RAIRO. RECHERCHE OPÉRATIONNELLE

A. MKHADRI
F. MARCHETTI

Pondération des variables pour la classification de données binaires

RAIRO. Recherche opérationnelle, tome 25, n° 4 (1991), p. 381-401

<http://www.numdam.org/item?id=RO_1991__25_4_381_0>

Recherche opérationnelle/Operations Research
(vol. 25, n° 4, 1991, p. 381 à 401)

PONDÉRATION DES VARIABLES POUR LA CLASSIFICATION DE DONNÉES BINAIRES (*)

par A. MKHADRI (1) et F. MARCHETTI (2)

Résumé. – *Cet article traite de la classification de données binaires et nous nous intéressons particulièrement à la pondération des variables. Nous considérons le problème de la classification en tenant compte du caractère spécifique des données binaires. Nous examinons l'influence d'un système de pondération pour classer les données et nous en tirons une variante de l'algorithme MNDBIN de Govaert* [10]. *Puis nous présentons et analysons des méthodes de pondération adaptatives des variables dans le cadre de la classification. Nous terminons par une étude comparative de cette approche avec l'approche probabiliste sur deux échantillons de données simulées et réelles.*

Mots clés : Classification binaire, distance $L1$ pondérée, pondérations adaptatives, modèle probabiliste.

Abstract. – *This paper deals with binary clustering and we are particularly concerned with the ponderation of the variables. We place the binary clustering problem in the framework defined by Govaert* [10]. *He defined a specific algorithm for this type of data, called MNDBIN, that we briefly describe in section 2. Next, we show that choosing a ponderation vector for the variables can be judicious to classify the data. Then we present and analyze some adaptable ponderation methods for clustering. We end by a comparative study of this approach with a probabilistic model using two samples of simulated and real data.*

Keywords : Binary clustering, weighted $L1$ distance, adaptable ponderations, probabilistic model.

INTRODUCTION

Les méthodes de classification automatique non hiérarchiques ont pour but de construire des partitions d'ensembles d'objets décrits par des variables de telle sorte que les classes soient le plus homogènes. Elles reposent essentiellement sur la définition d'une métrique et d'un critère associé à celle-ci.

(*) Reçu juillet 1990.
(1) I.N.R.I.A.-Rocquencourt, Domaine de Voluceau, 78153 Le Chesnay Cedex, France.
(2) L.R.I.M. Université de Metz, Ile-du-Saulcy, 57045 Metz Cedex.

Recherche opérationnelle/Operations Research, 0399-0559/91/04 381 21/$ 4.10

La Méthode des Nuées Dynamiques (Diday *et al.* [4]) utilise de plus la notion de noyau associé à chaque classe. Dans le cas de p variables quantitatives, ces noyaux sont des éléments de $\mathbf{R}^p$ contenant l'ensemble à classifier. Lorsque les variables ne sont pas quantitatives, il est possible de se ramener au cas précédent au moyen d'un codage (l'autre représentation possible des variables qualitatives est la représentation ensembliste et combinatoire). Les noyaux fournis par la méthode ont alors une structure différente des données initiales.

Pour lever cet inconvénient, Marchetti [15] et Govaert [10] proposent un cadre particulier aux données binaires, que nous rappelons dans le premier paragraphe. En suivant cette approche, un algorithme (MNDBIN) adapté aux données binaires a été développé. Celui-ci est caractérisé par l'utilisation de la distance $L1$ et de noyaux ayant la même structure que les données initiales. Nous examinons ensuite, au paragraphe 2, comment un système de pondération peut intervenir pour classer les données. Ainsi, des méthodes de pondération adaptatives ont été développées dans le même cadre. Nous les présentons et analysons au paragraphe 3. Le paragraphe 4 présente l'approche probabiliste définie par Govaert [10] qui permet de justifier, *a posteriori*, l'approche métrique. Nous terminons par une étude comparative de ces deux approches sur deux applications de données réelles et simulées.

1. INERTIE SUR DES DONNÉES BINAIRES

Ω étant un ensemble de n individus mesurés par p variables binaires, on cherche une partition de Ω en K classes « homogènes », K fixé *a priori*. Les données étant binaires, on peut, sans restriction, supposer que les valeurs prises par les variables sont 0 et 1. L'ensemble à classifier est contenu dans $\mathbf{B}^p = \{0, 1\}^p$. La proximité entre deux points est mesurée par la distance $L1$ ou distance city-block, notée d et définie par :

$$\forall\, x \text{ et } y \in \mathbf{B}^p, \quad d(x, y) = \sum_{j=1}^{p} |x^j - y^j|.$$

Une première approche consiste à munir chaque élément x de Ω d'une pondération $\alpha(x) \in \mathbf{R}^+$. On considère alors le nuage binaire pondéré de n points de $\mathbf{B}^p$:

$$\mathbf{N}(\Omega) = \{(x, \alpha(x)),\ x \in \Omega\}$$

A chaque variable j, on peut associer a^j la valeur majoritaire 0 ou 1 de l'ensemble $\{x^j,\ x \in \Omega\}$ compte tenu des pondérations. Cette valeur correspond

à la notion usuelle de médiane qui est ici toujours binaire. Dans le cas particulier où la somme des pondérations des 0 est la même que celle des 1, tout l'intervalle [0, 1] convient et on peut donc prendre arbitrairement une des deux valeurs binaires. On appellera centre médian de Ω le vecteur $a=(a^1, \ldots, a^p)$ appartenant à $\mathbf{B}^p$, ainsi défini.

DÉFINITION 1.1 : *On définit l'inertie au sens de la distance L1 du nuage $N(\Omega)$ par rapport à un point b de $\mathbf{B}^p$ par :*

$$\Im(\Omega/b)=\sum_{x\in\Omega}\alpha(x)\,d(x,\,b)=\sum_{x\in\Omega}\alpha(x)\sum_{j=1}^{p}|x^j-b^j|$$

En utilisant la propriété d'optimalité de la médiane pour la distance $L1$, on peut alors montrer que le centre médian est le point d'inertie minimale.

Cependant, cette approche ne permet pas d'établir des propriétés analogues à celles existant sur l'espace $\mathbf{R}^p$ muni de la distance euclidienne usuelle (propriété de conservation du centre de gravité, relation de Huyghens et relation de décomposition de l'inertie). Pour cette raison, nous proposons de remplacer les pondérations précédentes par des vecteurs de pondération $\alpha(x)=(\alpha^1(x), \ldots, \alpha^p(x))$ appartenant à $\mathbf{R}^{p+}$ associés à chaque sommet observé de $\mathbf{B}^p$. On peut alors définir une extension des notions de centre médian et d'inertie que l'on vient de définir.

A chaque variable j, on peut maintenant associer a^j la valeur majoritaire 0 ou 1 minimisant $\sum_{x\in\Omega}\alpha^j(x)|x^j-a^j|$. La différence avec la situation précédente est que l'ensemble des pondérations varie suivant la variable. On appellera toujours centre médian de Ω le vecteur $a=(a^1, \ldots, a^p)$ ainsi défini. Si toutes les pondérations des différentes composantes d'un même élément de Ω sont égales, on retrouve la notion précédente.

On associe à ce centre médian le vecteur de pondération suivant :

$$\alpha^j(a)=|n^j_\Omega(1)-n^j_\Omega(0)|$$

où

$$\mathrm{n}^j_\Omega(1)=\sum_{x\in\Omega}\alpha^j(x)\,x^j \qquad \text{et} \qquad n^j_\Omega(0)=\sum_{x\in\Omega}\alpha^j(x)\,(1-x^j).$$

La composante j du vecteur de pondération exprime la différence entre le nombre d'éléments de Ω ayant pris la valeur 1 pour la variable j et le nombre d'éléments de Ω ayant pris la valeur 0 (compte tenu des pondérations). Cette

approche permet de montrer (*cf.* Marchetti [15]) une propriété de conservation du centre médian analogue à celle du centre de gravité.

PROPRIÉTÉ 1.1 : *Si* $(P_1, P_2, \ldots, P_K)$ *est une partition de* Ω, *alors le centre médian de l'ensemble des centres médians des parties* $P_1, \ldots, P_K$ *munies de leurs pondérations respectives est le centre médian de* Ω.

Nous définissons maintenant une nouvelle inertie prenant en compte les vecteurs de pondération.

DÉFINITION 1.2 : *On appelle inertie d'un nuage* $\mathbf{N}(\Omega)$ *par rapport à un point* b *de* $\mathbf{B}^p$ *la quantité suivante :*

$$\mathfrak{I}(\Omega/b) = \sum_{x \in \Omega} \sum_{j=1}^{p} \alpha^j(x) \, | x^j - b^j |$$

Remarquons que si toutes les pondérations d'un élément de Ω sont égales, on retrouve là aussi la première définition. Il est également facile de montrer que le centre médian est toujours le point d'inertie minimale. On appellera alors inertie du nuage $\mathbf{N}(\Omega)$ l'inertie de ce nuage par rapport à son centre médian. On la notera $\mathfrak{I}(\Omega)$.

Il est maintenant possible d'établir une relation entre l'inertie d'un nuage par rapport à un point quelconque et l'inertie de ce nuage par rapport à son centre médian. On obtient alors une relation proche du théorème de Huyghens.

PSEUDO-THÉORÈME DE HUYGHENS : *Soient* a *le centre médian du nuage* $\mathbf{N}(\Omega) = \{ (x, \alpha(x)), x \in \Omega \}$ *et* b *un point quelconque de* $\mathbf{B}^p$. *On a :*

$$\mathfrak{I}(\Omega/b) = \mathfrak{I}(\Omega) + \mathfrak{I}(\{a\}/b) \qquad \text{où} \quad \mathfrak{I}(\{a\}/b) = \sum_{j=1}^{p} \alpha^j(a) \, | a^j - b^j |$$

On retrouve que l'inertie est minimale pour le centre médian du nuage. De plus, ce pseudo-théorème permet de démontrer une relation de décomposition de l'inertie.

PROPRIÉTÉ 1.2 : *Si* $P = (P_1, P_2, \ldots, P_K)$ *est une partition de* Ω *et si* $(a_1, a_1, \ldots, a_K)$ *sont les centres médians des classes, alors :*

$$\mathfrak{I}(\Omega/b) = \sum_{k=1}^{K} \mathfrak{I}(P_k) + \sum_{k=1}^{K} \mathfrak{I}(\{a_k\}/b)$$

Si on exprime cette relation par rapport au centre médian du nuage $\mathbf{N}(\Omega)$, *on obtient la relation habituelle :*

$$\textit{Inertie totale} = \textit{Inertie intraclasse} + \textit{Inertie interclasse}$$

A l'aide de ces propriétés, la méthode de classification pour données binaires MNDBIN va apparaître sous une forme plus conventionnelle. C'est ce que nous montrons dans le paragraphe suivant.

2. ALGORITHMES NON ADAPTATIFS

Rappelons rapidement le principe des méthodes des Nuées Dynamiques (Diday *et al.* [4]). On suppose que Ω est inclus dans un ensemble E (par exemple $\mathbf{R}^p$), on définit un ensemble $\mathbf{L}$ de noyaux, une distance $\mathbf{D}$ entre les éléments de E et les noyaux de $\mathbf{L}$. Le critère de classification $\mathbf{W}$ est alors le suivant :

$$\mathbf{W}(P, L) = \sum_{k=1}^{K} \sum_{x \in P_k} \mathbf{D}(x, a_k)$$

où $P = (P_1, \ldots, P_K)$ et $\mathbf{L} = (a_1, \ldots, a_K)$ avec $a_k \in \mathbf{L}$.

L'algorithme construit itérativement une suite $P^0, L^0, P^1, L^1, \ldots, P^n, L^n$ de partitions et de noyaux en minimisant à chaque étape le critère $\mathbf{W}$. On obtient ainsi à la convergence une partition avec comme résumé pour chaque classe le noyau associé.

Pour utiliser ce type d'algorithme dans le cas de données binaires, on peut, soit considérer que l'ensemble des données appartient à $\mathbf{R}^p$ muni de la distance euclidienne et appliquer la méthode des Nuées Dynamiques en prenant comme noyau des éléments de $\mathbf{R}^p$ (méthode des centres mobiles), soit se placer simplement dans l'ensemble Ω muni d'une distance quelconque et utiliser la méthode des Nuées Dynamiques sur tableau de distances (ce qui revient à imposer aux noyaux d'appartenir à Ω). Ces deux situations présentent des inconvénients. Dans le premier cas, le résumé de chaque classe et le critère sont difficilement interprétables par rapport aux données initiales; on ne tient pas compte de la forme particulière des données. Dans le second cas, cette fois l'appartenance des noyaux aux données initiales peut sembler trop restrictive et d'autre part cela va conduire à une certaine perte d'efficacité de l'algorithme en place mémoire et en temps (construction et utilisation du tableau des distances sur Ω).

Nous allons nous situer entre ces deux approches. Pour ceci, nous utiliserons la possibilité d'imposer des contraintes aux noyaux. En fait, nous allons respecter un principe d'homogénéité : les données à classer et les noyaux doivent être de même nature. L'ensemble Ω étant contenu dans $\mathbf{B}^p$, nous allons donc imposer aux noyaux d'appartenir à cet espace. La distance retenue représente le nombre de fois où les coordonnées correspondantes ne sont pas identiques. Sur l'espace $\mathbf{B}^p$, il s'agit exactement de la distance $L1$. On la notera d.

Pour pouvoir respecter ce principe d'homogénéité des résultats par rapport aux données, nous avons proposé un cadre d'étude. Celui-ci est donc initialement fixé et les propriétés précédentes vont trouver une application en matière de classification. Notre problème n'est pas de rechercher un espace de représentation euclidien ou non des données. De nombreux travaux vont en ce sens : à partir d'un ensemble d'individus et d'une dissimilarité, il s'agit de rechercher un espace de représentation qui soit le mieux possible adapté aux données. Si le cadre euclidien est le plus souvent utilisé, il ne convient pas toujours, notamment lorsque les données sont qualitatives. Certains auteurs proposent alors des représentations en terme de distance city-block (G. Le Calvé [13], B. Fichet [5]). A ce type de travaux, nous pouvons ajouter ceux de B. Fichet & G. Le Calvé [6] et S. Joly & G. Le Calvé [12] qui étudient la nature géométrique des dissimilarités et notamment du point de vue de leur structure métrique et euclidienne

2.1. L'algorithme MNDBIN

Cet algorithme de type Nuées Dynamiques fournit une solution au problème suivant :

Trouver une partition $P=(P_1, \ldots, P_K)$ *de* Ω *et un ensemble de* K *noyaux de* $\mathbf{B}^p$ *notés* $L=(a_1, \ldots, a_K)$ *minimisant le critère :*

$$\mathbf{W}(P, L)=\sum_{k=1}^{K} \sum_{x \in P_k} d(x, a_k)$$

L'algorithme se construit de la façon habituelle. La fonction d'affectation consiste à ranger chaque élément de Ω dans la classe dont le noyau est le plus proche. La fonction de représentation mise en évidence est particulière : les noyaux sont les centres médians des classes. Autrement dit, chaque variable dans chaque classe est représentée par la valeur 0 ou 1 la plus souvent choisie par les individus de la classe considérée. Chaque classe est ainsi décrite par un vecteur binaire facilement interprétable. Le critère obtenu à la convergence

représente simplement le nombre de fois où la situation obtenue s'écarte de la situation « idéale » (correspondant au cas où les individus sont identiques aux noyaux des classes auxquelles ils appartiennent).

A partir des propriétés énoncées dans le paragraphe 1, il est possible de donner un nouveau sens au critère **W**. Pour cela, associons à chaque point du nuage $\mathbf{N}(\Omega)$ un vecteur de pondérations toutes égales à $1\,[\alpha^j(x)=1$ pour tout $x\in\Omega$ et $j=1, \ldots, p]$. La relation de décomposition de l'inertie nous donne dans ce cas :

$$\mathfrak{I}(\Omega)=\mathbf{W}(P, L)+\mathfrak{I}(\{a_1, \ldots . a_K\})$$

Le critère $\mathbf{W}(P, L)$ optimisé par l'algorithme apparaît ici comme l'inertie intraclasse de la partition. De plus, un problème équivalent est celui de la maximisation du critère d'inertie interclasse du nuage des centres médians des classes. Ainsi présentée, cette méthode est analogue à une méthode des Nuées Dynamiques utilisant la distance euclidienne usuelle et les centres de gravité comme noyaux.

Remarquons que nous pouvons écrire une méthode plus générale prenant en compte des pondérations initiales quelconques. Nous montrons alors (*cf.* Marchetti [15]) que la méthode de classification croisée pour données binaires CROBIN (Govaert [9]) utilise l'algorithme MNDBIN généralisé comme algorithme intermédiaire. Ainsi, après avoir défini une mesure d'information binaire, notre approche permet de replacer CROBIN dans le cadre général de la classification croisée.

2.2. L'algorithme MNDBVP

Nous proposons maintenant d'utiliser un système de pondération particulier pour les variables.

Dans l'algorithme MNDBIN, on suppose (implicitement) que les variables ont le même poids, ce qui statistiquement peut être vu comme un défaut. Lerman [14] a proposé un indice de similarité entre objets décrits par p variables qualitatives nominales à deux modalités. Nous montrons, dans ce paragraphe, que l'utilisation, dans le cas des données qualitatives, d'un indice de dissimilarité associé à l'indice de similarité proposé par Lerman, revient à appliquer l'algorithme MNDBIN sur un tableau de données binaires avec une pondération particulière (et ayant un sens statistique précis) des variables.

L'indice de similarité global $S(i, i')$, défini par référence à une hypothèse d'absence de lien due à Lerman, entre deux individus i et i', qui tient compte

de toutes les variables, peut s'écrire :

$$S(i, i') = \sum_{j=1}^{p} \frac{s_j(i, i') - M^j}{\sigma_j}$$

où $s_j(i, i')$ est la contribution brute de chaque variable j à la « ressemblance » de deux individus i et i', qui est égale à 1 si i et i' ont même modalité et à 0 sinon. M^j et σ_j sont respectivement la moyenne et la variance empiriques de $s_j(i, i')$ qu'on peut écrire sous la forme (*cf.* Mkhadri [16]) :

$$M^j = \frac{n_{1j}(n_{1j}-1)}{n(n-1)} + \frac{n_{0j}(n_{0j}-1)}{n(n-1)} \quad \text{et} \quad \sigma_j^2 = \frac{2n_{1j}n_{0j}}{n^2}\left(1 - \frac{n_{1j}n_{0j}}{(n-1)^2}\right)$$

où n_{1j} est le nombre d'individus ayant pris la modalité 1 pour la variable j et n_{0j} est son complémentaire sur Ω (*i.e.* $n_{0j} = n - n_{1j}$). L'indice de dissimilarité D_1 associé à S est défini par :

$$D_1(i, i') = \sum_{j=1}^{p} \{1 - s_j(i, i')\}/\sigma_j$$

Il est clair que D_1 n'est autre que la distance $L1$ pondérée pour chaque coordonnée par $1/\sigma_j$. Ainsi, au tableau binaire Y, obtenu par transformation simple du tableau initial des variables qualitatives à deux modalités, on associe la distance Δ de type $L1$ pondérée par $1/\sigma_j$ pour chaque variable binaire j. On montre facilement, pour tout couple d'individus (i, i'), que $D_1(i, i') = \Delta(i, i')$. Ainsi, la recherche d'une solution au problème suivant :

Trouver une partition $P = (P_1, \ldots, P_K)$ de Ω, caractérisée par p variables qualitatives à deux modalités, et un ensemble de K noyaux de B^{2p} notés $L = (a_1, \ldots, a_K)$ minimisant le critère :

$$\mathbf{W}_1(P, L) = \sum_{k=1}^{K} \sum_{x \in P_k} D_1(x, a_k)$$

est équivalent à la résolution du problème de l'algorithme de type Nuées Dynamiques, noté MNDBVP, suivant :

Trouver une partition $P = (P_1, \ldots, P_K)$ de Ω, caractérisée par p variables binaires, et un ensemble de K noyaux de B^p notés $L = (a_1, \ldots, a_K)$ minimisant le critère :

$$\mathbf{W}_2(P, L) = \sum_{k=1}^{K} \sum_{x \in P_k} \Delta(x, a_k).$$

Donc la fonction de représentation est strictement la même que dans le cas non pondéré, seule change la fonction d'affectation du fait de la pondération $1/\sigma_j$ ($j=1, \ldots, p$) qui favorise les variables déséquilibrées (*cf.* Mkhadri [16]), mais ne varie pas au cours de l'algorithme.

Remarques 2.1 : Une manière classique de pondérer les variables consiste à travailler sur les variables réduites. Ici, les variables étant de Bernoulli, cela conduira donc à prendre comme poids $\alpha^j = 1/(V^j)^{1/2}$, où $V^j = n_{1j} n_{0j}/n^2$ qui est la variance de la loi de Bernoulli associée à la variable j. Soit L^q la variance associée à l'indice de Lerman, on peut faire l'approximation $L^q \cong 2 V^q \{1 - V^q\}$. L'étude de la variation de L^q en fonction de V^q montre que la pondération $1/(L^q)^{1/2}$ conduira à des poids moins contrastés par rapport à la pondération classique.

Par rapport à MNDBIN, MNDBVP vise à tenir compte des différences des fréquences entre les variables binaires.

3. ALGORITHMES ADAPTATIFS

Jusqu'à présent nous avons vu des critères qui fixaient la pondération des variables une fois pour toute et indépendamment de la classification. Dans ce paragraphe, nous allons examiner comment il est possible de faire dépendre cette pondération de la classification. L'intérêt d'une telle démarche est d'adapter au mieux le rôle des variables dans la formation des classes. Pour ce faire nous utilisons le même point de vue géométrique que l'on retrouve dans les algorithmes de type distances adaptatives pour la classification de variables continues (Friedman et Rubin [7], Govaert [9]) que nous décrivons rapidement ci-après.

Cas continu

Lorsque les objets de l'ensemble à classifier, sont décrits par p variables quantitatives, ils sont caractérisés par des vecteurs de $\mathbf{R}^p$ que l'on munit généralement d'une distance euclidienne d_M telle que :

$$\forall (x, y) \in \mathbf{R}^p \times \mathbf{R}^p, \quad d_M(x, y) = (x-y)^t M (x-y).$$

L'un des algorithmes le plus utilisé est l'algorithme des centres mobiles où la métrique M est fixée ($M = I_p$). Mais ce choix a tendance à donner des classes sphériques de même volume, ce qui limite son champ d'application. Certains auteurs se sont attachés à dépasser cette limitation en supprimant la contrainte

pour la métrique M d'être fixée tout au long de l'algorithme tout en restant dans le cadre euclidien de façon à obtenir une dissemblance qui s'adapte aux données traitées. Ces auteurs ont été obligés de fixer des contraintes sur la métrique M. La solution proposée est l'utilisation de MND avec la distance adaptative d_M unique telle que $|M|=1$. En fait, cette contrainte est assez naturelle, néanmoins on pourrait en envisager d'autres. Dans le cas où les classes ont le même type de dispersion mais possédant des directions d'allongement inconnues, ils utilisent à chaque itération la métrique $M=|W|^{1/P} W^{-1}$, où

$$W=\sum_{k=1}^{K} \sum_{x_i \in P_k} (x_i-g_k)(x_i-g_k)^t$$

et g_k est le centre de gravité de la classe P_k. Par ailleurs, le critère de l'algorithme des distances adaptatives (Govaert [9]) élaboré dans un cadre géométrique pour permettre de reconnaître des classes de « formes » différentes, associe à chaque classe P_k et pour chaque itération la métrique $|W_k|^{1/P} W_k$, où

$$W_k=\sum_{x_i \in P_k} (x_i-g_k)(x_i-g_k)^t.$$

Cas binaire

Nous avons cherché à ce que le choix des pondérations des variables dépende des classes. Pour cela on adopte une démarche sensiblement identique à celle des distances adaptatives (Govaert [9]). Comme le signale Lerman [14] « nous avons déjà signalé que la prise en compte *a priori* d'une pondération des variables est tout à fait problématique et discutable, même si cette pondération est basée sur une méthode objective (...). Néanmoins, il faut laisser la porte ouverte aux possibilités expérimentales de l'analyse classificatoire des données ». Gnanadesiken *et al.* [8] arrivaient à une conclusion similaire dans leur excellente étude sur les problèmes et les perspectives de la recherche en discrimination et classification.

On se place dans le même cadre qu'au paragraphe 1 où $\mathbf{N}(\Omega)=\{(y_i, \alpha_i)/i\in\Omega\}$ est le nuage de points de $\{0, 1\}^p$ et α_i est un élément de $\mathbf{R}^{p+}$ représentant un vecteur de pondération associé au point y_i. On propose de résoudre le problème général suivant :

Trouver une partition $P=(P_1, \ldots, P_K)$, *de noyaux associés* $L=(a_1, \ldots, a_K)$ *et un système de poids* $(\alpha_k^j/j=1, \ldots, p;\ k=1, \ldots, K)$ *minimisant les critères*

(i) *ou* (ii) *définis ci-dessous, suivant qu'ils dépendent des classes ou non.*

(*i) $\sum_k \sum_{i \in P_k} \Sigma_j \alpha^j | y_i^j - \alpha_k^j |$, *avec une contrainte sur les* α^j

(*ii) $\sum_k \sum_{i \in P_k} \Sigma_j \alpha_k^j | y_i^j - a_k^j |$, *avec une contrainte sur les* α_k^j *pour tout* k.

Pour éviter la solution dégénérée qui consiste à prendre les α^j tous nuls, il faut imposer une contrainte de cohérence. Le problème du choix des contraintes est évidemment délicat. On examinera dans la suite trois différentes contraintes pour chacun des deux cas ci-dessus.

(*a*) *Contrainte multiplicative*

Cette contrainte est analogue à la contrainte de déterminant constant dans le cas des distances adaptatives. En effet, les pondérations définissent une matrice diagonale dont le déterminant n'est autre que le produit des poids. Pour le critère (i), on impose la contrainte $\prod_{j=1}^{p} \alpha^j = 1$. Ainsi, la fonction d'affectation (recherche des classes) devient : on affectera les points aux « centres » les plus proches au sens d'une distance $L1$ pondérée par les valeurs α^j. Pour la fonction de représentation (recherche des a_k^j et des α^j), quelles que soient les valeurs α^j, il est facile de montrer que les a_k^j sont nécessairement les valeurs majoritaires de chaque classe pour chaque variable. Il ne reste plus qu'à déterminer les α^j. On montre (*cf.* Mkhadri [16]), en utilisant les multiplicateurs de Lagrange, que la solution du problème est définie par :

$$\alpha^q = \left\{ \left(\prod_{j=1}^{p} e_j \right)^{1/p} \right\} \Big/ e_q, \qquad q = 1, \ldots, p,$$

où e_q est le nombre de fois où la valeur majoritaire n'a pas été prise pour la variable q dans une classe, qu'on suppose non nul pour tout q. On notera par MNBPR1 l'algorithme associé à cette dernière solution.

De même, on montre que la solution du problème pour le critère (ii) est :

$$\alpha_k^q = \left\{ \left(\prod_{j=1}^{p} e_k^j \right)^{1/p} \right\} \Big/ e_k^q$$

où e_k^q est le nombre de fois où la valeur majoritaire n'a pas été prise pour la variable q dans la classe k ($q = 1, \ldots, p$; $k = 1, \ldots, K$). Ce système de poids

favorise les variables déséquilibrées pour chaque classe k. MNBPR2 désigne l'algorithme associé à cette solution.

(*b*) *Contrainte additive*

Ce type de contrainte est analogue à la contrainte $\operatorname{tr} M = \text{Cte}$ du cas des distances adaptatives. Remarquons tout d'abord que si l'on impose $\Sigma_q \alpha^q = \text{Cte}$, il est facile de voir que notre problème n'aura pas de solution. Ainsi nous travaillons avec la contrainte $\Sigma_q (\alpha^q)^m = \text{Cte}$, avec $|m| > 1$ qui correspond au cas de non dégénérescence (Mkhadri [16]). Les solutions du problème pour les critères (i) et (ii) sont définies respectivement par :

$$\alpha^q = e_q^{1/(m-1)} / \{ \Sigma_j e_j^{m/(m-1)} \}^{1/m},$$

et

$$\alpha_k^q = (e_k^q)^{1/(m-1)} / \{ \Sigma_j e_k^{jm/(m-1)} \}^{1/m}, \qquad q = 1, \ldots, p; \quad k = 1, \ldots, K.$$

On appellera MNBSO1 et MNBSO2 les algorithmes respectifs associés à ces deux solutions.

Remarque 3.1 : Ce système de poids $(\alpha_k^q;\ q = 1, \ldots, p;\ k = 1, \ldots, K)$ favorise les variables équilibrées quand $m > 1$ contrairement au précédent, tandis qu'il favorise les variables déséquilibrées pour le cas où $m < 1$.

(*c*) *Contrainte logistique*

C'est une contrainte qui vise à favoriser, dans chaque classe k, les variables déséquilibrées, où l'on suppose que $\Sigma_j 1/(1 + \alpha^j) = 1$ pour le critère i [respectivement $\Sigma_j 1/(1 + \alpha_k^j) = 1$ pour le critère ii, pour tout $k = 1, \ldots, K$]. Les solutions fournies par ces contraintes sont données par :

$$\alpha^q = \{ \Sigma_j (e_j)^{1/2} / (e_q)^{1/2} \} - 1,$$

et

$$\alpha_k^q = \{ \Sigma_j (e_k^j)^{1/2} / (e_k^q)^{1/2} \} - 1; \qquad q = 1, \ldots, p; \quad k = 1, \ldots, K.$$

On note par MNBLO1 et MNBLO2 les algorithmes associés à ces deux pondérations.

Remarque 3.2 : Notre étude à partir de l'indice de similarité de Lerman nous a amené à poser le problème des pondérations adaptatives. Trois des contraintes examinées paraissent intéressantes à exploiter du fait qu'elles favorisent les variables déséquilibrées; il reste à évaluer la performance de chacune sur des données réelles.

Nous avons adopté une approche géométrique pour définir des pondérations. Une autre approche, probabiliste, est possible. Elle a été notamment étudiée par Celeux [3] dans un cadre général et par Govaert [10] pour les données binaires. Ce dernier a établi que l'algorithme MNDBIN était lié à un modèle de mélange de distributions de Bernoulli. L'extension possible est de comparer ces deux approches.

4. APPROCHE PROBABILISTE

Récemment, plusieurs auteurs se sont intéressés aux liens existant entre les méthodes de classification automatique et les modèles de la statistique inférentielle. Dans le cas quantitatif, le critère d'inertie intraclasse est associé à un mélange gaussien (Scott et Symons [18], Schroeder [17], Celeux [3]). Pour le cas qualitatif, Celeux [3] a montré que la recherche de la partition en K classes d'information maximale est équivalente à l'identification d'un modèle de K classes latentes sous l'approche classification. Dans le même cadre, Bock [2] montre que les critères classiques d'information s'interprètent comme des vraisemblances classifiantes de modèles log-linéaires.

Govaert [10] se propose de faire la même chose que ces derniers, lorsque les données sont binaires. Il a considéré un mélange (noté $M1$) de lois de Bernoulli multivariées de densité :

$$f(x) = \sum_{k=1}^{K} p_k f(x, a_k, \varepsilon)$$

où

$$f(x, a_k, \varepsilon) = \prod_{k=1}^{K} \varepsilon^{|x^j - a_k^j|} (1-\varepsilon)^{1-|x^j - a_k^j|}$$

avec $a_k = (a_1, \ldots, a_K)$, $0 \leqq p_k \leqq 1$ avec $\Sigma_k p_k = 1$ pour $k = 1, \ldots, K$ et $\varepsilon \in]0, 1/2]$.

Cette paramétrisation particulière des distributions de Bernoulli signifie que les données de loi $f(x, a_k, \varepsilon)$ sont égales à leur valeur « centrale » a_k avec la probabilité $(1-\varepsilon)$ et différentes avec la probabilité ε. La valeur de ε choisie sera estimée sur la base de maximisation de la vraisemblance classifiante. On a, pour $k = 1, \ldots, K$:

$$f(x, a_k, \varepsilon) = \varepsilon^{d(x, a_k)} (1-\varepsilon)^{p - d(x, a_k)}$$

où d est la distance $L1$ définie en paragraphe 1.

La vraisemblance classifiante associé au mélange considéré est :

$$VC(p, a, \varepsilon) = \sum_{k=1}^{K} \sum_{x \in P_k} \operatorname{Log} f(x, a_k, \varepsilon)$$

$$VC(p, a, \varepsilon) = \sum_{k=1}^{K} \sum_{x \in P_k} \operatorname{Log} \left\{ \varepsilon^{d(x, a_k)} (1-\varepsilon)^{p - d(x, a_k)} \right\}$$

que l'on peut écrire :

$$VC(p, a, \varepsilon) = \sum_{k} \sum_{x \in P_k} d(x, a_k) \operatorname{Log} \frac{\varepsilon}{1-\varepsilon} + np \operatorname{Log}(1-\varepsilon).$$

Donc, pour ε fixé dans $]0, 1/2[$, la maximisation de $VC(p, a, \varepsilon)$ en fonction de a est équivalente à la minimisation de $\mathbf{W}(P, a)$ (du fait que $\operatorname{Log}(\varepsilon/1-\varepsilon) < 0$). Enfin, une fois que l'optimisation de $\mathbf{W}$ réalisée, il est facile de voir que $\varepsilon = \mathbf{W}(P^*, a^*)/np$ maximise $VC(p^*, a^*, \varepsilon)$. [Ici P^* et a^* représentent les partitions et les centres qui minimisent $\mathbf{W}(P, a)$.]

On note par $\mathbf{C1}(P, L, \varepsilon)$ le critère $\mathbf{W}(P, a)$ à minimiser.

Le paramètre ε qui mesure l'écart d'une classe à son centre ne dépend ni des variables ni des classes ce qui, dans certaines situations, peut s'avérer irréaliste. Ainsi, Govaert [10] propose d'autres critères où ε pourra dépendre des variables et des classes. Il considère deux mélanges de Bernoulli (notés respectivement $M2$ et $M3$). La vraisemblance classifiante associé au premier ($M2$), où ε dépend des variables, est :

$$VC(P, L, \varepsilon) = \sum_{k} \sum_{x \in P_k} \left\{ \sum_{j=1}^{p} \operatorname{Log} \frac{\varepsilon_j}{1-\varepsilon^j} \left| x^j - a_k^j \right| \right\} + n \sum_{j=1}^{p} \operatorname{Log}(1-\varepsilon^j).$$

Maximiser ce critère revient à minimiser :

$$\mathbf{C2}(P, L, \varepsilon) = \sum_{k} \sum_{x \in P_k} d_\varepsilon(x, a_k) - A_\varepsilon$$

où d_ε est une distance $L1$ pondérée par $\operatorname{Log}(1-\varepsilon^j/\varepsilon^j)$ et $A_\varepsilon = n \Sigma_j \operatorname{Log}(1-\varepsilon^j)$.

Lorsqu'il fixe la partition et les noyaux, il obtient, à la convergence $\varepsilon^j = \mathbf{W}(P^*, a^*)/n$ pour tout j ($j = 1, \ldots, p$).

Pour le deuxième mélange ($M3$), les valeurs ε dépendent à la fois des classes et des variables, le critère qu'on minimise se met sous la forme :

$$\mathbf{C3}(P, L, \varepsilon) = \sum_k \sum_{x \in P_k} \{ d_\varepsilon(x, a_k) - A_\varepsilon \}$$

où d_ε est une distance de type $L1$ pondérée par $\text{Log}\{(1-\varepsilon_k^j)/\varepsilon_k^j\}$ et où A_ε est la quantité $\Sigma_j \text{Log}(1-\varepsilon_k^j)$. Lorsqu'il fixe la partition et les noyaux, il obtient à la convergence $\varepsilon_k^j = \mathbf{W}(P_k^*, a^*)/n_k$ pour tout j ($j=1, \ldots, p$) et $k=1, \ldots, K$. On montre facilement que le système de pondération ($\text{Log}\{(1-\varepsilon_k^j)/\varepsilon_k^j\}$, $j=1, \ldots, p$; $k=1, \ldots, K$) favorise, pour chaque classe, les variables déséquilibrées comme pour les poids associés aux distances adaptatives du paragraphe 3.

Remarques 4 : Nous venons de voir que l'identification d'un mélange de lois de Bernoulli avec le même paramètre pour toutes les classes et toutes les variables correspond au critère de classification optimisé par l'algorithme MNDBIN. La généralisation de ce modèle en considérant différents paramètres permet de proposer de nouveaux critères utilisant des distances adaptatives de type $L1$.

Cette approche permet de justifier, *a posteriori*, l'utilisation, pour les données binaires, d'une part de la distance $L1$, d'autre part de noyaux binaires.

Les expressions des critères de l'approche « géométrique » du paragraphe 3 et de l'approche « modèle » sont difficiles à comparer algébriquement. Seules, les applications pratiques des deux approches sur des données réelles permettront de situer les deux types de critère l'un par rapport à l'autre.

Govaert [11] complète cette étude en partant d'une forme générale de critère métrique (distance $L1$ sur données binaires). Il montre alors comment cette approche rejoint l'approche précédente fondée sur des critères probabilistes définis à partir de distributions de lois de Bernoulli.

5. APPLICATIONS ET COMPARAISONS DES MÉTHODES

Dans cette section, nous considérons deux échantillons de données pour tester la performance des deux approches modèle et géométrique. Le but est d'étudier l'efficacité de toutes ces méthodes et de bien cerner leurs champs d'application.

5.1. Démarche

Pour réduire l'influence de la partition d'initialisation des algorithmes sur les résultats, nous avons opté pour la démarche suivante.

Variante 1 :

– Soit $\mathscr{P}$ un ensemble de 20 partitions aléatoires. Chaque partition de $\mathscr{P}$ est utilisée pour initialiser la méthode fondée sur le critère **C1**. Parmi les 20 essais effectués, on ne retient que le meilleur et on note P^* la meilleure partition obtenue.

– On opère de la même manière pour MNDBVP et on note Q^* la meilleure partition obtenue.

– On applique ensuite les méthodes fondées sur les critères **C2** et **C3** en les initialisant avec la partition P^* fournie par **C1**.

– On applique également MNBPR1, MNBLO1, MNBSO1, MNBPR2, MNBLO2, MNBSO2 en les initialisant avec la partition Q^* fournie par MNDBVP.

Variante 2 :

– Chaque partition de $\mathscr{P}$ est utilisée pour initialiser la méthode fondée sur le critère **C2**, Parmi les 20 essais effectués, on ne retient que le meilleur et on note P^{**} la meilleure partition obtenue.

– On opère de la même manière pour MNBPR1 (resp. MNBLO1 et MNBSO1) et on note Q_1^{**} (resp. Q_2^{**} et Q_3^{**}) la meilleure partition obtenue par MNBPR1 (resp. MNBLO1 et MNBSO1).

– On applique ensuite **C3** en l'initialisant avec la partition P^{**} fournie par **C2**.

– On applique également MNBPR2 (resp. MNBLO2 et MNBSO2) en l'initialisant avec la partition Q_1^{**} (resp. Q_2^{**} et Q_3^{**}) fournie par MNDPR1 (resp. MNBLO2 et MNBSO2).

Variante 3 :

– Chaque partition de $\mathscr{P}$ est utilisée pour initialiser la méthode fondée sur le critère **C3**. Parmi les 20 essais effectués, on ne retient que le meilleur et on note P^{***} la meilleure partition obtenue.

– On opère de la même manière pour MNBPR2 (resp. MNBLO2 et MNBSO2) et on note Q_1^{***} (resp. Q_2^{***} et Q_3^{***}) la meilleure partition obtenue par MNBPR2 (resp. MNBLO2 et MNBSO2).

TABLEAU I.

Les paramètres a_k^k des échantillons $M1$ et $M2$.

Composant	Variable									
	1	2	3	4	5	6	7	8	9	10
1	1	1	1	0	1	1	0	1	1	1
2	0	0	1	1	0	0	0	0	0	0
3	1	0	1	0	1	0	1	1	0	0

TABLEAU II

Les paramètres ε^j de l'échantillon $M2$.

Variable									
1	2	3	4	5	6	7	8	9	10
.01	.20	.01	.48	.01	.30	.40	.20	.01	.20

TABLEAU III

Les paramètres a_k^j des échantillons $M3$.

Composant	Variable									
	1	2	3	4	5	6	7	8	9	10
1	1	0	1	0	1	0	1	1	1	0
2	0	0	0	0	0	1	1	1	1	1
3	0	1	0	0	0	1	0	0	0	0

TABLEAU IV

Les paramètres ε_k^j de l'échantillon $M3$.

Composant	Variable									
	1	2	3	4	5	6	7	8	9	10
1	.00	.10	.20	.01	.20	.01	.50	.30	.01	.20
2	.10	.10	.01	.30	.10	.10	.10	.50	.35	.20
3	.50	.01	.01	.30	.01	.20	.50	.01	.01	.01

5.2. Données simulées

Il s'agit de trois fichiers de données simulées respectivement à partir des modèles $M1$, $M2$ et $M3$. Chaque fichier est constitué de 100 individus caractérisés par 10 variables binaires. Ces jeux de données nous ont été fournis par Govaert pour tester notre approche par rapport à la sienne (où les paramètres des modèles sont définis respectivement par ε, ε^j et ε_k^j, $j=1, \ldots, p$ et $k=1, \ldots, K$). Le nombre de classes demandé est fixé à trois. Ces données sont issues d'un mélange de lois de Bernoulli à trois composants de proportions égales (en fait, les tailles des vraies classes sous-jacentes $P1$, $P2$ et $P3$

TABLEAU V

Tableaux de confusion des algorithmes des deux approches modèle (à gauche) et géométrique (à droite) sur données simulées M1, M2 et M3.

(a) Résultats obtenus par la variante 1							
	*M*1	*M*2	*M*3		*M*1	*M*2	*M*3
C1	2	10	8	mndbvp	2	10	8
				mnbpr1	3	3	12
C2	2	3	6	mnblol.	2	3	11
				mnbsol	3	17	8
C3	2	3	1	mnbpr2	2	5	3
				mnblo2	2	5	3
(b) Résultats obtenus par la variante 2							
	*M*1	*M*2	*M*3		*M*1	*M*2	*M*3
				mnbpr1	3	3	7
C2	2	3	6	mnblo1	2	2	8
				mnbso1	3	10	8
C3	2	3	4	mnbpr2	2	4	3
				mnblo2	2	5	8
(c) Résultats obtenus par la variante 3							
	*M*1	*M*2	*M*3		*M*1	*M*2	*M*3
C3	2	4	0	mnbpr2	2	3	10
				mnblo2	2	3	11

N. B. : la case (i, j) de chaque tableau contient le nombre d'individus qui ont été mal classés par l'algorithme i (**C2**, **C3**, . . .) sur le fichier j ($M1$, $M2$, $M3$).

sont respectivement 34, 33 et 33). Les valeurs des paramètres a_k^j pour le modèle $M1$ sont données dans le tableau I; le paramètre ε vaut 0,1. Les valeurs des paramètres a_k^j du modèle $M2$ sont les mêmes que ceux du modèle $M1$. Le tableau II donne les valeurs des ε^j pour le modèle $M2$. Enfin, les tableaux III et IV fournissent les valeurs des paramètres a_k^j et ε_k^j pour le modèle $M3$.

Les résultats des différents algorithmes sont regroupés dans les tableaux de confusion du tableau V où la case (i, j) contient le nombre d'individus qui ont été mal classés par l'algorithme i (**C1**, **C2**, **C3**, . . .) sur le fichier j ($j = M1$, $M2$, $M3$) par rapport à la partition idéale. La performance de l'algorithme MNBSO2 est relativement moins bonne et ses résultats ne sont pas mentionnés.

On constate que les deux approches donnent des résultats relativement identiques pour la variante 1, sauf pour l'algorithme MNBSO1 sur les données $M2$ et $M3$. Pour la variante 2, on peut faire la même remarque que pour la variante 1, bien que dans ce cas MNBPR2 et MNBLO2 soient moins

TABLEAU VI

Tableaux de confusion et pondérations des variables associées aux algorithmes des approches modèle et géométrique sur le fichier DKS.

(a) Résultats obtenus par la variante 1

	C2	**C3**	mnbpr1	mnbpr2	mnblo1	mnblo2
DKS.	1	7	4	6	3	4

Pondérations associées à **C2**:

2,5	1,9	2,7	1,8	1,3	1,8	0,7	1,3	2,9	1,3

Pondérations associées à **C3**:

2,9	3,6	3,6	2,9	2,5	3,6	1,1	3,6	3,6	2,9
1,2	1,4	3,2	1,2	0,0	1,0	0,2	0,3	3,2	0,2
2,4	1,6	2,4	0,3	0,3	1,1	1,6	0,7	1,1	1,1

Pondérations associées à mnbpr1 :

1,2	1,0	1,4	0,8	0,5	0,9	0,4	0,6	9,9	0,7

Pondérations associées à mnbpr2 :

0,7	0,7	1,0	1,0	0,7	2,9	0,2	1,5	2,9	1,0
1,4	0,7	1,4	0,7	0,5	0,9	0,9	0,7	2,7	1,4
1,5	1,5	4,6	0,9	0,4	0,7	0,4	0,5	4,6	0,4

(b) Résultats obtenus par la variante 2

	C3	mnbpr2	mnblo2
DKS.	5	5	14

(c) Résultats obtenus par la variante 3

	C3	mnbpr2
DKS.	20	13

N. B. : la case (i, j) de chaque tableau de confusion contient le nombre d'individus qui ont été mal classés par l'algorithme j (**C2**, **C3**, . . .) sur le fichier i (*DKS*).

performants sur $M3$. Enfin, pour la variante 3, MNBPR2 et MNBLO2 présentent les mêmes résultats que **C3** sur les données $M1$ et $M2$.

Donc, bien que les données aient été tirées suivant les modèles $M1$, $M2$ et $M3$, les résultats de l'approche géométrique sont sensiblement identiques à ceux de l'approche probabiliste.

5.3. Données DKS

Nous reprenons la même étude sur des données nommées « Diagnosis of Keratoconjunctivitis Sicca », en abrégé DKS, traitées par Anderson *et al.* [1]. Il s'agit de 77 individus caractérisés par 10 variables binaires. Les pondérations des différents algorithmes sont indiquées sur le tableau VI. Les résultats des différents algorithmes sont regroupés dans les tableaux de confusion du tableau VI où la case (i, j) contient le nombre d'individus qui ont été mal classés par l'algorithme j (**C2**, **C3**, . . .) sur le fichier DKS.

On remarque pour le critère **C2** qu'il y a un seul élément qui a été mal classé par rapport au modèle de base, alors que pour les algorithmes MNBPR1 et

MNBLO1, 4 et respectivement 3 éléments ont été mal classés. Par contre, le critère **C3** et les algorithmes MNBPR2 et MNBLO2 fournissent des résultats très proches. Pour les pondérations associées à **C2** et MNBPR1, on constate que leurs valeurs sont, bien sûr, différentes, mais elles ont la même influence sur toutes les variables (sauf sur la variable 9 pour l'algorithme MNBPR1 qui a exagéré sa valeur). De même, l'effet de pondération des variables du critère **C3** et l'algorithme MNBPR2 est à peu près similaire.

Lorsque les algorithmes (**C2**, MNBPR2 et resp. MNBLO2) de chaque approche sont initialisées par la meilleure partition obtenue par (**C1**, MNBPR1 et resp. MNBLO1), les résultats sont identiques, sauf pour l'algorithme MNBLO2. Dans le dernier cas, où **C3** et MNBPR2 sont initialisés par une partition tirée au hasard parmi 20 tirages aléatoires, 20 éléments par le critère **C3** de l'approche modèle, ont été mal classés alors que pour l'approche géométrique 13 éléments seulement ont été mal classés.

CONCLUSION

A notre avis les algorithmes MNDBIN et MNDBVP se situent au même niveau (puisque la pondération reste inchangée au cours de l'algorithme) et les algorithmes adaptatifs doivent être vus comme leurs compléments. Ainsi, en présence d'un tableau binaire quelconque qu'on veut partitionner, on appliquera d'abord l'un des algorithmes non adaptatifs. Partant de la partition obtenue, on peut ensuite regarder si l'un des algorithmes adaptatifs la modifie. Si c'est le cas, il est probable que l'algorithme adaptatif ait ainsi mis en évidence une structure difficile à découvrir. Maintenant, le choix de la contrainte est évidemment important. Si on veut donner de l'importance aux variables déséquilibrées, on utilisera la contrainte multiplicative (ou logistique) ou le critère de vraisemblance classifiante **C2** ou **C3**. Par contre si on veut atténuer le rôle des variables déséquilibrées, on choisira la contrainte additive.

BIBLIOGRAPHIE

1. J. A. Anderson, A Statistical Aid to the Diagnosis of Keratoconjunctivitis Sicca, *Quart. J. Med.*, 1972, *41*, p. 175-189.
2. H.-H. Bock, Loglinear Models and Entropy Clustering Methods for Qualitative Data, Classification as Tool of Research, W. Gaul and M. Schader éd., *Elsevier Science Publishers B. V.*, North-Holland, 1986.
3. G. Celeux, Classification et modèles, *Rev. Statist. Appl.*, 1988, *36*, n° 4, p. 43-58.

4. E. DIDAY *et al.*, Optimisation et classification automatique, I.N.R.I.A., Rocquencourt, 1980.
5. B. FICHET, The Role Played by L_1 in Data Analysis. Statistical Data Analysis on the L_1-Norm and Related Methods, Y. DODGE éd., *Elsevier Science Publishers B.V.*, North-Holland, 1987.
6. B. FICHET et G. LE CALVE, Structure géométrique des principaux indices de dissimilarité sur signes de présence-absence, *Statist. Anal. Données*, 1984, *9*, n° 3, p. 11-44.
7. H. FRIEDMAN et J. RUBIN, On Some Invariant Criterion for Grouping Data, *JASA*, 1967, *62*, p. 159-178.
8. R. GNANADESIKAN *et al.*, Discriminant analysis and clustering, *Statist. Sc.*, 1989, *4*, n° 1, p. 34-69.
9. G. GOVAERT, Classification croisée, *Thèse de doctorat d'État,* Pierre-et-Marie-Curie, Paris-VI, 1983.
10. G. GOVAERT, Classification binaire et modèles, *Rev. Statist. Appl.*, 1990, *38*, N1.
11. G. GOVAERT, Modèle de classification et distance dans le cas discret, Rapport de Recherches I.N.R.I.A., 1990 (à paraître).
12. S. JOLY et G. LE CALVE, Étude des puissances d'une distance, *Statist. Anal. Données,* 1986, *11,* n° 3, p. 30-50.
13. G. LE CALVE, L_1-Embedding of a Data Structure (*I, D*). Statistical Data Analysis on the L_1-Norm and Related Methods, Y. DODGE éd., *Elsevier Science Publishers B.V.*, North Holland, 1987.
14. I. C. LERMAN, Construction d'un indice de similarité entre objets décrits par des variables d'un type quelconque. Application au problème du consensus en classification. *Rev. Statist. Appl.*, 1987, *35*, N2.
15. F. MARCHETTI, Contribution à la classification de données binaires et qualitatives, *Thèse,* Université de Metz, 1989.
16. A. MKHADRI, Pondération des variables pour la classification binaire, Rapport de Recherches I.N.R.I.A., n° 1079, 1989.
17. A. SCHROEDER, Analyse d'un mélange de distribution de probabilité de même type, *Rev. Statist. Appl.*, 1976, *24*, N1.
18. A. SCOTT et M. SYMONS, Clustering methods based on likelihood ratio criteria, *Biometrics, 27*, 1971.