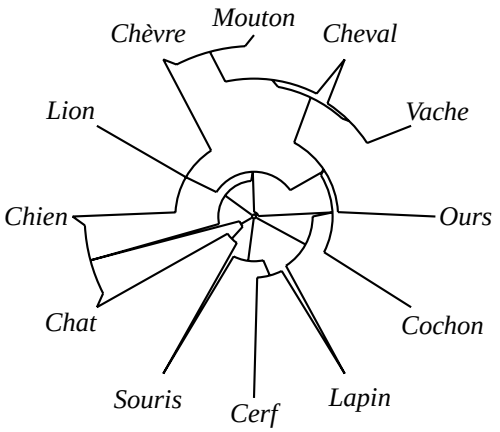
Nous avons au maximum $\mathcal{O}(n^2)$ dissimilarités à considérer, chacune nécessitant un traitement en $\mathcal{O}(n)$ plus un test en $\mathcal{O}(n^2)$ pour la dernière condition, soit une complexité globale en $\mathcal{O}(n^4)$.

Approximation lorsqu'aucun cycle n'est fixé

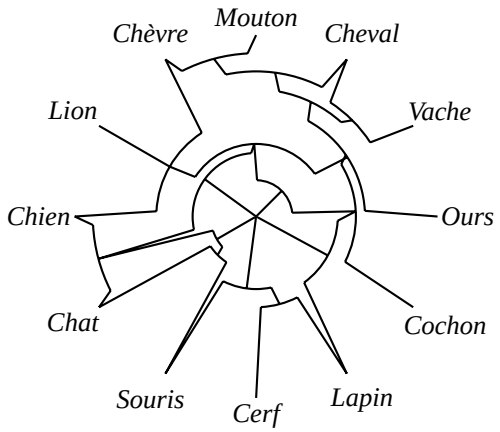
Lorsque la dissimilarité d n'est pas circulairement compatible, l'algorithme 1 conduit à un ordre partiellement compatible, *ie.* compatible avec un certain nombre des classes de d (dans notre cas, des 2-boules de d). Nous utilisons alors le second algorithme, couplé à une mesure de $||d - \delta_C||$ où δ_C est la quasi-ultramétrique obtenue par l'algorithme précédent sur le cycle C, en explorant le voisinage (au sens des transpositions) du cycle utilisé.

4. Exemple

L'application de cette méthode aux données de Henley (Barthélemy et Guénoche, 1988) donne ces résultats :



La hauteur des classes est proportionnelle au diamètre



Seul l'ordre total des diamètres est conservé

5. Bibliographie

[BAN 89] BANDELT H.-J., DRESS W. M., Weak hierarchies associated with similarity measures – an additive clustering technique, *Bulletin of Mathematical Biology*, vol. 51, n° 1, 1989, p. 133-166.

[BAR 88] BARTHELÉMY J.-P., GUÉNOCHE A., *Les arbres et les représentations de proximité*, Masson, Paris, 1988.

[BAR 01] BARTHÉLEMY J.-P., BRUCKER F., NP-hard approximation problems in overlapping clustering, *Journal of Classification*, vol. 18, n° 2, 2001, p. 159-183.

[BER 00] BERTRAND P., Set systems and dissimilarities, *European Journal of Combinatorics*, vol. 21, 2000, p. 727-743.

[BRU 01] BRUCKER F., Modèles de classification en classes empiétantes, PhD thesis, EHESS, juillet 2001.

[DIA 94] DIATTA J., FICHET B., *New approaches in classification and data analysis*, Chapitre From Asprejan hierachies and Bandelt-Dress weak-hierarchies to quasi-hierarchies, p. 111-118, Springer-Verlag, Berlin, 1994.

[DIA 96] DIATTA J., Une extension de la classification hiérarchique : les quasi-hiérarchies, PhD thesis, Université de Provence, mai 1996.

[HUB 98] HUBERT L., ARABIE P., MEULMAN J., Graph-theoretic representations for proximity matrices through strongly-anti-robinsonian or circular strongly-anti-robinsonian matrices, *Psychometrica*, vol. 63, n° 4, 1998, p. 341-358.

Interprétation des résultats de SVM

François Poulet

ESIEA Recherche
38, rue des Docteurs Calmette et Guérin
Parc Universitaire de Laval-Changé
53000 Laval - France
poulet@esiea-ouest.fr

RESUME. Nous présentons dans cet article deux outils graphiques visant à aider l'utilisateur à comprendre les résultats obtenus par les algorithmes de SVM (Séparateur à Vaste Marge ou "Support Vector Machine").Ces algortihmes sont de plus en plus souvent utilisés et donnent de bons résultats en classification, mais ils sont utilisés comme des boîtes noires. Ils cherchent à trouver la meilleure séparatrice entre les éléments de deux classes. Cette séparatrice est un hyperplan dans le cas le plus simple et une surface quelconque autrement. Le principal problème auquel on se trouve confronté est alors de trouver une représentation 2D ou 3D d'une telle frontière n-dimensionnelle sans perdre trop d'information.

MOTS-CLES : Fouille de données, Classification, SVM, Interprétation de résultats

1. Introduction

Les algorithmes de SVM (Support Vector Machine) font partie des méthodes de noyaux. Les premières publications [VaCh 64] utilisant ces méthodes datent du milieu des années 1960 dans le domaine de la reconnaissance des formes. Elles connaissent ces dernières années un fort regain d'intérêt (notamment en fouille / analyse de données). De nombreux logiciels sont disponibles à l'heure actuelle (cf. la liste disponible sur www.kernel-machines.org). De nombreuses versions différentes de SVM ont fait l'objet de publications et leur champ d'application s'élargit de jour en jour. Nous nous intéressons ici à une utilisation particulière des algorithmes de SVM : la classification supervisée. Dans ce cadre d'utilisation, les algorithmes de SVM permettent de trouver la meilleure séparatrice entre les éléments de deux classes. La nature de cette séparatrice varie avec le choix de la fonction de noyau utilisée, elle peut être un hyperplan, une fonction polynomiale de degré d ou une fonction sigmoidale, …

Habituellement, les seuls résultats fournis par les algorithmes de SVM sont l'équation de la séparatrice et le taux de bonne classification obtenu. Ici, nous nous intéressons à essayer d'expliquer ces résultats. Une première méthode utilisable dans le cas d'une frontière linéaire va consister à visualiser les projections en 2D de l'hyperplan de séparation. Bien entendu, l'information obtenue de cette manière est approximative (comme l'est toute projection 2D d'une primitive graphique de dimension supérieure à 2). Une autre solution (valable pour une frontière linéaire ou non) est d'afficher la distribution des individus en fonction de leur classe et de leur distance à la séparatrice calculée.

2. Visualisation des projections de l'hyperplan

Pour effectuer la visualisation de l'hyperplan de séparation, nous allons partir de l'ensemble des projections 2D suivant toutes les paires possibles d'attributs. Puis nous allons calculer toutes les intersections de l'hyperplan obtenu par le SVM avec les matrices 2D, c'est à dire une droite pour chaque matrice. Comme on peut le voir sur le schéma de la partie gauche de la figure 1, cette information est approximative. Dans le pire des cas, on a affaire à un hyperplan dont tous les coefficients sont non nuls (c'est à dire que la séparatrice est une combinaison linéaire de tous les attributs, il n'y a aucun coefficient nul ou proche de zéro). Les projections de l'hyperplan obtenues sur les matrices 2D ne sépareront pas visuellement les données des deux classes, bien que l'hyperplan, lui, sépare réellement les données. Dans le cas le plus favorable, il y aura bien une séparation visuelle des données par la droite projetée. Dans le cas général, on a un mélange des deux cas précédents, c'est par exemple le cas sur la partie gauche de la figure 1, où deux projections montrent une droite séparant les deux classes alors que la troisième montre les deux classes du même côté du plan.

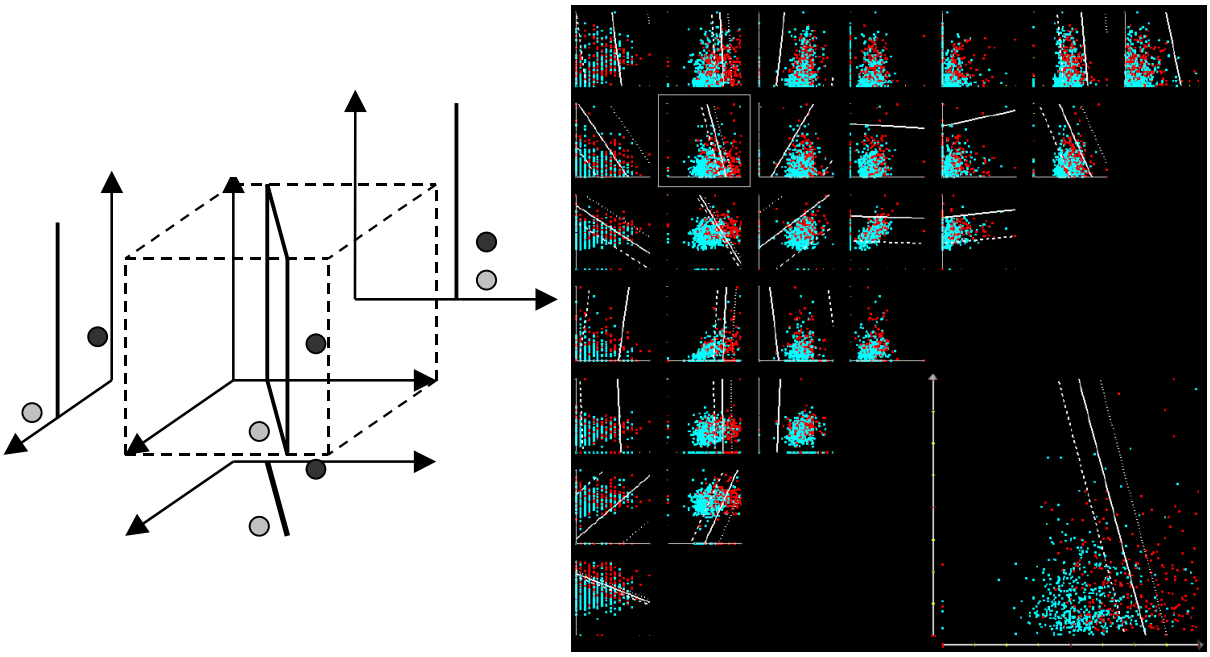


Figure 1. *Visualisation de l'hyperplan de séparation des deux classes sur les projections 2D*

Ce type de visualisation doit être vu plus comme un moyen de situer approximativement la position de l'hyperplan de séparation des données, que comme un moyen d'évaluer la qualité de la séparatrice obtenue. On ne peut aisément évaluer cette qualité puisque des points séparés par l'hyperplan peuvent apparaître du même côté de la droite après projection, et inversement, des points non séparés par l'hyperplan peuvent apparaître séparés après la projection. La partie droite de la figure 1 montre un exemple d'un telle représentation sur le set de données diabetes de l'UCI [BlMe 98] dans l'environnement graphique de fouille de données que nous avons développé [Poul 02].

3. Visualisation de la distribution des individus par rapport à l'hyperplan

Une toute autre méthode va consister à calculer (en même temps que la classification est effectuée par l'algorithme de SVM) pour chaque individu, sa distance perpendiculairement à l'hyperplan de séparation. On affiche ensuite la distribution de ces points, sous la forme d'un histogramme, avec en positif les points bien classés (du bon côté du plan) et en négatif, les points mal classés. Ceci permet dans un premier temps de voir de manière très intuitive la qualité de la séparatrice obtenue. Un exemple d'une telle visualisation est donné sur la figure 2 sur les données Segment de l'UCI avec en gris clair la classe 5 et en gris foncé les autres classes. On remarque immédiatement que les individus de la classe "le reste" sont plutôt bien classés alors que ceux de la classe 5 sont plutôt très mal classés.

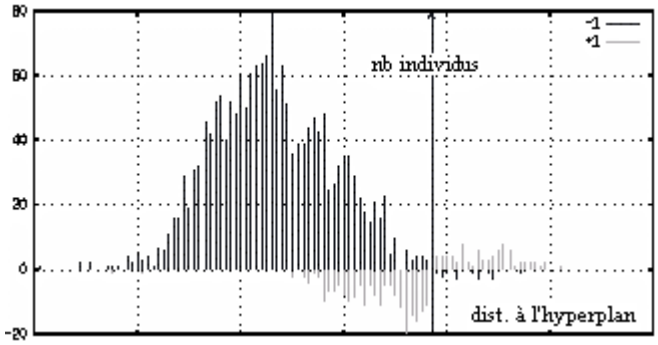


Figure 2. *Visualisation de la distribution des données en fonction de leur distance à l'hyperplan*

Cette représentation peut aussi servir de base pour aller voir quels sont les individus qui "posent problème". On peut sélectionner l'une des barres de l'histogramme et les individus correspondant sont alors mis en évidence dans l'ensemble des matrices 2D. Ceci permet de localiser les individus pour lesquels la classification ne donne pas des résultats satisfaisants.

Sur l'exemple de la figure 3 on représente le résultat de la classification de la classe 5 contre le reste pour les données Segment de l'UCI et on sélectionne les deux barres les plus proches de l'hyperplan obtenu, correspondant aux individus mal classés (en dessous de l'axe horizontal). Les individus en question sont alors représentés en gras sur la matrice 2D. On peut voir dans ce cas, que l'hyperplan de séparation est presque perpendiculaire à la matrice 2D. Les 6 points en surbrillance les plus à droite correspondent aux individus mal classés des autres classes (gris clair) et tous les autres points en gras correspondent aux individus mal classés de la classe 5 (gris foncé). L'hyperplan de séparation des données en deux classes se situe entre ces deux groupes de points. A l'inverse, on peut tout aussi facilement choisir de visualiser les individus les plus représentatifs de la classe.

4. Conclusion - Perspectives

Nous avons présenté deux techniques graphiques de visualisation / interprétation des résultats des algorithmes de SVM. La première est basée sur un ensemble de projections 2D (selon toutes les paires possibles d'attributs) de l'hyperplan de séparation des données en deux classes. Bien qu'approximative, cette visualisation permet d'avoir de manière simple une

information sur la position de l'hyperplan. La seconde technique présentée permet quant à elle d'évaluer la qualité de la frontière obtenue par le biais d'un histogramme représentant la distribution des individus en fonction de leur distance au plan de séparation. Liée aux matrices 2D ce système permet de plus de localiser les individus mal classés ou les plus représentatifs de la classe. Cette dernière représentation peut être naturellement généralisée à n'importe quelle type de frontière (linéaire ou non).

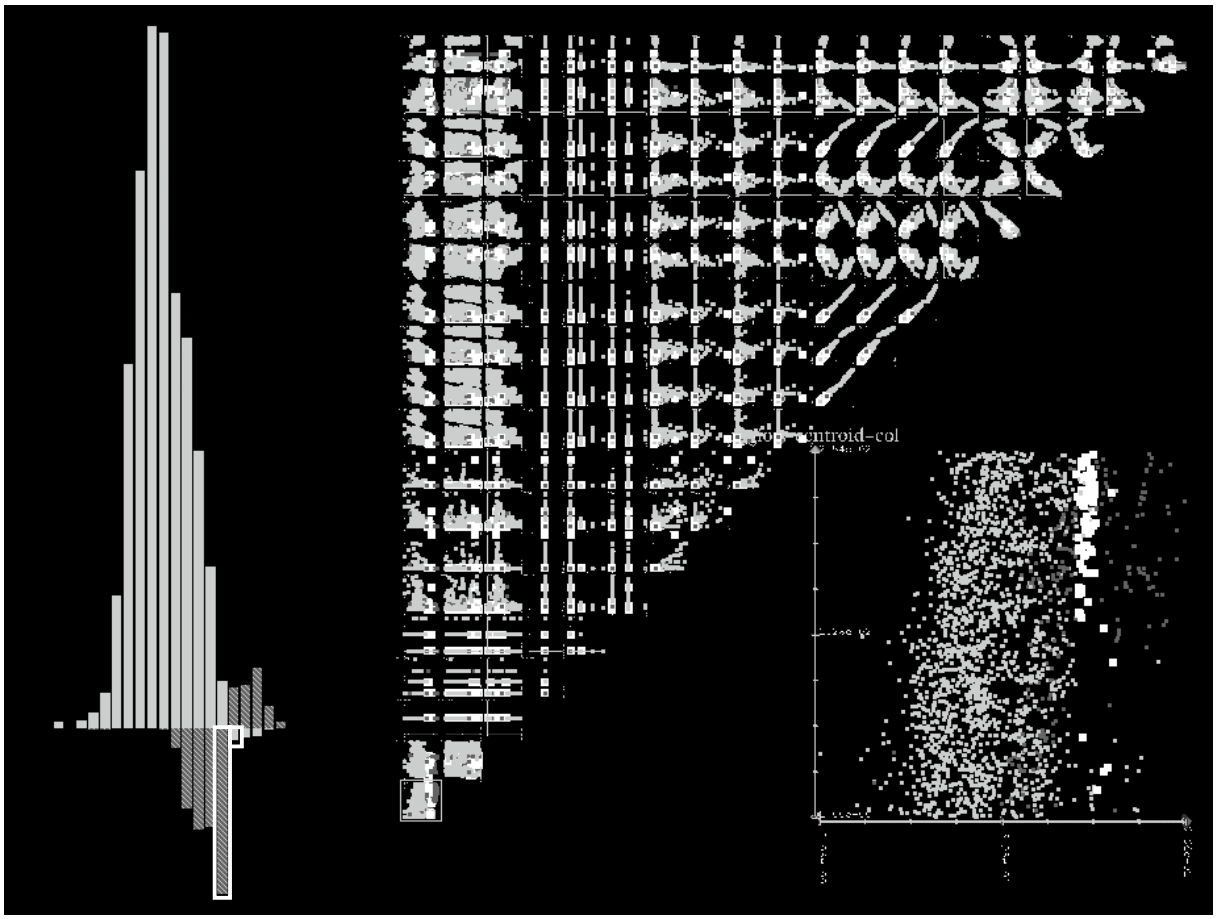


Figure 3. *Visualisation des individus mal classés les plus proches de l'hyperplan*

5. Bibliographie

[BlMe 98] C.Blake, C.Merz, UCI Repository of machine learning databases, [www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, University of California, Department of Information and Computer Science, (1998).

[Poul 02] F.POULET, "FullView: A Visual Data Mining Environment", *International Journal of Image and Graphics, 2(1),* 127-144, Jan.2002.

[VaCh 64] V. VAPNIK, A.CHERVONENKIS, "A note on one class of perceptrons" in *Automation and Remote Control*, 25, 1964.

Reconnaissance dynamique de formules mathématiques

Marcel Rémon

Département de Mathématique
Université de Namur B-5000 Namur
marcel.remon@fundp.ac.be

RÉSUMÉ. *Les algorithmes de reconnaissance d'objets procèdent habituellement en comparant l'objet à reconnaître à une collection d'objets de référence, appelée base d'entraînement. Cette dernière ne comporte qu'un nombre fini d'éléments. Dans plusieurs applications, telle que la reconnaissance de formules mathématiques, le nombre de situations plausibles est quasiment illimité, vu la structuration propre de ce type d'expression.*
Il est important de développer des systèmes dynamiques de reconnaissance qui puissent générer, à la demande, des nouvelles situations de référence. L'aspect le plus ardu de cette recherche est l'automatisation du choix des nouvelles références que devra proposer le générateur (optimisation dynamique), en fonction de l'étape où le programme de reconnaissance se trouve.

MOTS-CLÉS : *Classification, Reconnaissance de caractères, Discrimination, Analyse d'images*

1. Introduction

Dans le domaine de la reconnaissance de documents, celle de formules mathématiques a déjà fait l'objet de nombreuses études[FAU 00, LAV 00, KAC 01]. L'extraction automatique de formules mathématiques est un problème plus difficile qu'il n'y paraît à première vue, car la complexité d'une formule peut être très grande. L'idée préconisée dans cette présentation est de permettre à l'algorithme de reconnaissance de générer, à la demande, de nouveaux objets de référence, afin que la base d'entraînement (les formules de référence) ne soit plus limitée en taille.

2. Segmentation de la formule

Soit $\mathcal{F}$ l'ensemble des formules mathématiques possibles. Cet ensemble est infini. Soit F la formule à reconnaître.

La première étape de reconnaissance est classique, mais peut être complexe. Il s'agit de l'étape de **segmentation** de l'équation en entités simples. On découpe la formule en symboles atomiques (lettres, opérateur arithmétique). Pour cela, on procède d'abord à une séparation des symboles mathématiques selon l'axe vertical, puis selon l'axe horizontal, et ainsi de suite jusqu'à l'obtention de zones non séparables ni verticalement, ni horizontalement. Dans ces zones, on procède à une identification (qui n'est pas encore de la reconnaissance) des symboles atomiques par connexité. Une entité est définie comme un ensemble de pixels connexes. Le problème des symboles non connexes est traité (signe $=$, lettres i et j). Chaque entité est localisée afin de reconstruire par après l'arbre de la formule.

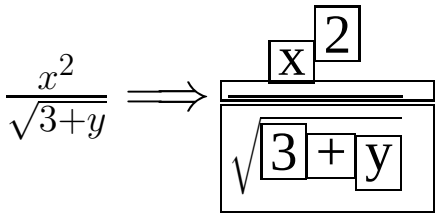


Figure 1. Segmentation des entités mathématiques

3. Reconnaissance des entités

Chaque symbole atomique ainsi identifié est comparé avec les éléments b_i d'une base d'entraînement simplifiée, soit $\mathcal{B}$. Pour cela, il nous faut un générateur de telles entités de référence. Ce générateur est un petit script utilisant LaTeX. Pour que cette reconnaissance soit efficace, il nous faut une distance robuste, peu sensible à la casse ou à la police de caractère utilisée (italique, gras, police générée par Word ou LaTeX).

Après plusieurs tests, la distance retenue est définie comme suit : on génère un grand nombre de droites aléatoires dans la fenêtre contenant l'entité et on compte la fréquence de droites rencontrant l'entité $0, 1, 2, \cdots, n$ fois [1]. Ensuite on compare (distance euclidienne dans $\mathbb{R}^n$) ce vecteur de fréquences avec ceux trouvés pour les entités b_i de référence. Cette mesure est très robuste aux dilatations, translations, rotations. Elle permet de discrimer entre des symboles très proches. A chaque entité identifiée est associé un élément de $\mathcal{B}$, ou plusieurs, selon le seuil de discrimination accepté entre les entités de référence.

Figure 2. Exemple de mesure morphologique.
(f_1, f_2, f_3, f_4) où f_i est la fréquence des droites rencontrant i fois l'entité.
Pour la figure : $(0, \frac{4}{8}, \frac{4}{8}, 0), (0, \frac{3}{8}, \frac{4}{8}, \frac{1}{8})$ et $(\frac{1}{8}, \frac{2}{8}, \frac{3}{8}, \frac{2}{8})$.

b_i	0	1	2	3	4	$\|b_i - \mathbf{x}\|$
$-$	0,0	1,0	0,0	0,0	0,0	0,6124
$+$	0,2834	0,5391	0,1774	0,0	0,0	0,1825
1	0,1264	0,7880	0,0852	0,0	0,0	0,3687
2	0,1255	0,6273	0,2289	0,0182	0,0	0,1571
3	0,1238	0,5751	0,2826	0,0184	0,0	0,0923
5	0,1201	0,5743	0,2933	0,0122	0,0	0,0923
x	0,18	0,4671	0,3081	0,0428	0,0019	0,0355
$\mathbf{x}$	0,1652	0,4970	0,3062	0,0310	0,0004	

Figure 3. Comparaison entre $\mathbf{x}$ et plusieurs entités de référence b_i.

1. on en a généré 20 000 pour la figure 3, mais 2 000 droites donnent déjà une bonne estimation de la probabilité pour une droite quelconque de rencontrer i fois l'entité. Le nombre n est choisi de telle façon que la fréquence de rencontrer $n + 1$ fois l'entité est nulle.

4. Première estimation de la formule

A partir de la localisation des entités lors de l'étape de segmentation, nous construisons, de manière récursive, un arbre associé à la formule. L'arbre ainsi construit se base sur la grammaire des formules mathématiques. Il est important que l'arbre soit mathématiquement interprétable, et non pas une suite de symboles atomiques.

Dans un premier essai, nous avions comme résultat un arbre composé de fonctions interprétables par LaTeX, ce qui nous permettait de redessiner la formule. Mais, cela n'est pas suffisant, car il n'y a pas de différence en LaTeX entre `2 5` et `25`. De plus, LaTeX ne peut manipuler les formules.

C'est pourquoi nous sommes en train de travailler à la construction d'arbres qui soient interprétables mathématiquement, par exemple, en Matlab. La reconnaissance de la formule est alors quasiment effective, car on peut la reprendre dans un algorithme de calcul.

Pour cela, nous avons besoin d'une grammaire des opérateurs et des termes. Cette grammaire est également utilisée pour affiner notre reconnaissance de la formule globale, car les expressions non valides sont éliminées d'office.

Lorsque le découpage n'est pas clair - par exemple, dans le cas d'exposant ou d'indice -, l'algorithme garde en mémoire les structurations possibles de la formule étudiée. Par exemple, nous retenons comme choix possibles xy, x_y et x^y si l'entité y n'est pas exactement au même niveau que x, à condition que cela ait un sens en arithmétique (on ne peut avoir $+_y$).

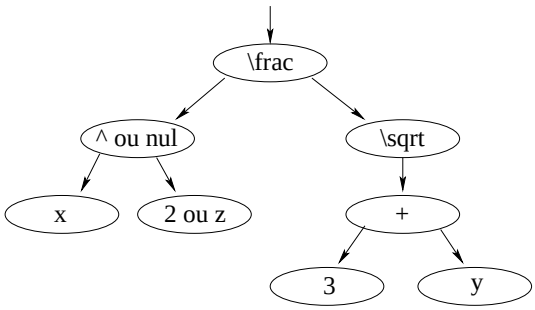


Figure 4. Exemple d'arbre avec mémoire des choix possibles.

5. Reconnaissance dynamique de la formule globale

Il s'agit d'améliorer l'estimation de F. Grâce au script générateur de formules et à l'utilisation de la mesure exposée précédemment, il est aisé de comparer n'importe quelle formule mathématique avec la formule à reconnaître.

Le problème est de fournir au générateur de formules de bons candidats pour les estimations successives. Lors de l'étape de reconnaissance des entités, nous avons gardé, par chaque entité, un ou deux candidats dans l'ensemble $\mathcal{B}$, et nous avons fait de même lors de la création de l'arbre représentant la formule. Nous testons alors systématiquement toutes les combinaisons apparaissant dans l'arbre de la formule. Ainsi, pour l'exemple de la figure 1, l'algorithme générera les formules suivantes : $\frac{x2}{\sqrt{3+y}}$, $\frac{x^2}{\sqrt{3+y}}$, $\frac{xz}{\sqrt{3+y}}$ et $\frac{x^z}{\sqrt{3+y}}$.

Ce n'est qu'à partir des formules globales générées par l'algorithme que se fera la reconnaissance ultime de la formule. Là encore, la distance basée sur un ensemble aléatoire de droites est utilisée, vu sa robustesse. Et les résultats sont vraiment encourageants, que ce soit pour des formules générées en LaTeXou en Word.

6. Bibliographie

[FAU 00] FAURE C., *http ://www.tsi.enst.fr/∼cfaure/math.html*, Liste de références pour la reconnaissance de formules mathématiques, 2000.

[KAC 01] KACEM A., BELAÏD A., BAN AHMED M., Automatic extraction of printed mathematical formulas using fuzzy logic and propagation of context, *International journal on Document Analysis and Recognition*, vol. 4, 2001.

[LAV 00] LAVIROTTE S., *Reconnaissance structurelle de formules mathématiques typographiées et manuscrites*, Thèse défendue à l'Université de Nice Sophia-Antipolis, 2000.

Sur des indices de comparaison de deux classifications

Genane Youness

CEDRIC-CNAM
BP 114661
Beyrouth, Liban
genane99@hotmail.com

Gilbert Saporta

Chaire de Statistique Appliquée et CEDRIC-CNAM
292 rue Saint Martin
75141 Paris Cedex 03
saporta@cnam.fr

RÉSUMÉ. On étudie la ressemblance entre deux partitions sur les mêmes individus à l'aide du coefficient de corrélation vectorielle RV et du kappa de Cohen (dans le cas où les partitions ont le même nombre de classes). On montre que le RV s'identifie à l'indice J de Janson et Vegelius. On étudie la distribution d'échantillonnage de ces indices pour des paires de partitions proches (issues d'un modèle de classes latentes) afin de donner des valeurs critiques sous des hypothèses réalistes.

MOTS-CLÉS : kappa de Cohen, corrélation vectorielle, indice de Janson et Vegelius, classification, classes latentes.

1. Introduction

Dans des travaux antérieurs [SAP 01, 02, 03], nous avons étudié la distribution du coefficient de Rand et d'indices voisins dans le but de comparer deux classifications provenant d'un même ensemble de données, afin de répondre aux questions suivantes : lors de deux enquêtes portant sur les mêmes individus, comment mesurer l'accord entre les deux classifications? Est-ce que les configurations de ces deux classifications se ressemblent ?

On présente ici les écritures logiques et relationnelles d'un indice obtenu à partir du coefficient de corrélation vectorielle RV introduit par P. Robert et Y. Escoufier [ROB 76] qui se révèle identique au coefficient J de S. Janson et J. Vegelius [JAN 82] ainsi que leurs distributions d'échantillonnage sous une hypothèse nulle d'absence de liaison.

Le coefficient kappa de Cohen, fournit une autre façon de mesurer l'accord entre deux partitions ayant le même nombre de classes, provenant d'un même échantillon. Cet indice, contrairement au précédent dépend de la numérotation des classes : on identifie la permutation des classes d'une des deux partitions en maximisant la valeur du kappa.

2. Notations

Soient P_1 et P_2 deux partitions des mêmes individus (ou deux variables qualitatives) à p et q classes. On notera K_1 et K_2 les tableaux disjonctifs associés et N le tableau de contingence croisant P_1 et P_2 de terme général n_{ij}. On a $N=K'_1K_2$

Lorsque l'on croise deux partitions, on s'intéresse également aux paires d'individus qui restent ou non dans les mêmes classes. On a en tout $n(n-1)/2$ paires d'individus.

A chaque partition P_k est associé un tableau relationnel C^k, de dimension nxn, dont le terme général $c^k_{ii'}$ est défini par :

$$c_{ii'}^{k} = \begin{cases} 1 \text{ si les deux individus i et i' sont dans la meme classe de la partition } P_k \\ 0 \text{ sinon} \end{cases}$$

On a $C^1 = K_1 K_1'$ et $C^2 = K_2 K_2'$

3. RV ou J

Le coefficient de corrélation vectorielle RV introduit par P. Robert et Y. Escoufier [ROB 76] permet de mesurer la ressemblance entre deux tableaux de données numériques X_1 et X_2 sur les mêmes observations en comparant les produits scalaires inter-individus associés aux deux tableaux .

Ces matrices de produits scalaires $X_i X'_i$ notées W_i sont de dimension *nxn* . Le coefficient RV est défini par :

$$RV(X_1, X_2) = \frac{\text{trace}(W_1 W_2)}{\sqrt{\text{trace}(W_1^2)\,\text{trace}(W_2^2)}}$$

Les travaux de A. Lazraq et R.Cleroux [LAZ 01,02] donnent la possibilité de tester des hypothèses concernant RV mais pour des données numériques.

Si on applique ce coefficient à deux tableaux disjonctifs K_1 et K_2 , on trouve :

$$RV(P_1,P_2) = \frac{\text{trace}(C^1 C^2)}{\sqrt{\text{trace}(C^1)^2\,\text{trace}(C^2)^2}} = \frac{\sum_{i,i'} (c^1_{ii'})(c^2_{ii'})}{\sqrt{\sum_{i,i'} (c^1_{ii'})^2 \sum_{i,i'} (c^1_{ii'})^2}}$$

Si RV est suffisamment grand, les classifications obtenues seront voisines.

En centrant les $c^k_{ii'}$ on retrouve la forme relationnelle de l'indice J de Janson et Vegelius établi par [IDR 00]:

$$J(P_1,P_2) = \frac{pq\sum\sum n_{ij}^2 - p\sum n_{i.}^2 - q\sum n_{.j}^2 + n^2}{\sqrt{[p(p-2)\sum n_{i.}^2 + n^2][q(q-2)\sum n_{.j}^2 + n^2]}} = \frac{\sum_{i,i'}(c^1_{ii'} - \frac{1}{p})(c^2_{ii'} - \frac{1}{q})}{\sqrt{\sum_{i,i'}(c^1_{ii'} - \frac{1}{p})^2 \sum_{i,i'}(c^2_{ii'} - \frac{1}{q})^2}}$$

4. Le kappa de Cohen

Introduit par [COH 60], le coefficient kappa est une mesure d'accord entre deux variables qualitatives pour des données appariées : pour deux partitions à même nombre de classes, il mesure l'écart à la diagonale du tableau de contingence :

$$\kappa = \frac{n\sum_{i=1}^{k} n_{ii} - \sum_{i=1}^{k} n_{i.} n_{.i}}{n^2 - \sum_{i=1}^{k} n_{i.} n_{.i}}$$

La concordance observée P_o est la proportion d'individus classés dans les cases diagonales de concordance du tableau de contingence, soit la somme des effectifs diagonaux divisés par la taille de l'échantillon *n* : $P_0 = \frac{1}{n}\sum_{i=1}^{k} n_{ii}$

La concordance aléatoire P_e est égale à la somme des produits des effectifs marginaux divisés par le carré de la taille de l'échantillon $P_e = \frac{1}{n^2}\sum_{i=1}^{k} n_{i.} n_{.i}$. Le kappa exprime la différence relative entre la proportion d'accords observés P_o et la proportion d'accords aléatoires P_e qui est la valeur espérée, sous l'hypothèse nulle d'indépendance des variables, divisée par le complément à un de l'accord aléatoire.

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

L'identification des classes est nécessaire pour utiliser le coefficient kappa, car quand on a à comparer deux partitions des mêmes individus obtenues par des méthodes de classification, la numérotation des classes est arbitraire : il est alors logique d'identifier les classes des partitions qui conduisent à une valeur maximale de κ . On prend alors la permutation des classes qui maximise le kappa d'où leur renumérotation.

5. Distributions d'échantillonnage

On utilise la méthodologie présentée en [SAP 02] pour étudier la distribution d'échantillonnage de ces deux indices pour des paires de partitions proches.

Rappelons ici que le but n'est pas d'étudier si les deux partitions sont indépendantes, mais si elles sont concordantes : la difficulté étant de formuler correctement l'hypothèse nulle de concordance.On procéde alors comme suit: à partir d'une partition initiale basée sur un modèle de classes latentes, on obtient deux partitions par une méthode classique type k-means obtenue en séparant les variables en deux blocs. On calcule alors les indices ci-dessus pour les deux partitions : en itérant le procédé on obtient par simulation la distribution d'échantillonnage de RV (ou J) et de kappa.

6. Application numérique

On a obtenu deux partitions de 1000 individus P_1 et P_2 à 4 classes chacune par la méthode des k-means selon deux groupes de variables

Le tableau de contingence croisant les deux partitions est :

1	**2**	**3**	**4**
248	0	0	2
1	198	27	9
2	6	43	202
0	58	192	12

On trouve une valeur de l'indice Kappa égale à 0.335, et une valeur de J (ou RV) égale à 0.648.

On réordonne ensuite les colonnes selon le kappa maximal (il y a 4! permutations) pour pouvoir identifier les classes de P_2 à celles de P_1 : la valeur maximale du kappa est de 0.787 obtenue en permutant les deux dernières colonnes d'où le tableau réordonné :

1	**2**	**4**	**3**
248	0	2	0
1	198	9	27
2	6	202	43
0	58	12	192

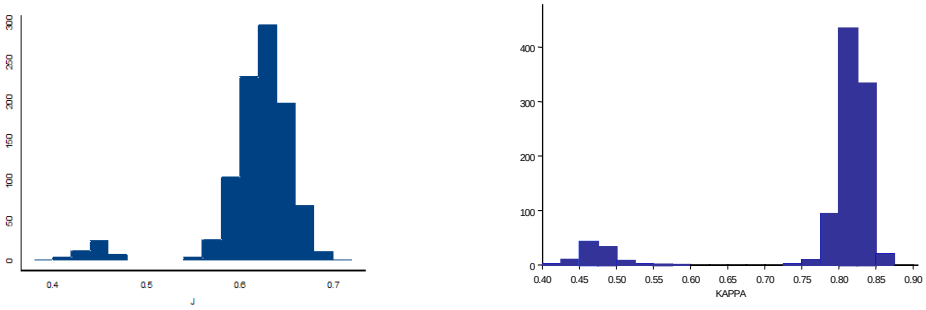


Figure 1 *: Distribution des coefficients de Janson et Vegelius et de kappa pour des partitions en 4 classes de 1000 individus avec 1000 itérations.*

Par simulation on trouve que le coefficient J varie entre 0.4 et 0.7. La valeur la plus fréquente est de 0.63 et la moyenne du coefficient J est égale à 0.617. Le coefficient kappa de Cohen varie entre 0.4 et 0.875, et est de moyenne 0.82. La bimodalité est due à la présence d'optimums locaux engendrés par la méthode des k-means. Dans l'exemple, on en déduit que les deux partitions sont suffisamment proches car les deux coefficients prennent des valeurs proches de la moyenne sous l'hypothèse de partitions identiques.

7. Conclusion

Nous avons montré l'identité des coefficients RV et J pour des partitions. J et le kappa de Cohen (mais ce dernier pour des nombres de classes identiques et après permutation) permettent de tester la similitude entre deux partitions en les comparant à leur distribution simulée dans le cas de données provenant d'une même partition « mère ».

Bibliographie

[COH 60] COHEN J., A coefficient of agreement for nominal scales., *Educ. Psychol. Meas.*, vol 20, 1960, p.27-46.

[IDR 00] IDRISSI A., *Contribution à l'unification de Critères d'Association pour Variables Qualitatives*, Thèse de doctorat de l'Université de Paris 6, 2000.

[LAZ 02] LAZRAQ, A., CLEROUX R., Inférence Robuste sur un indice de Redondance, *Revue de Statistique Appliquée*, vol. (4), p.39-54, 2002.

[JAN 82] JANSON S., VEGELIUS J., The J-index as a measure of association for nominal scale response agreement, *Applied psychological measurement*, vol. 16, 1982, p.243-250.

[ROB 76] ROBERT P., ESCOUFIER, Y. A unifying tool for linear multivariate statistical methods: the RV-coefficient, *Appl. Statist.*, vol. 25**,** 1976, p.257-65.

[SAP 01] SAPORTA G., YOUNESS G., Concordance entre deux partitions: quelques propositions et expériences, *in Actes des 8èmes rencontres de la SFC,* 2001, Pointe à Pitre.

[SAP 02] SAPORTA G., YOUNESS G., Comparing two partitions: some proposals and Experiments, *Proceedings in Computational Statistics edited by Wolfgang Härdle, Physica- Verlag*, 2002, Berlin.

[SAP 03] SAPORTA G., YOUNESS G., Une méthodologie pour la comparaison de partitions, *Revue de Statistique Appliquée*, à paraître.

Les differences financieres en Italie

Giorgio Skonieczny et Benedetto Torrisi

Faculté d'Economie
Université de Catane (Italie)
Corso Italia, n° 57
95127 CATANIA
giorgios@unict.it - btorrisi@unict.it

RESUME: Le procès de restructuration et de redéfinition de la fonction d'intermédiaire bancaire a entraîné de profonds changements dans le système financier italien. Dans cette situation les banques ont essayé une rapide réorganisation de leurs structures dans le but de parvenir à un meilleur rangement sur le plan de la rentabilité. L'objectif principal de notre étude consiste à fournir une représentation sur les déséquilibres financiers des régions italiennes, en utilisant les techniques d'analyse multivariée. L'analyse particulière des «cluster » (cluvar) a permis d'obtenir un tableau sur les discordances du système financier des régions italiennes en mettant en évidence les relations entre les régions et les variables plus significatives qui ont contribué à la détermination de ces relations elles-mêmes.
MOTS-CLÉS : banques, cluster analisys, PCA, régionalisation.

1. Introduction

Depuis des années le marché financier et monétaire résulte soumis à des transformations remarquables qui peuvent être imputées à des raisons différentes. Entre autres on remarque une intense «disintermédiation des créances », un processus de restructuration et de redéfinition de la fonction d'intermédiation financière des banques, avec la conséquence d'une naissance de nouveaux opérateurs financiers non bancaires spécialisés (Gallo G., 1990). Parallèlement le débouché progressif des mouvements de capital et la prospective du marché unifié européen ont déterminé la plus grande pression de la concurrence à l'intérieur des systèmes financiers (Parillo F., 1998). Les banques ont essayé une rapide réorganisation dans le but de parvenir à un meilleur rangement sur le plan de la rentabilité (Alessandrini, 1992).

On en trouve l'épreuve dans le grand nombre d'opérations de fusion et d'incorporation, la croissante diversification de l'activité bancaire avec la progressive ouverture aux services bancaires, financiers et immobiliers, outre le recours à des modèles de gestion variés. La concurrence a encouragé en plus la naissance et une diffusion rapide de nouveaux instruments dans l'emploi de l'épargne. Le tout dans le but d'une plus grande efficacité du système (Parrillo F., 1955). L'évolution des marchés financiers constitue un deuxième secteur de

recherche surtout pour ce qui concerne les relations entre l'économie réelle et l'intermédiation financière, entre le crédit et le développement, outre l'articulation du phénomène en relation aux différentes réalités du territoire. Ce dernier aspect constitue le sujet de notre recherche dont l'objectif est celui de tracer un plan de la diversification du territoire à l'intérieur du système financier italien sur la base d'indicateurs spécifiques. C'est ainsi qu'on essayera de déterminer les différences ou les ressemblances entre les régions italiennes, de le classer en des groupes ayant des caractéristiques homogènes et de constater la présence et la nature d'éventuelles différences financières.

2. Les indicateurs

Le sectionnement du territoire en des aires homogènes en relation au phénomène étudié implique la localisation d'indicateurs spécifiques qui soient à même de synthétiser l'information statistique qu'on espère pouvoir obtenir.

Suivant ce procédé, on a localisé, en relations aux hypothèses, un ensemble exhaustif de variables et l'on a construit les macro indicateurs concernant les aspects du phénomène ci-après mentionnés : économie réelle (7 indicateurs simples) - densité bancaire (5 indicateurs simples) - productivité bancaire (5 indicateurs simples) - financement de l'activité de production (6 indicateurs simples) -intermédiation du crédit (3 indicateurs simples) - analyse dimensionnelle des agences de crédit (4 indicateurs simples) - risque et taux d'intérêts (9 indicateurs simples) - flux financiers inter-régionaux (4 indicateurs simples) - activités financières des ménages (4 indicateurs simples) - système des règlements (8 indicateurs simples).

3. Analyse empirique

Dans le but de tâter la répartition du territoire italien en relation à chaque aspect marquant, selon les hypothèses qu'on a faites, les différentes typologies du comportement du phénomène, nous avons utilisé singulièrement des groupes de variables conformément à ce que nous avons indiqué plus haut. La technique que nous avons proposée a pour but le repérage des groupes homogènes et des relations entre les unités (parcours) (relations entre chaque unité et les cluster tout en identifiant parmi les variables celles qui ont le plus contribué aux liens entre les couples d'unités).Dans le détail, étant donné un barème de données où x_{ij} représente la détermination de la variable X_j étant J=1,...m sur les unités U_i étant i=1,... n. Soit Sn,m le correspondant barème des valeurs standardisées et di l'écart plus opportun. On établit m matrices des écarts. Dans ce but on a utilisé la technique du regroupement « Cluvar » (Skonieczny G, 1995).

L'utilisation des groupes de variables a permis une analyse thématique du phénomène et de ses articulations sur le territoire ne permettant pas, toutefois, une vision d'ensemble considérée comme analyse de chaque aspect et des éventualités produites par l'intégration entre celles-ci (Torrisi, 2000).

Souvent l'accumulation de causes concomitantes produit des manifestations et des comportements tout à fait écartées des communs barèmes d'interprétation. On a engagé des indicateurs composés construits en utilisant des moyennes pondérées à travers les pois factoriels des 55 variables (indicateur simples) prises en examen. La cluster analisys a été appliquée sur le total des 10 macro indicateurs composés dans le but d'établir un tableau sur la diversification territoriale du système économique-financier aussi bien que sur les groupes de variables (indicateurs simples) de chaque macro indicateur, le tout dans le but de construire les différences régionales relatives aux différents aspects pris en examen par chaque macro indicateur : sphère réelle, densité et productivité bancaire, capacité de financement des activités de production, intermédiation bancaire, etc.

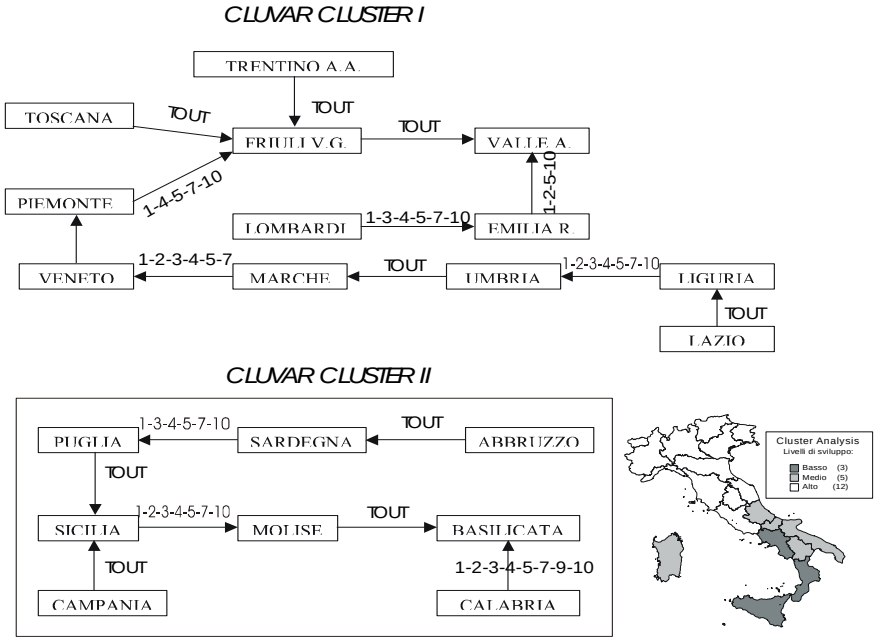
5. Conclusions

L'analyse des cluster (appliquée aux composantes principales obtenues par le moyen de l'analyse factorielle conduite sur 10 groupes d'indicateurs régionaux) a permis d'obtenir un tableau du système financier national qui met en évidence la divergence entre le Nord et le Midi de notre Pays. Les éléments structuraux qu'on a quantifiés dans la phase précédente, offrent un tableau de synthèse sur les différents aspects du système financier des régions italiennes. Généralement, après avoir fait un premier examens des groupes, il résulte, aussi dans la sphère financière, la subdivision traditionnelle du pays en trois aires bien distinguées, d'où prévaut un Nord caractérisé par un système de crédit compétitif en train d'augmenter sensiblement. La technique de classement appliquée sur le total des indicateurs financiers a mis en évidence une Italie à trois vitesses. S'il y a d'un côté un partiel changement des groupes, par contre les résultat d'ensemble sont généralement les mêmes : Ce phénomène contribue partiellement à confirmer la validité des facteurs déterminants, en tant qu'indicateurs synthétiques représentatifs des divergences financières entre les différentes régions. Le premier groupe englobe les régions Centre-Nord, caractérisées dans leur ensemble par la présence d'un grand nombre de banques très productives. Ce sont les régions du NEC qui ont une importance décisive, étant donné les valeurs très élevées du facteur « dimension des entreprises de crédit ». L'activité d'intermédiation bancaire résulte être bien élevée parce qu'elle consent à la clientèle de choisir à l'intérieur d'une vaste gamme d'instruments financiers et de moyens de règlement, même parmi les plus nouveaux et qu'elle garantit une seule base pour les nécessités financières des nombreuses entreprises présentes dans l'aire.

On trouve une situation diamétralement opposée dans les régions regroupées dans le deuxième et le troisième cluster prenant en analyse l'aire du Midi. Le système financier de cette répartition est caractérisé par un relatif gap entre la structure réelle et le système financier. Les entreprises de crédit sont largement dispersées sur le territoire régional. Chaque guichet sert en moyenne un nombre considérables de clients et englobe un territoire très étendu. Dans l'aire est prédominante l'activité des banques de grandes dimensions, ayant peut-être leurs sièges dans les centres financiers du Nord. En outre il résulte une faible productivité bancaire aussi bien que l'intermédiation des entreprises de crédit. Tout cela peut contribuer à expliquer la faible activité de financement de l'économie, particulièrement dangereuse,

confiée surtout aux grandes entreprises financières du Nord (à remarquer la haute dépendance du Midi du crédit extérieur), et à expliquer aussi le fautif recours aux investissements en des activités financières (telles que titres de l'Etat) et enfin une insuffisante utilisation des plus modernes instruments d'encaissement et de règlement. L'analyse comparée du secteur financier a confirmé l'existence d'un persistant écart entre le Nord et le Midi, en mettant en évidence un type de classement qui, dans l'ensemble, résulte presque analogue à celui qui a été obtenu pour le développement de l'économie.Cette divergence ne manifeste pas des signes évidents de changement.

Figure 1 – Diagram cluvar cluster



6. Bibliographie

[GAL 90] GALLI G. *Il sistema finanziario nel mezzogiorno,* (a cura di), Banca d'Italia, Roma(1990).

[PAR 95] PARILLO F. *I nuovi orizzonti del sistema finanziario italiano: sfide del mercato e cambiamento,* in *"sistema finanziario e governo del cambiamento"* 50°anniversario della rivista bancaria - ed. Minerva Bancaria(1995).

[PAR 98] PARILLO F. *Condizioni, strategie e prospettive per al crescita e la stabilità del sistema creditizio italiano,* in "rivista bancaria", ed. Minerva Bancaria n.3 (1998).

[SKO 95] SKONIECZNY G. *Cluvar plus: un metodo di classificazione con la selezione delle variabili,* in Giornate di analisi dei dati multidimensionale, (1995)

[TOR 00] TORRISI B. *Economie di scala e di scopo. Un'applicazione della Hybrid traslog cost function al sistema bancario siciliano,* in Bancaria Editrice N.7/8 luglio/agosto(2000).

Analyse des données de l'expression génomique par la classification

Pourquoi et comment ?

TALLUR Basavanneppa

IRISA, Université de Rennes 1,
Campus Universitaire de Beaulieu,
Avenue de Gen. Leclerc, 35042 Rennes cedex, France
E-mail: tallur@irisa.fr

ABSTRACT. *After sequencing and identification of genes, Biology has entered the post-genomic era. Thanks to the reccent advances in microarray technology and new developments made in hybridization techniques, the biologists are able to study the expression patterns of several thousands of genes in a variety of experimental conditions and in different types of cells simultaneously. The analysis of a huge quantity of data thus produced is useful for providing a general view of the basic biological processes and to identify genes of interest. Cluster analysis (also called unsupervised classification) is useful in reducing the complexity of the data and in formulating hypotheses that can be tested.*

RÉSUMÉ. *Après le séquençage et l'identification du génome, la biologie est entrée dans l'ère post-génomique. Grâce aux énormes progrès réalisés récemment dans la technologie des "puces à ADN" ou des "biopuces" et de l'hybridation, les biologistes peuvent maintenant réaliser des expérimentations sur plusieurs milliers de gènes, exprimés dans différentes cellules, dans des conditions d'expérimentations variées et à différents instants. L'analyse de la quantité impressionnante des données ainsi générées doit aider à comprendre les mécanismes de régulation des génes. La classification, aussi bien des gènes que des expériences (conditions), se révèle un outil efficace pour ces analyses. Nous faisons dans cet article un "inventaire" des différentes méthodes utilisées par les biologistes ainsi que des outils logiciels disponibles notamment sur le web. Nous insistons surtout sur les problèmes liés aux données elles-mêmes :erreurs systématiques, données manquantes.*

KEYWORDS: *microarray experiments, gene expression, hierarchical classification, data normalisation, missing data*

MOTS-CLÉS : *microarray, biopuces, données de l'expression, data mining, classification hiérarchique, normalisation des données, données manquantes*

1. Introduction

Grâce aux énormes progrès réalisé dans la technologie des biopuces et dans les techniques de l'hybridation il est possible d'étudier simultanément l'expression de plusieurs milliers de gènes dans des conditions expérimentales diverses ainsi que dans des cellules de types différents. Il est devenu nécessaire d'adapter les méthodes d'analyse des données pour analyser les données de l'expression en vue de comprendre le fonctionnement et les mécanismes de régulation de gènes. Les données sont présentées dans un tableau rectangulaire dont les lignes sont indexées par les gènes et les colonnes par les conditions, d'une façon générale. Parmi les objectifs de l'analyse des données de l'expression figure celui d'identifier les gènes co-exprimés. La classification des gènes est donc un outil qui se révèle très efficace car il existe une forte association entre le "profil d'expression" et la fonction du produit de gène et il est possible de déterminer la fonction des gènes regroupés par la classification ([BRO 99]). L'application de la classification des données de biopuce a été utilisée au début pour ré-ordonner les lignes et/ou les colonnes de façon à pouvoir visualiser les associations en coloriant les valeurs de l'expression (ce tableau colorié est souvent appelé "heat map") ([EIS 98]). Le regroupement des gènes dans une même classe permet de faire des hypothèses sur leur fonction commune.

2. Prétraitement des données

Compte tenu de la nature des expérimentations et des technologies des biopuces, les données obtenues ne peuvent être analysées directement sans avoir été préalablement ”pré-traitées”. Il s’agit de transformation, ”standardisation” ou ”normalisation”, correction de biais, filtrage et, éventuellement, traitement des données manquantes. Au départ, on se contentait de centrer et de réduire les données par rapport à la moyenne (ou la médiane) et l’écart type d’échantillons (colonnes) et/ou moyennes (ou médiane) et écart type des gènes (lignes). Plusieurs chercheurs se sont penchés sur des problèmes de biais systématique et ont proposé des méthodes de correction (cf [XIA 02], [QUA 02], [AL 03], et [TRO 01] pour n’en citer que quelques uns). Il faut noter que les données sont obtenues à partir des expérimentations utilisant l’une des 2 technologies :

- ”oligo microarrays” générés par la technique de ”photolithographie” et fabriqué et commercialisé par Affymetrix Inc.;

- ”cDNA arrays” utilisant une technique de ”maillage” mécanique où chaque ”spot” reçoit du matériel génétique (séquence d’ADN) par un système de ”jet d’encre”. Les données sont obtenues sous forme d’intensité de la luminance des différentes couleurs (employées pour étiqueter les échantillons ”hybridisés”) par les techniques d’analyse d’image.

Les données obtenues (rapports de fluorescence de couleurs, typiquement rouge/vert) doivent être ”standardisées” ou ”normalisées” différemment en fonction de la technologie biopuce utilisée pour les produire. Il semble y avoir de cofusions quant à l’emploi des termes ”standardisation” et ”normalisation”. Pour certains, le terme ”normalisation” désigne l’étape de correction d’image avant d’évaluer l’intensité de fluorescence (ces logiciels sont fournis par le fabricant de microarrys), et ”standardisation”, celle de transformer les données de façon à rendre les échantillons comparables par l’analyses statistique. Pour d’autres, les deux termes sont synonymes de ”standardisation”. La première transformation consiste à utiliser la fonction log (base 2) des rapports (appelé ”log-ratio”) afin d’accorder la même importance aux sur-expressions qu’aux sous-expressions. Ces log-ratios sont fortement dépendantes de l’intensité du signal: la variance des log-ratios augmente avec l’intensité du ”spot”. On peut énumerer quelques sources de biais dans les données de l’expression : les quantités variables d’ARN présentes dans les échantillons de départ, les différences d’étiquetage et d’efficacité de détection des couleurs employées, biais systématiques des mesures, et le biais spatial sur les ”microarrays”. Les biais dus à l’intensité sont détectés à l’aide d’un graphique, appelé ”R-I plot”, et corrigés à l’aide d’un facteur de correction estimé par la méthode ”lowess” (cf [TWU 01]). Après avoir ”corrigé” les données on doit les standardiser (gène par gène ou échantillon par échantillon) en vue de les rendre comparables statistiquement. Il s’agit en général de les centrer par rapport à la moyenne (ou la médiane) et de réduire par rapport à l’écart type.

Le filtrage des données a pour but la détection et l’élimination des gènes non exprimés ou exprimés de façon ”plates” à travers les échantillons (conditions). On utilise la méthode classique du test de Student ou d’analyse de la variance, ou tout simplement en calculant les seuils basés sur la variance du niveau d’expression.

Un autre aspect important du pré-traitement des données concerne la gestion des données manquantes. Plusieurs stratégies sont proposées :

- filtrer (éliminer) les gènes ayant plus d’un certain pourcentage de données manquantes (ce n’est pas idéal) ;

- remplacer les données manquantes par des zéros (très dangeureux) ;

- remplacer les données manquantes par des moyennes (ou des médianes) des données disponbles ;

- estimer les données manquantes par la méthode des voisins les plus proches ;

- estimer les données manquantes par la méthode de la ”décomposition en valeurs singulières” (formule de reconstitution de données à partir des facteurs)

3. Les méthodes de classification et ressources sur le web

La classification des gènes est un outil très utile pour obtenir une vue générale des processus biologiques de base dans une cellule donnée ou dans un tissu donné. Le regroupement de gènes dans une même classe permet aux biologistes d'identifier les gènes co-exprimés et de formuler l'hypothèse que ces gènes ont une fonction similaire.

Les méthodes d'analyse classificatoire utilisées pour des données de l'expression sont nombreuses et variées et certaines d'entre elles sont très originales. Il existe de nombreux serveurs web qui proposesnt des outils d'analyse. Il existe des sites web qui proposent d'analyser vos données après avoir effectué le pré-traitement.

Parmi de très nombreux sites proposant soit des serveurs soit des logiciels à télécharger, on peut citer les suivants :

- EPCLUST de E.B.I. (http://www.ebi.ac.uk/EP/EPCLUST) ;
- Eisen Lab (http://rana.lbl.gov/EisenSoftware.htm) ;
- C.N.I.O. (http://gepas.bioinfo.cnio.es) ;
- Genesis (http://genome.tugraz.at) (cf [AST 00]) ;
- Engene (http://www.engene.cnb.uam.es) (cf [NAVA 02]) ;
- Stanford Medical informatics (http://smi-web.stanford.edu/projects/helix/pubs/impute) (cf [TRO 01]) ;
- J-Express (http://www.molmine.com/frameset/frm_jexpress.htm) ;
- Clusfavor (http://mbcr.bcm.tcm.edu/genepi/) ;
- Clustarray (http://www.cbs.dtu.dk/services/DNAarray/index.html)

La plupart de ces sites proposent des méthodes de classification hiérarchique classiques, utilisant la distance euclidienne ou celle de la corrélation et les critères d'agrégation tels que le lien minimum, le diamètre, le lien moyen et le critère de Ward. Certains sites proposent aussi les méthodes non-hiérarchiques (telles que k-means, k-means floues (proposée par [GE 02]), cartes de Kohonen et réseaux neuroneaux). Nous avons proposé une méthode de classification hiérarchique originale, AVL (Analyse de la Vraisemblance du Lien), basée sur un indice de similarité probabiliste reflétant la "vraisamblance" de la ressemblance pour analyser les données de l'expression (voir [LER 94] pour une présentation de la méthode ainsi que le logiciel "chavl"). Cette méthode a été employée pour la classification des séquences protéiques appartenant aux diverses familles (cf par exemple, [LNTP 94], [TN 98], [TCA 99]). Les expériences sont en cours sur des données de l'expression en vue de comparer les résultats avec d'autres méthodes.

4. Conclusion

Nous avons fait un état de l'art dans le traitement et l'analyse des données issues des expériences biopuces et nous proposons notre solution, en ce qui concerne la classification basée sur la méthode "AVL" (Analyse de la Vraisemblance du Lien). Nous avons souligné l'importance du pré-traitement des données de l'expression issues des expériences biopuce. Il est nécessaire d'adapter les méthodes d'analyse des données pour tenir compte de la particularité de la technologie des biopuces produisant ces données, et qui va continuer d'évoluer.

5. References

[AL 03] BOLSTAD B. M.,IRIZARRY R. A.,ASTRAND M., SPEED T. P. "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias", *Bioinformatics*, vol. 19, 2003, p. 185–193.

[BRO 99] BROWN P., BOTSTEIN D., "Exploring the new world of genome with DNA microarrays", *Nature genetics*, vol. 21, 1999, p. 33–37.

[EIS 98] EISEN M., SPELLMAN P., BROWN P., BOTSTEIN D., "Cluster analysis and display of genome-wide expression patterns", *Proc. Natl. Acad. Sci.*, vol. 95, 1998, p. 14863–14868.

[LER 94] LERMAN I. C., PETER P., LEREDDE H., “Principes et calculs de la méthode implantée dans le programme CHAVL (Classification Hiérarchique par Analyse de la Vraisemblance des Liens)”, *La revue de Modulad*, vol. 12, 1994.

[QUA 02] QUACKENBUSH J., “Microarray data normalization and transformation”, *Nature genetics*, vol. 32, 2002, p. 496–501.

[TRO 01] TROYANSSKAYA O., ET AL, “Missing value estimation methods for DNA microarrays”, *Bioinformatics*, vol. 17, 2001, p. 520–525.

[XIA 02] XIANG C., KERR M., CHURCHILL G., “Data transformation for cDNA microarray data”, report , 2002, The Jackson Laboratory, Bar Harbour, Maine, USA.

[TWU 01] WU THOMAS D., “Analysing gene expression data from DNA microarrays to identify candidate genes ”, *Journal of Pathology*, vol. 195, 2001, p. 53–65.

[AST 00] STURN A., “Genesis : cluster analysis of microarray data ”, *Bioinformatics*, vol. 18, 2000, p. 207–208.

[NAVA 02] NAVA G., SANTAELLA D. F., ALBA J. C., CARAZO J. M., TRELLES O., PASCUAL-MONTANO A. “Engene : the processing and exploratory analysis of the gene expression data ”, *Bioinformatics*, vol. 19, 2002, p. 657–658.

[LNTP 94] LERMAN I. C., NICOLAS J., TALLUR B. , PETER PH., “Classification of aligned biological sequences”, *New approaches in classification and data analysis. Springer-Verlag*, 1994, p. 370–377.

[TN 98] TALLUR B., NICOLAS J., “A method for classifying unaligned biological sequences”, *Data science, classification and related methods. Springer-Verlag, Tokyo*, 1998, p. 758–765.

[TCA 99] TALLUR B., NICOLAS J., FROGER A., THOMAS D., DELAMARCHE C. “Sequence classification of water channels and related proteins in view of functional predictions ”, *Theoretical Chemistry Accounts*, vol. 101, 1999, p. 77–81.

[GE 02] GASCH A., EISEN M. B. “Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering ”, *Genome Biology*, vol. 3(11), 2002.

Extraction de motifs fréquents multi-supports Application aux données symboliques

Tao Wan, Karine Zeitouni

Laboratoire PRiSM
Université de Versailles
45 avenue des Etats-Unis,
78000 Versailles Cedex
Tao.Wan@prism.uvsq.fr, Karine.Zeitouni@prism.uvsq.fr

RÉSUMÉ. *Cet article propose une nouvelle méthode d'analyse de données symboliques. Cette méthode cherche des sous-ensembles fréquents inter-dimensionnels à partir des données symboliques et permet aux utilisateurs de spécifier des seuils de support différents pour certaines dimensions en fonction des besoins des applications. Ainsi, elle combine les avantages de l'analyse de données symboliques et de la fouille de données. En outre, la méthode assiste l'utilisateur dans le choix des différents seuils de support.*

MOTS-CLÉS : *fouille de données, analyse de données symboliques, sous-ensembles fréquents.*

1. Introduction

Nouveau paradigme de description de données, les données symboliques permettent de définir des structures de données complexes, car elles permettent de capturer les variations internes comme les distributions ou les intervalles. En entré, les données symboliques peuvent prendre en compte de nombreux types de sources et résumer de grands ensembles de données en données de taille plus petite et plus gérable. "L'analyse de données symboliques" [BOC 00] est une extension de l'analyse de données standard appliquée aux données symboliques.

Face à l'explosion récente de sources de données, ce formalisme est bien adapté au support de sources complexes et multimédia. Dans ce contexte, nous nous intéressons au problème de recherche de motifs fréquents couramment répandue en fouille de données, mais non encore développée pour des données symboliques. La seconde extension est l'aspect *multi-support*. En effet, la recherche de motifs fréquents conventionnelle recherche les sous-ensembles fréquents satisfaisant un seuil de support uniforme. Cette contrainte d'uniformité pose souvent des problèmes car d'un côté, le choix d'un seuil de support trop élevé ne retourne pas certains motifs fréquents intéressants et de l'autre, le choix d'un seuil de support trop bas produit une quantité de résultats trop grande à charge de l'utilisateur d'en retenir les motifs intéressants.

La méthode de recherche de motifs fréquents multi-supports proposée a pour objectifs:

- de traiter les données symboliques
- de mieux prendre en compte les besoins utilisateurs en leurs permettant d'affiner le choix des seuils de support par variable.

L'intérêt de l'extension de cette méthode de fouille de données aux données symboliques offre au moins deux avantages. Puisque le paradigme des données symboliques permet d'agréger les données

en une description plus synthétique du fait à analyser en minimisant la perte d'informations, extraire des motifs fréquents de ces données donne un résultat plus significatif pour ces faits que s'ils n'étaient pas agrégés. Le second avantage est que la sortie de l'analyse de données symboliques est elle-même symbolique et peut donc être prise en entrée d'autres méthodes d'analyses.

1.1. *Extraction de sous-ensembles fréquents*

L'extraction d'associations introduite dans [AGR 93] permet de retrouver les fréquences de co-occurrences d'articles dans des listes d'articles. L'application par excellence est l'analyse dite du panier de la ménagère où l'on s'intéresse aux règles d'association entre articles achetés dans les transactions (ou « paniers »). Elle se base sur la recherche d'ensembles fréquents dans ces listes en calculant deux principales mesures : le support donné par la fréquence de l'ensemble rapportée au nombre de listes et la confiance mesurée par la fréquence d'un sous-ensemble de l'ensemble d'article rapportée à celle de son complément.

Depuis son introduction, le problème d'extraction de règles d'association dans de grandes bases de données et le problème plus général de recherche d'ensembles fréquents a été l'objet de nombreuses études. Ces études peuvent être classées en deux familles (cf. [LAK 00]) :

- visant le passage à l'échelle (*scalabilité* en anglais) : la question centrale est d'optimiser le coût de l'algorithme de recherche des règles d'association pour de gros volumes de données.
- étendant la fonctionnalité : la question centrale est de savoir quelle sorte de règles calculer en étendant la méthode traditionnelle ou en proposant d'autres mesures de pertinences que le support et la confiance.

Généralement le seuil de support est fixe, hormis l'extraction de règles d'association multi-niveaux [HAN 95] où l'utilisateur peut spécifier des seuils de supports à différents niveaux d'une hiérarchie de concepts prédéfinie. Dans le problème que l'on se pose, l'utilisateur peut donner un seuil de support différent par type d'article et non par niveau de concept. En effet, dans les applications réelles, il est rare que tous les types d'articles aient des fréquences proches.

2. Extraction de motifs fréquents de données symboliques

Cette méthode traite les données symboliques catégorielles, que ce soit des valeurs simples ou des multi-valeurs. Elle comprend trois phases décrites dans la suite. Mais tout d'abord, on définit la notion de fréquent multi-supports.

2.1. Fréquent multi-supports

Exemple1 : Soit un sous-ensemble d'articles $I=\{l_1, l_2, l_3\}$ et les supports minimums de chaque article appelés *min-support* ou MIS: $MIS(l_1) = 10\%$, $MIS(l_2) = 20\%$, $MIS(l_2) = 30\%$. I est fréquent multi-supports ssi $support(l_1) \geq 10\%$, $support(l_2) \geq 20\%$, $support(l_3) \geq 30\%$, $support(l_1, l_2, l_3) \geq 10\%$ et $support(l_2, l_3) \geq 20\%$.

Définition 1 (fréquent multi-supports) : Un sous-ensemble d'articles I est fréquent multi-supports ssi chaque sous-ensemble de I satisfait le plus petit *min-support* de ses articles.

2.2. *Conversion des données symboliques en données transactionnelles*

L'extraction de motifs fréquents a largement été étudiée dans les recherches en fouille de données donnant lieu à des algorithmes de référence telles que **Apriori**, **FP-growth**, **TreeProjection**, etc.

Toutes ces méthodes se basent sur une représentation en base de données transactionnelles[1]. Afin de réutiliser et d'adapter ces algorithmes, une première phase de la méthode proposée consiste à convertir les données symboliques en données transactionnelles.

2.3. *Extension de la structure de données FP-Tree (eXtended Frequent Pattern : XFP-tree)*

Une des méthodes d'extraction des motifs fréquents, FP-growth, consiste à construire d'abord une base de données compressée appelée FP-tree [HAN 00] dans laquelle chaque transaction est représentée par un chemin dans l'arbre et les comptages de fréquences sont reportés avec les articles dans les nœuds. Cette caractéristique est bien adaptée à notre problème car elle distingue le support par article.
Nous construisons d'abord une extension de FP-tree, appelée XFP-tree, dont l'organisation des nœuds est définie par un ordre relatif qui nous permet d'optimiser l'exploration des sous-ensembles fréquents.

Définition 2 (Ordre relatif) : Un ordre relatif dans un XFP-tree est une organisation ordonnée de ses nœuds. Il est établi en plaçant les nœuds (articles) dans l'arbre selon l'ordre ascendant des MIS de ces articles; si deux articles ont le même MIS, ils sont alors ordonnés par ordre descendant de leurs fréquences.

Exemple 2 : Soit un XFP-tree tel que les MIS des articles sont : $MIS(a_1) = MIS(a_3) = 40\%$, $MIS(b_2) = 20\%$, $MIS(c_1) = 45\%$, $MIS(d_1) = 30\%$. On prend comme ensemble d'articles $L = [a_1{:}8, c_1{:}6, b_2{:}5, a_3{:}5, d_1{:}5]$ dans lequel chaque article satisfait son seuil de support. L'ordre relatif des éléments fréquents donne : $b_2 > d_1 > a_1 > a_3 > c_1$.

2.4. *Extraction des motifs fréquents à l'aide d'un XFP-tree*

La construction d'un XFP-tree comprenant un ordre relatif d'éléments fréquents fait que cette structure est moins compacte que celle du FP-tree d'origine. Cependant, elle s'avère très efficace lors de l'extraction des motifs fréquents grâce à l'ordre relatif. Afin d'explorer les informations contenues dans le XFP-tree lorsque les seuils de supports (MIS) ne sont pas uniformes, nous proposons d'étendre l'algorithme FP-growth pour extraire l'ensemble des motifs fréquents. Pour ce faire, nous nous basons sur le lemme suivant :

Lemme 1 : Soit *a* un sous-ensemble fréquent multi-supports, B un ensemble de chemins de l'arbre XFP-tree incluant *a* et soit β un élément de B. Si *a* est fréquent et le support de β satisfait MIS (β), alors l'ensemble d'éléments $<a,\beta>$ est également fréquent.

Ainsi, à partir d'un sous-ensemble fréquent multi-supports, on génère des sous-ensembles fréquents multi-supports plus grands. Pour plus de détail, se référer à l'article complet dans [ref BDA].

L'exemple ci-dessous illustre la procédure générant tous les sous-ensembles fréquents multi-supports comprenant C1 :

[1] Une base de données transactionnelles est un ensemble de sous-ensembles d'articles (données catégorielles) d'une même transaction.

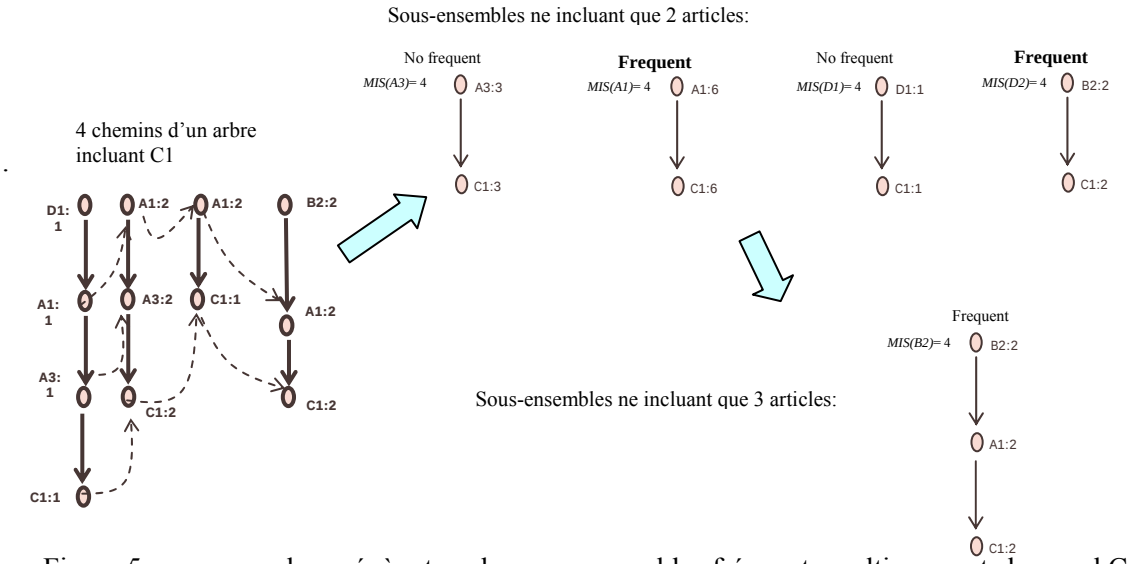


Figure 5: une procedure génère tous les sous-ensmebles fréquents multi-supports le nœud C1

3. Expérimentation

Cette méthode a été implémentée et testée sur des données symboliques relatives aux accidents routiers et dont certaines variables telles que les impliqués ou le voisinage sont multi-valuées. En fonction de la variable, la fréquence des valeurs peut varier énormément. Lors d'une première phase, l'outil affiche ces fréquences afin d'aider l'utilisateur à estimer les bons supports par variable. Ensuite, l'algorithme retourne les seuls fréquents vérifiant les critères de support donnés par l'utilisateur. Le temps de réponse est quasi-instantané sur près de 30000 individus. Les résultats des tests ont bien validé la méthode d'un point de vue fonctionnel et sur le plan des performances.

Bibliographie

[BOC 00] H.-H.BOCK, E. DIDAY. Analysis of Symbolic Data, Springer-Verlag Edition, 2000.

[AGR 93] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. *In Proc.* 1993 SIGMOD, pp. 207-216.

[LAK 00] L. V. S. Lakshmanan, C. K.-S. Leung, R. T. Ng. The Segment Support Map: Scalable Mining of Frequent Itemsets. *In Proc*. 2000 SIGKDD, pp. 21-27.

[HAN 95] J. Han and Y. Fu. Discovery of multiple-level association rules from large databases. In Proc. 1995 VLDB, pp. 420-431.

[HAN 00] J. Han, J. Pei, and Y.Yin. Mining Frequent Motifs without Candidate Generation. In SIGMOD, 2000.