

REVUE DE STATISTIQUE APPLIQUÉE

M. NADIF

F. MARCHETTI

Classification de données qualitatives et modèles

Revue de statistique appliquée, tome 41, n° 1 (1993), p. 55-69

http://www.numdam.org/item?id=RSA_1993__41_1_55_0

© Société française de statistique, 1993, tous droits réservés.

L'accès aux archives de la revue « Revue de statistique appliquée » (<http://www.sfds.asso.fr/publicat/rsa.htm>) implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques

<http://www.numdam.org/>

CLASSIFICATION DE DONNÉES QUALITATIVES ET MODÈLES

M. Nadif, F. Marchetti

*Université de Metz, Laboratoire de Recherche en Informatique
Ile du Saulcy 57045 Metz*

RÉSUMÉ

Dans cet article, nous proposons des méthodes de classification pour données qualitatives. Nous montrons que les critères métriques utilisés par ces méthodes sont associés à des modèles probabilistes. Dans notre approche, les données initiales ne sont pas dénaturées, et chaque classe est caractérisée par un vecteur de modalités. Nous montrons que cette étude est une généralisation des travaux déjà réalisés dans le cas des données binaires. Enfin, nous présentons des applications sur des données simulées et comparons ces méthodes avec la méthode de classification utilisant la distance du Chi2.

Mots-clés : *Classification, Données qualitatives, Noyaux de modalités, Mélange de lois de probabilités.*

ABSTRACT

In this paper, we propose clustering methods for discrete data. We show that the criteria used by these methods are associated with probabilistic models. In our approach, the data are not denatured and each cluster is characterized by a vector of categories. We show that this study is a generalization of works realized with binary data. Finally, we present some numeric experiences on simulated data and we compare these methods with the cluster method using the Chi2 distance.

Key-words : *Clustering, Discrete data, Kernels of categories, Mixture.*

1. Introduction

Les modèles de statistique inférentielle apportent souvent un éclairage nouveau des méthodes de classification automatique reposant sur l'optimisation d'un critère défini à partir d'une distance. Ainsi, lorsque les données sont quantitatives, le critère de l'inertie correspond à l'hypothèse d'une population issue d'un mélange de lois gaussiennes (Celeux 1988). Lorsque les données sont binaires, le critère défini à partir de la distance L1 (Marchetti 1989) rejoint l'hypothèse d'un mélange de lois de Bernoulli comme le montre Govaert (1990). Par contre, pour

des données qualitatives, l'approche probabiliste n'apparaît pas aussi clairement. En effet, la méthode optimisant le critère du Chi2, nommée MNDQAL par Ralambondrainy (1988) n'est associée à aucun modèle. Cependant, Celeux (1988) constate par expérience que, sous l'approche classification, le critère d'information mutuelle (Benzécri 1973) et le critère du Chi2 pour des données qualitatives sont associés au modèle des classes latentes (Goodman 1974, Everitt 1981). En outre, la méthode MNDQAL, utilisant la distance du Chi2, dénature les données initiales en les transformant en vecteurs de $\mathbb{R}^p$. La conséquence immédiate est la difficile interprétation du critère final en regard des données initiales.

Nous proposons ici une méthode de classification automatique (avec plusieurs variantes) pour données qualitatives nominales qui ne possède pas les inconvénients mentionnés ci-dessus. Dans le deuxième paragraphe, nous présentons cette méthode dans sa version la plus simple. Elle est nommée MNDDIJ (Marchetti 1989) et est caractérisée par la définition d'un espace de vecteurs de modalités muni d'une distance particulière qui est égale au nombre de différences entre les deux éléments considérés. Cette approche permet d'une part le respect de la nature initiale des données, d'autre part une interprétation simple des résultats. Dans le troisième paragraphe, nous rappelons comment l'approche classification peut être utilisée pour identifier un mélange de lois de probabilité (Scott et Symons 1971, Schroeder 1976), puis nous proposons un modèle de mélange de lois reposant sur la donnée d'une modalité et d'une probabilité d'erreur. Nous montrons alors que la méthode de classification MNDDIJ correspond au cas le plus simple du modèle : les variables possèdent toutes le même nombre de modalités et la probabilité d'erreur est la même pour toutes les variables. La méthode apparaît alors restrictive, mais comme nous le montrons dans les paragraphes 3 et 4, il est néanmoins possible de construire des variantes en différenciant les lois des variables (une probabilité d'erreur propre à chaque variable) ou en différenciant les lois des variables dans chacune des classes de la partition. Cette dernière approche correspond alors au cas le plus général du modèle. Nous terminons notre étude par une expérimentation des méthodes proposées sur des données simulées à partir des modèles probabilistes associés. Les résultats fournis sont alors très satisfaisants comparativement à la méthode MNDQAL qui est souvent mise à défaut pour ces mêmes données simulées.

2. Méthode de classification

2.1 Notations

Soit $X(I, Q)$ le tableau de modalités croisant un ensemble $I = \{1, 2, \dots, n\}$ de n individus et un ensemble $Q = \{1, 2, \dots, p\}$ de p variables qualitatives nominales. Nous notons :

$$X(I, Q) = [x_i^q]$$

où x_i^q représente la modalité de la variable q choisie par l'individu i .

A chaque variable q correspond l'ensemble de modalités $J_q = \{1, 2, \dots, m_q\}$, m_q étant le nombre de modalités de la variable q . Nous définissons alors l'espace E comme le produit $J_1 \times J_2 \times \dots \times J_p$, que nous munissons de la distance d ,

égale au nombre de composantes différentes entre les deux éléments considérés. Son expression sur E est la suivante :

$$\forall (x, y) \in E^2 \quad d(x, y) = \sum_{q \in Q} \delta(x^q, y^q)$$

$$\text{où } \delta(x^s, y^q) = \begin{cases} 1 & \text{si } x^q \neq y^q \\ 0 & \text{sinon} \end{cases}$$

A partir des lignes du tableau $X(I, Q)$, nous définissons le nuage $N(I)$, inclus dans l'espace E , par :

$$N(I) = \{x_i, i \in I\}$$

$$\text{où } x_i = (x_1^i, x_1^2, \dots, x_i^p)$$

2.2 Problème

Il s'agit de déterminer une partition de l'ensemble I des individus en K classes, K étant fixé *a priori*. Nous allons écrire un algorithme basé sur le principe des Nuées Dynamiques mais respectant le principe d'homogénéité suivant : l'ensemble à classer étant inclus dans E , nous imposons aux noyaux d'appartenir à ce même espace E . Le problème à résoudre est alors le suivant :

Trouver une partition $P = (P_1, P_2, \dots, P_K)$ de I et un ensemble $L = (a_1, a_2, \dots, a_K)$ de K noyaux de E tels que le critère :

$$W(P, L) = \sum_{k=1}^K \sum_{i \in P_k} d(x_i, a_k) = \sum_{k=1}^K \sum_{i \in P_k} \sum_{q \in Q} \delta(x_i^q, a_k^q)$$

soit minimum.

La méthode nommée MNDDIJ (Marchetti 1989) fournit une solution à ce problème. Notons que si les variables ont toutes 2 modalités (cas binaire), nous retrouvons le critère défini à partir de la distance L_1 .

2.3 Algorithme

Il s'agit de construire la fonction d'affectation f telle que $W(f(L), L)$ soit minimum, et la fonction de représentation g telle que $W(P, g(P))$ soit minimum.

• Fonction d'affectation

Chaque individu i est affecté à la classe P_k dont il est le plus proche au sens de la distance d .

• *Fonction de représentation*

Il s'agit de déterminer l'ensemble L des K noyaux optimisant le critère $W(P, g(P))$. Pour cela, pour toute classe P_k , il suffit de rechercher le noyau a_k appartenant à E et minimisant la quantité :

$$\sum_{i \in P_k} d(x_i, a_k)$$

La solution est de choisir pour composante q de a_k , la modalité de la variable q la plus souvent choisie dans la classe P_k :

$$a_k^q = \text{modalité majoritaire relative de } \{x_i^q, i \in P_k\}$$

L'algorithme ainsi construit fournit un ensemble de noyaux de même nature que les éléments à classer. Par définition, ces noyaux sont de plus facilement interprétables. Il nous reste à préciser la signification du critère obtenu à la convergence de l'algorithme.

2.4 Expression du critère à la convergence

À la convergence, nous pouvons exprimer le critère uniquement par rapport à la partition de sorte que :

$$W(P) = \sum_{k=1}^K \sum_{i \in P_k} \sum_{q \in Q} \delta(x_i^q, a_k^q)$$

Ce critère représente le nombre d'éléments du tableau $X(I, Q)$ qui sont différents des éléments du tableau correspondant à la situation «idéale» (situation où tous les individus sont identiques aux noyaux des classes auxquelles ils appartiennent).

2.5 Indices d'aides à l'interprétation

La valeur du critère à la convergence constitue une première indication car elle mesure la qualité de la partition fournie. Cependant, cette valeur qui est égale au nombre de différences entre situation obtenue et situation «idéale» est à étudier en regard de la taille du tableau initial. À partir de ce critère, nous définissons un premier indice égal au pourcentage de valeurs identiques entre noyaux et données initiales. Ainsi, l'utilisateur peut juger rapidement de la qualité de la partition obtenue. Sa définition est la suivante :

$$100 \times \frac{(np - W(P))}{np} \quad \text{où} \quad W(P) = \sum_{k=1}^K \sum_{q \in Q} D_k^q$$

et où D_k^q est le nombre de modalités différentes de la modalité majoritaire relative dans la classe P_k pour la variable q .

De plus, pour chaque classe de la partition, nous proposons un second indice permettant de juger de sa qualité. Il représente simplement le pourcentage de valeurs identiques au noyau de la classe. Il est alors possible de juger de manière indépendante de l'homogénéité de chacune des classes obtenues. Pour une classe P_k de cardinal n_k , la définition de l'indice est la suivante :

$$100 \times \frac{n_k p - D_k}{n_k p} \quad \text{où} \quad D_k = \sum_{q \in Q} D_k^q$$

Les noyaux de modalités permettent d'étudier rapidement les particularités de chaque classe de la partition. En complément, pour chaque couple (classe, variable), nous proposons un indice égal au pourcentage d'individus ayant choisi la modalité majoritaire relative. L'utilisateur dispose ainsi de $K \times p$ indicateurs directement interprétables. Pour une classe P_k et une variable q , l'expression de l'indice est :

$$100 \times \frac{n_k - D_k^q}{n_k}$$

2.6 Exemple d'application

Soit un ensemble de 10 individus identifiés par les nombres de 1 à 10 et décrits par un ensemble de 5 variables qualitatives nominales identifiées par les lettres de a à e. Supposons que ces variables possèdent chacune 3 modalités représentées par les valeurs 1, 2 et 3. Les données initiales sont organisées sous la forme d'un tableau de modalités (tableau 1). Nous appliquons alors la méthode MNDDIJ en demandant 3 classes. Après plusieurs essais, nous aboutissons finalement à la partition $\{A, B, C\} = \{\{3, 7, 9, 10\}, \{1, 4, 5, 8\}, \{2, 6\}\}$. Nous représentons le tableau initial réordonné suivant cette partition (tableau 2). Nous indiquons également les indices d'aides à l'interprétation que sont :

- les effectifs et les indices d'homogénéité des classes (tableau 3);
- les noyaux de modalités fournis par la méthode (tableau 4);
- la mesure d'homogénéité de chaque couple (classe, variable) (tableau 5).

La valeur du critère à la convergence étant de 12, cela indique que sur 50 valeurs initiales, 12 ne sont pas égales à la valeur idéale. La situation «idéale» obtenue permet donc d'expliquer 76% des données initiales.

TABLEAU 1
tableau initial

	a	b	c	d	e
1	1	2	2	3	2
2	3	2	1	1	1
3	2	3	3	1	1
4	1	1	2	3	3
5	1	2	1	3	3
6	3	2	1	1	2
7	3	3	2	1	1
8	1	1	1	3	3
9	2	2	2	1	1
10	2	3	3	2	2

TABLEAU 2
tableau réordonné

	a	b	c	d	e
3	2	3	3	1	1
7	3	3	2	1	1
9	2	2	2	1	1
10	2	3	3	2	2
1	1	2	2	3	2
4	1	1	2	3	3
5	1	2	1	3	3
8	1	1	1	3	3
2	3	2	1	1	1
6	3	2	1	1	2

TABLEAU 3
caractéristiques des classes

	effectifs	homogénéité
A	4	70
B	4	75
C	2	90

TABLEAU 4
les noyaux

	a	b	c	d	e
A	2	3	2	1	1
B	1	1	1	3	3
C	3	2	1	1	1

TABLEAU 5
homogénéité par classe et par variable

	a	b	c	d	e
A	75	75	50	75	75
B	100	50	50	100	75
C	100	100	100	100	50

3. Classification et modèle

De nombreuses méthodes de classification reposent essentiellement sur la définition d'une distance (ou plus généralement sur celle d'une mesure de dissimilarité) et d'un critère associé, sans faire explicitement référence à des modèles probabilistes. Cependant, il est souvent possible de montrer qu'il existe un modèle sous-jacent qui permet alors de donner une interprétation du critère et de justifier de son choix. Dans ce paragraphe, nous rappelons comment l'approche classification peut être utilisée puis nous proposons un modèle associé aux données qualitatives.

3.1 Modèle général et approche classification

Le tableau initial X est considéré comme un échantillon Ω de taille n d'une variable aléatoire à valeurs dans E dont la loi de probabilité admet la fonction de densité f :

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

avec $\forall k = 1, K \quad p_k \in]0, 1[\quad \text{et} \quad \sum_{k=1}^K p_k = 1$

où $f(., a)$ est une distribution de probabilité sur E appartenant à une famille de distributions de probabilités (dépendant du paramètre a élément de $\mathbb{R}^s$, $s \geq 1$) et où p_k est la probabilité qu'un point de l'échantillon suive la loi $f(., a_k)$, c'est-à-dire le poids spécifique du composant dans la population générale. Nous appelons ces p_k les proportions du mélange, K étant le nombre de composants de ce mélange.

Dans l'approche «classification» (Scott et Symons 1971, Schroeder 1976), le problème initial d'estimation est remplacé par le problème suivant :

Rechercher une partition $P = (P_1, \dots, P_K)$, K étant supposé connu, telle que chaque classe P_k soit assimilable à un sous-échantillon qui suit une loi $f(., a_k)$.

L'espace de représentation d'une classe étant l'espace de définition des paramètres a dont dépendent les densités $f(., a)$, la méthode vise alors à maximiser le critère de **vraisemblance classifiante** suivant :

$$W(P, a) = \sum_{k=1}^K \text{Log } L(P_k, a_k) \quad (3.1.1)$$

où a est le p -uplet $(a_1, \dots, a_K)$ et $L(P_k, a_k)$ est la vraisemblance du sous-échantillon P_k suivant la loi $f(., a_k)$:

$$L(P_k, a_k) = \prod_{i \in P_k} f(x_i; a_k).$$

Pour maximiser ce critère (3.1.1), nous pouvons utiliser des algorithmes de type Nuées Dynamiques qui construisent à partir d'une partition initiale P^0 en K classes une suite de partitions en appliquant les deux fonctions suivantes :

- une fonction de représentation g définie par $g(P) = g(P_1, \dots, P_K) = (a_1, \dots, a_K)$ où a_k est l'estimation du maximum de vraisemblance du paramètre de la densité associée au sous-échantillon P_k ;
- une fonction d'affectation définie par $f(a) = f(a_1, \dots, a_K) = (P_1, \dots, P_K)$ où $P_k = \{x_i \in \Omega / f(x_i; a_k) \geq f(x_i; a_m) \text{ avec } k < m \text{ en cas d'égalité}\}$.

Cet algorithme fait croître le critère à chaque itération et nous obtenons à la convergence une partition P et une estimation des paramètres a_k . Les proportions p_k du mélange supposées constantes dans l'algorithme peuvent être alors estimées par les quantités $(\text{Card}(P_k)/n, k = 1, K)$

3.2 Modèle associé aux données qualitatives

Pour chaque composant k du mélange, nous supposons que les p variables sont indépendantes et que chacune d'elles (q) suit une loi de distribution dépendant d'un paramètre a et d'une modalité a_k^q :

$$\forall q \in Q \quad \begin{cases} a_k^q \text{ avec la probabilité } 1 - \alpha \\ \text{les autres modalités avec la probabilité } \frac{\alpha}{m_q - 1} \end{cases}$$

En nous inspirant de la méthode MNDDIJ, la valeur α peut être considérée comme la probabilité d'erreur que la variable q ne prenne pas la modalité a_k^q . Ainsi, nous imposons la contrainte que $\alpha < \frac{m_q - 1}{m_q}$ pour tout $q \in Q$ (en cas d'égalité, à chaque modalité est associée une même probabilité $\frac{1}{m_q}$, ce qui ne correspond plus à la donnée d'une modalité plus probable que les autres). De cette façon, la modalité a_k^q est choisie avec la plus forte probabilité et les autres modalités se partagent de manière équiprobable l'erreur α .

Nous pouvons alors écrire :

$$f(x_i; a_k) = \prod_{q \in Q} (1 - \alpha)^{1 - \delta(x_i^q, a_k^q)} \left(\frac{\alpha}{m_q - 1} \right)^{\delta(x_i^q, a_k^q)} \quad (3.2.1)$$

où $a_k = (a_k^1, \dots, a_k^p)$ et où les a_k^q indiquent la probabilité retenue :

$$\begin{cases} a_k^q = x_i^q & \text{pour la première probabilité } 1 - \alpha \\ a_k^q \neq x_i^q & \text{pour la seconde probabilité } \alpha / (m_q - 1) \end{cases}$$

3.3 Optimisation du critère

A partir de (3.2.1), nous pouvons alors déduire l'expression du critère de vraisemblance classifiante (3.1.1) qui s'écrit sous la forme suivante :

$$W(P, a, \alpha) = \sum_{k=1}^K \sum_{i \in P_k} \left\{ - \sum_{q \in Q} \text{Log} \left(\left(\frac{1 - \alpha}{\alpha} \right) (m_q - 1) \right)^{\delta(x_i^q, a_k^q)} \right\} + np \text{Log}(1 - \alpha)$$

Il s'agit ensuite de rechercher les différents paramètres du modèle maximisant ce critère. Notons d'abord que la recherche de α et celle des a_k sont indépendantes.

De plus, maximiser le critère $W(P, a, \alpha)$ par rapport à a revient à minimiser la quantité :

$$C(P, a) = \sum_{k=1}^K \sum_{i \in P_k} \sum_{q \in Q} \text{Log} \left(\left(\frac{1-\alpha}{\alpha} \right) (m_q - 1) \right) \delta(x_i^q, a_k^q)$$

La minimisation de cette quantité se réalise à l'aide des deux fonctions habituelles caractérisant la méthode des Nuées Dynamiques.

• *Fonction d'affectation (recherche des classes)*

Chaque individu i est affecté à la classe dont il est le plus proche au sens de la distance D définie par :

$$D(x_i, a_k) = \sum_{q \in Q} \text{Log} \left(\left(\frac{1-\alpha}{\alpha} \right) (m_q - 1) \right) \delta(x_i^q, a_k^q)$$

Il s'agit bien d'une distance puisque sous la contrainte imposée sur α (c'est-à-dire $\alpha \in]0, \min \left\{ \frac{m_q - 1}{m_q}; q \in Q \right\} [$), le terme $\text{Log}[(1 - \alpha)(m_q - 1)/\alpha]$ est positif. Notons que cette distance est du même type que celle utilisée dans la méthode MNDDIJ et est pondérée par les coefficients $\text{Log}[(1 - \alpha)(m_q - 1)/\alpha]$.

• *Fonction de représentation (recherche des a_k^q et de α)*

Quelle que soit la valeur du paramètre α , il est évident que chaque a_k^q correspond à la modalité majoritaire relative de la variable q pour la classe P_k .

Par ailleurs, la valeur du paramètre α maximisant $W(P, a, \alpha)$ est donnée par e/n où $e = \sum_{k=1}^K \sum_{i \in P_k} \sum_{q \in Q} \delta(x_i^q, a_k^q)$ exprime le nombre de fois où la valeur majoritaire n'a pas été prise par les individus dans toutes les classes. Il nous reste à vérifier si la valeur α respecte bien la contrainte imposée. Pour cela, si nous notons $e_k^q = \sum_{i \in P_k} \delta(x_i^q, a_k^q)$, nous avons $e_k^q \in [0, \frac{m_q - 1}{m_q} n_k]$ et par conséquent $e \in [0, n \sum_{q \in Q} \frac{m_q - 1}{m_q}]$. La valeur α appartient donc à l'intervalle $[0, \frac{1}{p} \sum_{q \in Q} \frac{m_q - 1}{m_q}]$ qui contient $]0, \min \left\{ \frac{m_q - 1}{m_q}; q \in Q \right\} [$. Nous proposons alors de choisir $\alpha = \min \left\{ \frac{e}{np}, \min \left\{ \frac{m_q - 1}{m_q}; q \in Q \right\} \right\}$.

• *Remarque*

Nous avons ainsi construit une première méthode de classification pour données qualitatives. Celle-ci peut paraître restrictive du fait de l'hypothèse d'une même probabilité d'erreur pour toutes les variables considérées. Cependant, cette approche admet les deux cas particuliers suivants :

i) Dans le cas binaire, pour tout q appartenant à Q , nous avons $m_q = 2$, la contrainte imposée sur α se restreint à α appartenant à $]0, 1/2[$ et la quantité à minimiser devient :

$$C(P, a) = \sum_{k=1}^K \sum_{i \in P_k} \sum_{q \in Q} \delta(x_i^q, a_k^q)$$

Nous rejoignons alors les travaux de Govaert (1990) qui montre comment l'identification d'un mélange de distributions de Bernoulli avec le même paramètre appartenant à $]0, 1/2[$ pour toutes les classes et toutes les variables correspond au critère de classification binaire utilisant la distance L_1 et des noyaux binaires (Marchetti 1989).

ii) Lorsque les m_q sont tous égaux (nous posons $m_q = m$ pour tout q), la quantité à minimiser devient :

$$\begin{aligned} C(P, a) &= \text{Log} \left(\left(\frac{1-\alpha}{\alpha} \right) (m-1) \right) \sum_{k=1}^K \sum_{i \in P_k} \sum_{q \in Q} \delta(x_i^q, a_k^q) \\ &= \text{Log} \left(\left(\frac{1-\alpha}{\alpha} \right) (m-1) \right) W(P, L) \end{aligned}$$

La minimisation de $C(P, a)$ est alors équivalente à la maximisation de $W(P, a)$, ce qui permet d'affirmer que le mélange de distributions proposé, avec la contrainte que tous les m_q soient égaux, correspond au critère de la méthode de classification MNDDIJ.

4. Généralisation

L'algorithme précédent peut être étendu en considérant deux autres modèles dépendant du choix du paramètre α . En effet, celui-ci peut être remplacé soit par $(\alpha^1, \dots, \alpha^p)$, soit par $\{(\alpha_k^1, \dots, \alpha_k^p); k = 1, K\}$ suivant que α dépend respectivement de chaque variable ou de chaque couple (classe, variable). Dans les deux cas, la contrainte imposée est la même : les α^q et α_k^q doivent appartenir à l'intervalle $]0, \frac{m_q-1}{m_q}[$. L'utilisation d'un paramètre dépendant à la fois de la variable et de la classe correspond à la version la plus générale de la méthode.

4.1 Paramètre dépendant de chaque variable

Dans ce cas, l'expression (3.2.1) s'écrit :

$$f(x_i; a_k) = \prod_{q \in Q} (1 - \alpha^q)^{1 - \delta(x_i^q, a_k^q)} \left(\frac{\alpha^q}{m_q - 1} \right)^{\delta(x_i^q, a_k^q)}$$

Le critère de vraisemblance classifiante (3.1.1) devient :

$$W_1(P, a, \alpha) = - \sum_{k=1}^K \sum_{i \in P_k} \sum_{q \in Q} \left\{ \text{Log} \left(\left(\frac{1 - \alpha^q}{\alpha^q} \right) (m_q - 1) \right) \delta(x_i^q, a_k^q) - \text{Log}(1 - \alpha^q) \right\}$$

Pour maximiser ce critère, nous procédons de la même manière que précédemment.

- *Fonction d'affectation (recherche des classes)*

Chaque individu i est affecté à la classe dont il est le plus proche au sens de la mesure de dissimilarité pondérée D_1 définie par :

$$D_1 = (x_i, a_k) = \sum_{q \in Q} \left\{ \text{Log} \left(\left(\frac{1 - \alpha^q}{\alpha^q} \right) (m_q - 1) \right) \delta(x_i^q, a_k^q) - \text{Log}(1 - \alpha^q) \right\}$$

- *Fonction de représentation (recherche des a_k^q et de α)*

Quelles que soient les composantes du paramètre α , il est évident que chaque a_k^q correspond à la modalité majoritaire relative de la variable q pour la classe P_k .

Par ailleurs, chaque composante α^q est donnée par $\frac{e^q}{n}$ où $e^q = \sum_{k=1}^K \sum_{i \in P_k} \delta(x_i^q, a_k^q)$

exprime le nombre de fois où la valeur majoritaire n'a pas été prise dans chacune des classes pour la variable q .

4.2 Paramètre dépendant de chaque variable et de chaque classe

Dans ce cas, l'expression (3.2.1) s'écrit :

$$f(x_i; a_k) = \prod_{q \in Q} (1 - \alpha_k^q)^{1 - \delta(x_i^q, a_k^q)} \left(\frac{\alpha_k^q}{m_q - 1} \right)^{\delta(x_i^q, a_k^q)}$$

et le critère de vraisemblance classifiante devient :

$$W_2(P, a, \alpha) = - \sum_{k=1}^K \sum_{i \in P_k} \sum_{q \in Q} \left\{ \text{Log} \left(\left(\frac{1 - \alpha_k^q}{\alpha_k^q} \right) (m_q - 1) \right) \delta(x_i^q, a_k^q) - \text{Log}(1 - \alpha_k^q) \right\}$$

Pour maximiser ce critère, nous procédons toujours de la même manière.

- *Fonction d'affectation (recherche des classes)*

Chaque individu i est affecté à la classe dont il est le plus proche au sens de la mesure de dissimilarité pondérée D_2 définie par :

$$D_2(x_i, a_k) = \sum_{q \in Q} \left\{ \text{Log} \left(\left(\frac{1 - \alpha_k^q}{\alpha_k^q} \right) (m_q - 1) \right) \delta(x_i^q, a_k^q) - \text{Log}(1 - \alpha_k^q) \right\}$$

• *Fonction de représentation (recherche des a_k^q et de α)*

Quelles que soient les composantes du paramètre α , chaque a_k^q correspond à la modalité majoritaire relative de la variable q pour la classe P_k .

Chaque composante α_k^q est donnée par $\frac{e_k^q}{n_k}$ où n_k représente l'effectif de la classe P_k et $e_k^q = \sum_{i \in P_k} \delta(x_i^q, a_k^q)$ exprime le nombre de fois où la valeur majoritaire n'a pas été prise dans la classe P_k pour la variable q .

Cette méthode correspond au cas le plus général et englobe toutes celles proposées dans ce travail. De plus, dans le cas où toutes les variables ont 2 modalités, nous rejoignons le modèle généralisé associé aux données binaires et proposé par Govaert (1990).

• *Remarque :*

Il est facile de vérifier que les α^q (respectivement α_k^q) maximisant $W_1(P, a, \alpha)$ (respectivement $W_2(P, a, \alpha)$) vérifient bien la contrainte (α^q , respectivement α_k^q appartiennent à $]0, \frac{m_q - 1}{m_q} [$) sauf dans le cas très particulier où $e^q = 0$ (respectivement $e_k^q = 0$). Pour pallier cet inconvénient, il suffit de prendre dans l'algorithme les valeurs de ces paramètres calculées dans l'étape précédente. Dans ce cas, nous ne maximisons pas le critère mais nous l'améliorons.

5. Expériences numériques

Pour illustrer les différentes variantes de notre méthode, nous avons procédé comme suit : nous avons simulé des données structurées en classes ; nous appliquons ensuite les différentes méthodes et nous comparons la partition obtenue avec la partition simulée. En outre, nous comparons les résultats obtenus avec ceux de la méthode MNDQAL.

5.1 Appellation des méthodes

méthodes	Caractéristiques des modèles associés
MNDDIJ	M1 : même nombre de modalités et α constant
MNDFIX	M2 : nombre quelconque de modalités et α fixe
MNDVAR	M3 : nombre quelconque de modalités et $\alpha = (\alpha^1, \dots, \alpha^p)$
MNDCLA	M4 : nombre quelconque de modalités et $\alpha = (\alpha_k^1, \dots, \alpha_k^p)$

5.2 Illustrations

Nous présentons maintenant une synthèse de toutes les simulations effectuées.

Première expérience

Nous avons tout d'abord comparé les différentes variantes entre elles à travers des applications sur des données simulées à partir des modèles M1, M2, M3 et M4. Les résultats obtenus sont ceux attendus. En effet, nous pouvons constater le caractère simpliste de la méthode MNDDIJ ne donnant des résultats acceptables que dans le seul cas où les données sont simulées à partir du modèle M1. La méthode MNDFIX convient pour les modèles M1 et M2, la méthode MNDVAR pour M1, M2 et M3. Enfin, nous avons apprécié le caractère général de la méthode MNDCLA qui donne de bons résultats quel que soit le modèle (M1, M2, M3 ou M4) utilisé pour la simulation.

Seconde expérience : MNDDIJ et MNDQAL

A l'aide du modèle M1 associé à l'algorithme MNDDIJ, nous avons simulé des données structurées en 3 classes en utilisant différentes valeurs du paramètre α . Pour chaque valeur de α , nous avons appliqué les deux algorithmes et les résultats obtenus pour un ensemble de 100 individus décrits par 5 variables (ayant toutes 5 modalités) sont résumés dans la figure 1 ci-dessous.

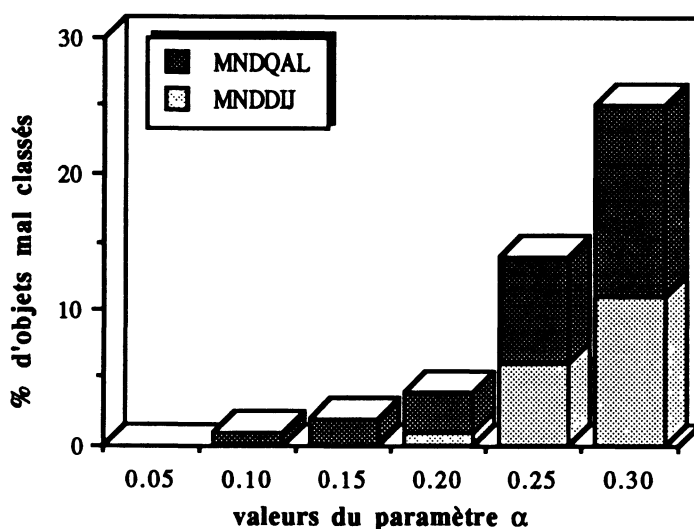


FIGURE 1

pourcentage d'objets mal classés par rapport à la partition simulée en fonction des valeurs prises par le paramètre α

Nous avons également réalisé de nombreux autres essais avec des tableaux de données et de classes plus importants. Et, toujours, la méthode MNDDIJ s'est avérée équivalente sinon meilleure que la méthode MNDQAL pour des données simulées.

Troisième expérience : MNDFIX et MNDQAL

Les données sont simulées suivant le modèle M2 associé à MNDFIX. Dans ce cas, la méthode MNDQAL est aussi mise à défaut par la méthode MNDFIX.

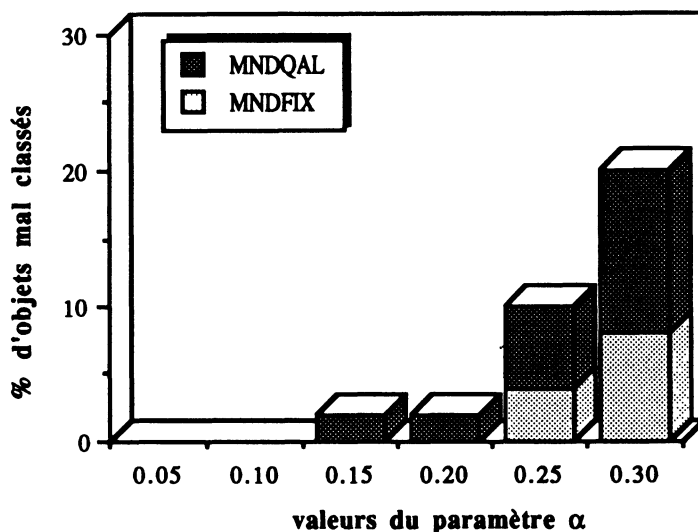


FIGURE 2

pourcentage d'objets mal classés par rapport à la partition simulée en fonction des valeurs prises par le paramètre α

Remarque

Nous avons également comparé MNDVAR et MNDCLA avec MNDQAL. Nous avons constaté, de façon plus marquée encore, l'inadéquation de la méthode MNDQAL sur les données simulées suivant les modèles M3 et M4.

6. Conclusion

Dans ce travail, nous avons proposé des méthodes de classification pour données qualitatives nominales, liées à un modèle précis de mélanges de distributions de probabilité. Ces liens justifient les bons résultats obtenus à l'aide de ces méthodes appliquées sur des données simulées suivant justement chacun des modèles associés. En effet, nous avons souligné par de nombreuses expériences l'adéquation des méthodes proposées et nous avons pu mettre à défaut la méthode de classification MNDQAL. Ces nouvelles méthodes permettent de caractériser les classes de la partition obtenue par un vecteur de modalités facilement interprétable. L'introduction de noyaux de modalités et l'existence d'un modèle probabiliste associé nous semblent constituer les deux points essentiels de notre étude.

Cette approche nous a permis d'apporter un éclairage nouveau de la méthode de classification MNDDIJ en lui associant un modèle probabiliste bien défini. De plus, nous avons montré que le modèle associé aux données binaires est un cas particulier du modèle proposé ici pour les données qualitatives nominales. Il nous reste cependant à confronter ces nouvelles méthodes à des données réelles.

7. Références

- BENZECRI J.P (1973) Théorie de l'information et classification d'après un tableau de contingence. L'Analyse des données tome 1, Dunod.
- CELEUX G. (1988) Classification et modèles. *R.S.A.* 4 : pp 43-58.
- EVERITT B. (1981) An introduction to latent variable models. Chapman and Hall.
- GOODMAN L. (1974) Exploratory latent structure models using both identifiable and unidentifiable models. *Biometrika*. 61.
- GOVAERT G. (1990) Classification binaire et modèles. *R.S.A.* 38 : pp 67-81.
- MARCHETTI F. (1989) Contribution à la classification de données binaires et qualitatives. Thèse de doctorat de l'Université. Université de Metz.
- RALAMBONDRAINY H. (1988) Etude des données qualitatives par les méthodes typologiques. Actes au congrès de l'Association Française de Marketing. Montpellier.
- SCHROEDER A. (1976) Analyse d'un mélange de distributions de probabilités de même type. *R.S.A.* vol 24, n°1 : pp 39-62.
- SCOTT A.J. et SYMONS M.J. (1971) Clustering methods based on likelihood ratio criteria. *Biometrics* 27 : pp 387-397.