REVUE DE STATISTIQUE APPLIQUÉE

YAMINA KHEMAL BENCHEIKH

Classification croisée et mélanges sur données quantitatives

Revue de statistique appliquée, tome 52, n° 2 (2004), p. 71-86

<http://www.numdam.org/item?id=RSA_2004__52_2_71_0>

Rev. Statistique Appliquée, 2004, **LII** (2), 71-86

CLASSIFICATION CROISÉE ET MÉLANGES SUR DONNÉES QUANTITATIVES

Yamina KHEMAL BENCHEIKH

Département de Mathématiques
Faculté des sciences
Université Ferhat Abbas de Sétif
Sétif 19000 – ALGÉRIE

RÉSUMÉ

La méthode de reconnaissance des composants d'un mélange nous permet d'interpréter des méthodes de classification simple. Nous proposons ici de le faire pour les méthodes de classification croisée et en particulier sur des tableaux de mesures. Ces derniers posent un problème car les deux ensembles décrivant ces tableaux sont de nature différente. Nous montrons alors que les algorithmes de partitionnements simultanés utilisant un critère d'inertie peuvent se présenter comme des méthodes pour identifier un mélange gaussien à l'aide d'une classification : suivant différentes hypothèses sur les variances des différents composants, on retrouve à chaque fois le critère d'inertie utilisé pour la classification de données quantitatives. Nous montrons aussi que dans le cas où les variances sont toutes les même pour tous les composants cela nous conduit au même critère qui correspond à la version la plus simple et la plus utilisée de la méthode CROEUC (méthode de classification croisée sur tableau de mesures optimisant un critère défini à l'aide d'une distance euclidienne).

Mots-clés : *Classification, Classification croisée, Critère d'inertie, Mélange gaussien, Tableaux de mesures.*

ABSTRACT

The method of identifying components of a mixture allows to interpret simple clustering methods. We propose it for cross clustering methods, in particular on measures tables, which present a problem for the different nature of the two sets describing these tables. Then we show that simultaneous partitioning algorithms using a criterion of inertia, can be considered as methods for identifying a gaussian mixture by using a clustering. According to different hypotheses on variances of different components, every time, we find again the criterion of inertia used for the quantitative data clustering. We also show, that where all variances are the same for all components, this leads to the same criterion which corresponds to the CROEUC method (cross clustering method on measures tables optimizing a defined criterion by using an Euclidean metric) most simple and used version.

Keywords : *Clustering, Cross clustering, Criterion of inertia, Gaussian mixture, Measures tables.*

Introduction

Les méthodes de classification reposent essentiellement sur la définition d'une métrique et d'un critère associé sans faire référence explicitement à des modèles probabilistes. En réalité comme Scott et Symons [14], Schroeder [13], Celeux [6] le proposent, il est souvent possible de montrer qu'il y a un modèle sous-jacent. Celui-ci permet alors de donner une interprétation du critère et de justifier son choix. La plupart de ces approches ont été faites dans le cas de la classification simple, nous proposons de le faire dans ce papier pour la classification croisée de données quantitatives, le cas de données binaires ayant déjà été traité par Bencheikh [3].

La méthode **CROEUC** (Govaert [9]) correspond à la classification croisée de tableaux de mesures optimisant un critère défini à l'aide d'une distance Euclidienne ; dans le premier paragraphe, nous décrivons cette méthode. Dans le second paragraphe, nous rappelons le principe de la méthode de reconnaissance des composants d'un mélange croisé (Bencheikh [2]). Cette dernière permet de placer les méthodes de classification croisée dans le même cadre que celui fait pour les méthodes de classification simple Schroeder [13], Celeux [6] et Govaert [10]. Dans le troisième paragraphe nous montrons comment la méthode précédente peut être interprétée comme l'approche classification associée à un mélange de lois gaussiennes unidimensionnelles ; suivant différentes hypothèses sur les variances des différents composants du mélange, on retrouve à chaque fois le critère d'inertie utilisé pour la classification de données quantitatives. La comparaison des résultats obtenus en particulier dans le troisième cas où les variances de chaque composant du mélange sont différentes entres elles, permet de mieux comprendre les liens qui existent entre l'algorithme des distances adaptatives et celui de la reconnaissance de mélange gaussien. La méthode des distances adaptatives correspond donc elle aussi à un modèle de reconnaissance de mélange gaussien.

1. La méthode CROEUC

L'objectif de la méthode de classification croisée sur données quantitatives est la recherche d'un couple de partitions, l'une sur les individus (les lignes du tableau étudié), l'autre sur les colonnes (variables), tel que la «perte d'information» due au regroupement soit minimale ; c'est-à-dire telle que la différence entre l'information apportée par le tableau initial et celle apportée par le tableau obtenu après regroupement soit minimale.

1.1. Notations

Soit $\mathbf{I}$ un ensemble de n individus décrits par p variables quantitatives. Les données sont rangées dans un tableau de description $\mathbf{X}$ à n lignes et p colonnes

$$\mathbf{X} = (x_i^j) \quad i \in \mathbf{I} \text{ et } j \in \mathbf{J}$$

où x_i^j est la valeur de la variable j pour l'individu i.

– $\mathbf{I}$, un sous ensemble fini de $\mathbf{R}^p$, contenant n éléments.

– $\mathbf{J}$, un sous ensemble fini de $\mathbf{R}^n$, contenant p éléments.

– $\mathbf{P} = (P_1, \ldots, P_K)$ représente une partition de $\mathbf{I}$ en K classes.

– $\mathbf{Q} = (Q^1, \ldots, Q^M)$ représente une partition de $\mathbf{J}$ en M classes.

– $\mathbf{L} = \{\lambda_k^m, k = 1, \ldots, K \text{ et } m = 1, \ldots, M\}$ l'ensemble des noyaux, ces noyaux seront associés aux partitions des deux ensembles $\mathbf{I}$ et $\mathbf{J}$.

D'autre part on supposera que les individus sont munis de poids p_i tels que : $\sum_{i \in I} p_i = 1$. Associons à chaque individu i le vecteur $x_i = (x_i^1, \ldots, x_i^p)$ de $\mathbf{R}^p$ correspondant à la ligne i du tableau X.

L'ensemble des x_i munis des pondérations p_i forme un nuage $N(\mathbf{I})$ contenu dans $\mathbf{R}^p$. De la même façon, à chaque variable j est associé le vecteur $x^j = (x_1^j, \ldots, x_n^j)$ correspondant à la colonne j du tableau $\mathbf{X}$.

L'ensemble des x^j munis des pondérations q_j forment un nuage $N(\mathbf{J})$ contenu dans $\mathbf{R}^n$.

Pour mesurer la proximité entre individus, on munit l'espace des individus de la métrique quadratique définie par la matrice diagonale de terme générale q_j correspondant à l'importance donnée à la variable j :

$$d^2(x_i, x_{i'}) = \sum_{j=1}^{p} q_j (x_i^j - x_{i'}^j)^2$$

De même pour mesurer la proximité entre variables, on munit l'espace des variables de la métrique des poids D_p :

$$d^2(x^j, x^{j'}) = \sum_{i=1}^{n} p_i (x_i^j - x_i^{j'})^2$$

1.2. Le problème

La méthode CROEUC fournit une solution locale au problème d'optimisation suivant :

Il s'agit de trouver une partition $(P_1, \ldots, P_k)$ *de* $\mathbf{I}$ *en* K *classes, une partition* $(Q^1, \ldots, Q^M)$ *de* $\mathbf{J}$ *en* M *classes et un ensemble* L *de noyaux tel que le critère d'inertie intraclasse suivant :*

$$\mathbf{W}(P \times Q, L) = \sum_{k=1}^{K} \sum_{m=1}^{M} \sum_{i \in P_k} \sum_{j \in Q^m} p_i q_j (x_i^j - g_k^m)^2 \qquad (1)$$

soit minimal.

g_k^m : qui est définie par la formule (1bis), étant la moyenne de la classe $P_k \times Q^m$ ($g_k^m \in \mathbf{R}$). $P \times Q = \{P_k \times Q^m, k = 1, \ldots, K \text{ et } m = 1, \ldots, M\}$ et $P_k \times Q^m = \{x_i^j \in \mathbf{R} / i \in P_k \text{ et } j \in Q^m\}$ $L = \{g_k^m, k = 1, \ldots, K \text{ et } m = 1, \ldots, M\}$

1.3. L'algorithme

Plusieurs algorithmes de classification croisée existent, on a retenu l'algorithme développé par Govaert [9] ; celui-ci utilise deux algorithmes voisins l'un de l'autre, tous les deux étant basés sur le principe des Nuées Dynamiques (Diday [7]). Le principe général de cet algorithme est le suivant : à partir d'une partition $(P^\circ \times Q^\circ)$ en $K \times M$ classes, on construit une suite de partitions en appliquant successivement les trois fonctions suivantes :

La fonction de représentation g

Cette fonction permet de déterminer les $K \cdot M$ noyaux minimisant le critère : $\mathbf{W}(P \times Q, g(P \times Q))$. On peut facilement voir que ces noyaux sont les moyennes des classes. Si nous notons $\{(g_k^m); k = 1, \ldots, K \text{ et } m = 1, \ldots, M\}$ l'ensemble de ces centres, on a :

$$g_k^m = \frac{1}{(\sum_{i \in P_k} \sum_{j \in Q^m} p_i \cdot q_j)} \sum_{i \in P_k} \sum_{j \in Q^m} p_i q_j x_i^j \qquad (1bis)$$

La fonction d'affectation f

Cette fonction minimise le critère $\mathbf{W}(f(Q, L) \times Q, L)$ en affectant chaque individu à la classe P_k du noyau g_k de laquelle il est le plus proche (au sens de la distance euclidienne) en supposant que la partition Q et le noyau L sont fixés. Considérons le nuage des individus obtenu après regroupement des variables selon la partition Q; il est défini par :

$$\{x_i = (x_i^1, \ldots, x_i^M), i \in \mathbf{I}\} \quad \text{où} \quad x_i^m = \frac{1}{\sum_{j \in Q^m} q_j} \sum_{j \in Q^m} q_j x_i^j$$

Considérons l'algorithme des Nuées Dynamiques (Diday [7]) suivant :

Le tableau de données est le tableau $\mathbf{X}(\mathbf{I}, Q)$ défini par les x_i^m, les individus sont en ligne et les variables en colonne. On a donc un ensemble de n éléments et M variables.

– Les vecteurs sont les lignes du tableaux $\mathbf{X}(\mathbf{I}, Q)$.

– Les pondérations des individus sont toujours les p_i.

– Les noyaux sont de la forme $(g_k^1, \ldots, g_k^M)$ qui représente le centre de gravité de la classe P_k.

– La métrique associée à ce nouveau tableau $\mathbf{X}(\mathbf{I}, Q)$ est définie par la matrice diagonale de terme général q'_m où $q'_m = \sum_{j \in Q^m} q_j$.

La fonction d'affectation f range alors chaque élément i dans la classe ayant le noyau le plus proche au sens de la métrique précédente.

La fonction d'affectation h

Cette fonction minimise le critère $\mathbf{W}(P \times h(P, L), L)$, en supposant que la partition P et le noyau L sont fixés. On applique la méthode des Nuées Dynamiques sur le tableau $\mathbf{X}(P, \mathbf{J})$ défini par les x_k^j. Les individus sont en colonnes et les variables en lignes. On obtient un ensemble de p éléments et K variables. La fonction d'affectation h range alors chaque élément $j \in \mathbf{J}$ dans la classe ayant le noyau le plus proche au sens de la métrique diagonale dont la diagonale a pour terme général $p'_k = \sum_{i \in P_k} p_i$.

La définition de ces trois fonctions entraîne que la suite $\mathbf{W}(P^n \times Q^n, L^n)$ est décroissante ; on retrouve les propriétés habituelles de convergence des Nuées Dynamiques (Diday [7]).

1.4. Cas particulier

Posons : $p_i = \frac{1}{n}$, $q_j = \frac{1}{p}$ $\forall i \in \mathbf{I}$ et $j \in \mathbf{J}$

Le critère (1) s'écrit :

$$\mathbf{W}(P \times Q, L) = \sum_{k=1}^{K} \sum_{m=1}^{M} \sum_{i \in P_k} \sum_{j \in Q^m} \frac{1}{n \cdot p} (x_i^j - g_k^m)^2 \qquad (2)$$

avec $g_k^m = \frac{1}{n_k \cdot q_m} \sum_{i \in P_k} \sum_{j \in Q^m} x_i^j$ où $n_k = \text{Card}\,(P_k)$ et $q_m = \text{Card}\,(Q^m)$

La minimisation du critère (2) revient à la minimisation du critère suivant :

$$\mathbf{W}(P \times Q, L) = \sum_{k=1}^{K} \sum_{m=1}^{M} \sum_{i \in P_k} \sum_{j \in Q^m} (x_i^j - g_k^m)^2 \qquad (3)$$

qui correspond à la version la plus simple et la plus utilisée de la méthode CROEUC.

2. Méthode de reconnaissance des composants d'un mélange croisé

La méthode de reconnaissance des composants d'un mélange croisé a été proposée par Bencheikh [1] en 1992 ; elle s'est intéressée aux méthodes de classification

croisée proposées par Govaert [9] et [11] ; elle montre que ces méthodes, comme les méthodes de classification simple, peuvent souvent être considérées comme une approche classification d'un modèle de mélange. Pour ceci, elle propose la notion de mélange croisé en se basant sur un exemple concret (Bencheikh [2]) et définit les notions de vraisemblance et de vraisemblance classifiante associées. Elle étudie au passage les liens avec les modèles de mélange simple proposés par Schroeder [13], Celeux [6] et Govaert [10] et montre que ces liens sont tout à fait analogues à ceux qui existent entre les méthodes de classification simple et les méthodes de classification croisée.

Avant de rappeler le principe général de cette méthode, notons que Bzioui [4] s'est aussi intéressé aux liens qui existent entre les méthodes de classification croisée et les modèles probabilistes ; il propose un modèle probabiliste qui va dans le même sens que celui étudié ici, en adoptant la même démarche probabiliste mais en définissant cette fois-ci un modèle de mélange croisé qui respecte la structure des lignes et des colonnes ; il considère le tableau de données comme un échantillon de taille 1 comme c'est souvent le cas en traitement statistique d'image. Notre approche diffère de celle de Bzioui [4], dans le sens où comme on va le voir les données du tableau seront considérées comme un échantillon de taille np.

2.1. Notion de mélange croisé

Les données sont toujours fournies sous la forme d'un tableau rectangulaire à n lignes et p colonnes. Rappelons que $\mathbf{I}$ est un ensemble de n individus et $\mathbf{J}$ est un ensemble de p variables. $\mathbf{I} \times \mathbf{J} = \{(i,j)/i \in \mathbf{I} \text{ et } j \in \mathbf{J}\}$ est un ensemble de np individus-variables; associons à chaque individu-variable la valeur $x_i^j \in \mathbf{R}$ qui correspond à la valeur prise par la variable j pour l'individu i.

On suppose maintenant que l'ensemble $\mathbf{I}$ des individus constitue un échantillon de taille n d'une population Ω, de même on considère que l'ensemble $\mathbf{J}$ des variables constitue un échantillon de taille p d'une population Ω'.

Soit $\mathbf{T} = \mathbf{I} \times \mathbf{J}$ le produit cartésien des deux ensembles $\mathbf{I}$ et $\mathbf{J}$; l'ensemble $\mathbf{T}$ peut être considéré comme un échantillon de taille $n \cdot p$ d'une population $\Omega \times \Omega'$. On peut toujours définir une variable aléatoire $\mathbf{Z}$ qui permet d'associer à chaque couple $(i,j) \in \mathbf{I} \times \mathbf{J}$ la valeur se trouvant à l'intersection de la ligne i avec la colonne j qui est x_i^j.

2.2. Identification d'un mélange « croisé »

Le modèle d'un mélange croisé est le modèle probabiliste qui cherche à reconnaître dans une population observée du type $\Omega \times \Omega'$ l'éventuelle présence d'échantillons de lois de probabilité connues (Bencheikh [2]) . Nous proposons alors le modèle suivant.

Le tableau de données de départ de dimension (n,p) est considéré comme un échantillon $\mathbf{T} = \mathbf{I} \times \mathbf{J}$ de taille $(n \cdot p)$ d'une variable aléatoire à valeur dans $\mathbf{R}$ dont

la loi de probabilité admet la fonction de densité :

$$f(x) = \sum_{k=1}^{K} \sum_{m=1}^{M} p_k^m f(x/\lambda_k^m) \tag{4}$$

$\forall x \in \mathbf{R}\ \forall k = 1, \ldots, K$ et $m = 1, \ldots, M \quad 0 \leqslant p_k^m \leqslant 1$ et $\sum_{k=1}^{K} \sum_{m=1}^{M} p_k^m = 1$

La formule (4) décrit le modèle d'un mélange de type donné $f(\cdot, \lambda_k^m)$ qui est une fonction de densité sur $\mathbf{R}$ appartenant à une famille paramétrée de fonctions de densité dépendant du paramètre λ_k^m et p_k^m est la probabilité d'apparition de l'observation $f(\cdot, \lambda_k^m)$ dans le mélange.

2.3. Problème à résoudre

Problème 1

Le problème consiste à estimer les nombres $\mathbf{K}$ *et* $\mathbf{M}$ *de composants du mélange et les paramètres inconnus* q_k^m *:*

$(q_k^m = (p_k^m, \lambda_k^m); k = 1, \ldots, K$ *et* $m = 1, \ldots, M)$ *au vu de l'échantillon* $\mathbf{T} = \mathbf{I} \times \mathbf{J}$.

Il s'agit d'un problème d'estimation de paramètres. Nous ne le traiterons pas par l'approche « estimation » du problème mais nous nous concentrerons sur l'approche « classification » pour l'identification d'un mélange « croisé ».

2.4. Approche classification

Dans cette approche on remplace le problème 1 d'estimation par le problème 2 suivant :

Problème 2

Rechercher une partition $P \times Q = \{P_k \times Q^m\ ;\ k = 1, \ldots, K$ *et* $m = 1, \ldots, M\}$, *K et M étant supposés connus, telle que chaque classe* $P_k \times Q^m$ *soit assimilable à un sous-échantillon qui suit une loi* $f(\cdot, \lambda_k^m)$.

En suivant l'approche modèle proposé par Schroeder [13] et la représentation de Celeux [6] qui transforme le problème d'optimisation de critère de vraisemblance en un problème d'optimisation de critère de vraisemblance classifiante, on se ramène également, dans le cas de mélange « croisé » , à la maximisation du critère de vraisemblance classifiante suivant :

$$\mathbf{VC}(P \times Q, L) = \sum_{k=1}^{K} \sum_{m=1}^{M} \mathrm{Log} R(P_k \times Q^m, \lambda_k^m) \tag{5}$$

où L est le $K \cdot M$-uple $(\lambda_k^m, k = 1, \ldots, K$ et $m = 1, \ldots, M)$ et $R(P_k \times Q^m, \lambda_k^m)$ est la vraisemblance du sous-échantillon $P_k \times Q^m$ qui suit la loi $f(\cdot, \lambda_k^m)$:

$$R(P_k \times Q^m, \lambda_k^m) = \prod_{x \in P_k \times Q^m} f(x/\lambda_k^m).$$

Le critère de vraisemblance classifiante s'écrit alors :

$$\mathbf{VC}(P \times Q, L) = \sum_{k=1}^{K} \sum_{m=1}^{M} \sum_{i \in P_k} \sum_{j \in Q^m} \mathrm{Log} f(x_i^j/\lambda_k^m) \tag{6}$$

On peut remarquer que la résolution du problème (6) correspond exactement à la résolution d'un problème de classification croisée (Govaert [9]). Nous proposons d'utiliser le même algorithme que celui défini pour la classification croisée et rappelé dans le paragraphe 1.3. Cet algorithme, que nous allons reprendre pour l'adapter à notre problème, utilise deux algorithmes voisins l'un de l'autre, tous les deux étant basés sur le principe des Nuées Dynamiques.

2.5. Algorithme

Le principe de cet algorithme est le suivant : en partant de deux nombres K et M et d'une partition initiale $(P \times Q)^\circ$ en $K \cdot M$ classes, l'algorithme construit une suite de partitions-noyaux jusqu'à l'obtention d'une partition stable en appliquant successivement les trois fonctions suivantes.

Une fonction de représentation g définie comme suit

Cette fonction permet de déterminer les $K \cdot M$ noyaux maximisant le critère (6) : $\mathbf{VC}(P \times Q, g(P \times Q)) = \mathrm{Max}\, VC(P \times Q, L)$.

On peut facilement voir que ces noyaux sont les estimateurs du maximum de vraisemblance des paramètres associés aux sous échantillons $\{P_k \times Q^m; k = 1, \ldots, K$ et $m = 1, \ldots, M\}$.

$g(P \times Q) = g(\{P_k \times Q^m; k = 1, \ldots, K$ et $m = 1, \ldots, M\}) = \{\lambda_k^m, k = 1, \ldots, K$ et $m = 1, \ldots, M\} = L$.

Une fonction d'affectation f définie comme suit

Cette fonction permet de déterminer, à Q et L fixés, une partition P de l'échantillon $\mathbf{I}$ améliorant le critère $\mathbf{VC}(P \times Q, L)$. Le critère (6) s'écrit alors :

$$\mathbf{VC}(P \times Q, L) = \sum_{k=1}^{K} \sum_{i \in P_k} \mathrm{Log} F(x_i/\lambda_k)$$

où $F(x_i/\lambda_k) = \prod_{m=1}^{M} \left(\prod_{j \in Q^m} f(x_i^j/\lambda_k^m) \right)$ et $\lambda_k = (\lambda_k^1, \ldots, \lambda_k^M)$

On retrouve ainsi la forme du critère de vraisemblance classifiante dans le cas de la classification simple. Les éléments de la classe P_k seront définis comme suit :

$$P_k = \{i \in \mathbf{I}/F(x_i/\lambda_k) \geqslant F(x_i/\lambda_{k'}) \quad \text{avec } k < k' \text{ en cas d'égalité}\}.$$

Une fonction d'affectation h définie comme suit

Cette fonction permet de déterminer, à P et L fixés, une partition Q de l'échantillon $\mathbf{J}$ améliorant le critère :

$$\mathbf{VC}(P \times Q, L) = \sum_{m=1}^{M} \sum_{j \in Q^m} \mathrm{Log} F(x^j/\lambda^m)$$

où $F(x^j/\lambda^m) = \prod_{k=1}^{K} (\prod_{i \in P_k} f(x_i^j/\lambda_k^m))$ et $\lambda^m = (\lambda_1^m, \ldots, \lambda_K^m)$

Les éléments de la partition Q de l'échantillon $\mathbf{J}$ seront déterminés comme suit :

$$Q^m = \{j \in \mathbf{J}/F(x^j/\lambda^m) \geqslant F(x^j/\lambda^{m'}) \quad \text{avec } m < m' \text{ en cas d'égalité}\}.$$

On peut montrer que sous certaines hypothèses (Govaert [9]), cet algorithme est convergent. On obtient à la convergence une partition $P \times Q$ et une estimation des paramètres λ_k^m. Les proportions p_k^m du mélange sont fournies par les fréquences des classes $P_k \times Q^m$.

3. Modèle associé aux données quantitatives

La méthode de reconnaissance des composants d'un mélange croisé rappelée au paragraphe 2 permet d'interpréter des méthodes de classification croisée qui traite les deux ensembles I et J décrivant le tableau de données de manière identique ; par exemple les tableaux binaires.

Les tableaux de mesures (données quantitatives) posent un problème, car comme nous venons de le signaler les ensembles décrivant ces tableaux sont de nature différente. Pour pouvoir appliquer la méthode de reconnaissance des composants d'un mélange croisé, nous nous sommes efforcés dans tout ce travail de considérer avec un peu « d'abus » que l'ensemble J des p variables quantitatives constitue un échantillon de taille p d'une population Ω' (où Ω' est l'ensemble de toutes les variables possibles).

3.1. Choix de la famille de distribution

On considère dans ce modèle que les données du tableau proviennent d'un mélange de lois gaussiennes unidimensionnelles, les paramètres λ_k^m s'écrivent $\lambda_k^m = (\mu_k^m, \sigma_k^m)$.

μ_k^m : espérance du composant (k, m)

σ_k^m : écart-type du composant (k, m)

Le composant (k, m) est le sous-échantillon associé à la classe $P_k \times Q^m$ ayant pour fonction de densité la loi $f(\cdot/\lambda_k^m)$. Posons $V_k^m = (\sigma_k^m)^2$ qui représente la variance du composant (k, m).

D'où : $f(x/\lambda_k^m) = (2\pi \cdot V_k^m)^{-\frac{1}{2}} \cdot \exp - \dfrac{(x - \mu_k^m)^2}{2 \cdot V_k^m}$

Les paramètres à estimer sont les μ_k^m et les V_k^m.

La maximisation du critère $\mathbf{W}(P \times Q, L)$ donné par la formule (6) revient à la minimisation du critère suivant :

$$\mathbf{C}(P \times Q, L) = \sum_{k=1}^{K} \sum_{m=1}^{M} \sum_{i \in P_k} \sum_{j \in Q^m} \left\{ \frac{(x_i^j - \mu_k^m)^2}{V_k^m} + \mathrm{Log} V_k^m \right\} \tag{7}$$

Pour optimiser ce critère, on va devoir étudier trois cas différents en faisant plusieurs hypothèses sur les variances des composants du mélange.

Premier cas

$V_k^m = V \quad \forall\, k = 1, \ldots, K$ et $m = 1, \ldots, M$ et V est supposé connu. La maximisation du critère (7) revient à la minimisation du critère suivant :

$$\mathbf{C_1}(P \times Q, L) = \sum_{k=1}^{K} \sum_{m=1}^{M} \sum_{i \in P_k} \sum_{j \in Q^m} (x_i^j - \mu_k^m)^2 \tag{8}$$

Les partitions étant fixées, l'estimateur du maximum de vraisemblance classifiante de μ_k^m est la moyenne de la classe $P_k \times Q^m$. Le critère (8) prend alors la forme :

$$\mathbf{W}(P \times Q, L) = \sum_{k=1}^{K} \sum_{m=1}^{M} \sum_{i \in P_k} \sum_{j \in Q^m} (x_i^j - g_k^m)^2$$

Ce critère correspond bien au critère minimisé par la méthode CROEUC dans le cas le plus simple (formule (3)).

Deuxième cas

Les variances de tous les composants du mélange sont les mêmes et supposées inconnues. Notons par W l'expression $\sum_{k=1}^{K} \sum_{m=1}^{M} \sum_{i \in P_k} \sum_{j \in Q^m} (x_i^j - \mu_k^m)^2$

Le critère à minimiser s'écrit : $\mathbf{C}(P \times Q, L) = \frac{1}{V} \cdot W + n \cdot p\mathrm{Log}V$

$$\frac{\partial C(P\times Q)}{\partial V}=\frac{\partial}{\partial V}(\frac{1}{V}W+np\mathrm{Log}V)=0 \quad \text{d'où} \quad V=\frac{W}{n\cdot p}$$

$$\mathbf{C}(P\times Q,L)=n\cdot p+n\cdot p\mathrm{Log}W-n\cdot p\mathrm{Log}(n\cdot p)$$

Cela revient à minimiser le critère :

$$\mathbf{C_2}(P\times Q,L)=\mathrm{Log}W \tag{9}$$

qui est équivalent à la minimisation du critère :

$$\mathbf{W}(P\times Q,L)=W=\sum_{k=1}^{K}\sum_{m=1}^{M}\sum_{i\in P_k}\sum_{j\in Q^m}(x_i^j-\mu_k^m)^2$$

Là aussi l'estimateur de μ_k^m n'est autre que la moyenne de la classe $P_k\times Q^m$.

Troisième cas

Les variances de chaque composante sont différentes entre elles et inconnues.

Posons : $L_k^m=\sum_{i\in P_k}\sum_{j\in Q^m}\left\{\frac{(x_i^j-\mu_k^m)^2}{V_k^m}+\mathrm{Log}V_k^m\right\}$ et

$W_k^m=\sum_{i\in P_k}\sum_{j\in Q^m}(x_i^j-\mu_k^m)^2$

Le critère à minimiser s'écrit alors :

$$C(P\times Q,L)=\sum_{k=1}^{K}\sum_{m=1}^{M}L_k^m=\sum_{k=1}^{K}\sum_{m=1}^{M}\frac{W_k^m}{V_k^m}+n_k\cdot q_m\cdot\mathrm{Log}V_k^m$$

$$\frac{\partial L_k^m}{\partial V_k^m}=\frac{\partial}{\partial V_k^m}\left(\frac{W_k^m}{V_k^m}+n_k\cdot q_m\cdot\mathrm{Log}V_k^m\right)=0 \quad \text{d'où} \quad V_k^m=\frac{W_k^m}{n_k\cdot q_m}$$

$$\frac{\partial L_k^m}{\partial \mu_k^m}=\frac{\partial}{\partial \mu_k^m}(n_k\cdot q_m+n_k\cdot q_m\cdot\mathrm{Log}\frac{W_k^m}{n_k\cdot q_m})=0$$

$$\text{d'où} \quad \mu_k^m=\frac{1}{n_k\cdot q_m}\sum_{i\in P_k}\sum_{j\in Q^m}x_i^j$$

Dans ce cas le critère (7) à minimiser se réduit à :

$$\mathbf{C_3}(P\times Q,L)=\sum_{k=1}^{K}\sum_{m=1}^{M}n_k\cdot q_m\cdot\mathrm{Log}V_k^m \tag{10}$$

Remarque. – Le critère de l'algorithme des distances adaptatives Govaert [8] élaboré dans un cadre purement géométrique pour permettre de reconnaître des classes de « formes » différentes s'écrit :

$$\mathbf{W}(P,L) = \sum_{k=1}^{K} n_k \cdot |V_k|^{\frac{1}{p}} \tag{11}$$

où V_k est la matrice de variance associée à la classe P_k (cas de la classification simple), et $|V_k|$ le déterminant de V_k. D'autre part le critère proposé par Celeux [6] pour l'identification d'un mélange gaussien à l'aide d'une classification simple s'écrit :

$$\mathbf{W}(P,L) = \sum_{k=1}^{K} n_k \cdot \log |V_k| \tag{12}$$

Celeux [6] affirme que les critères (11) et (12) ne sont pas identiques, mais sont très analogues et qu'en pratique, ces deux méthodes donnent des résultats quasi-identiques (*cf* Govaert [8] et Schroeder [13]) .

On peut facilement voir que l'expression du critère donnée par la formule (3), où les métriques sont toutes les deux de la formes $\gamma \cdot I_d$ où I_d est la matrice identité et γ est un réel, peut s'écrire sous la forme suivante :

$$\mathbf{W}(P \times Q, L) = \sum_{k=1}^{K} \sum_{m=1}^{M} W_k^m = \sum_{k=1}^{K} \sum_{m=1}^{M} n_k \cdot q_m \cdot V_k^m \tag{13}$$

et on a vu que pour optimiser ce critère, on utilise l'algorithme proposé au paragraphe 1.3 ; ce dernier utilise deux algorithmes voisins l'un de l'autre de classification simple, en fixant à chaque étape une partition et on améliore l'autre partition et *vice versa.* Ces deux algorithmes ne sont autre que les algorithmes correspondant a la méthode (11).

On remarque aussi que lorsqu'on fixe l'une où l'autre des deux partitions P et Q, les deux critères (10) et (13) se comportent respectivement comme les critères (12) et (11). En se basant sur la remarque faite ci-dessus, on peut affirmer qu'en pratique les méthodes (10) et (13) donnent à leur tour des résultats voisins.

En principe le critère (10) devrait donner des résultats meilleurs que ceux obtenus avec la méthode classique CROEUC qui utilise le critère (3), cette dernière utilise une distance fixe et identique pour toute affectation d'élément par contre la méthode correspondant au critère (10) utilise des distances adaptatives pour l'affectation des éléments aux différentes classes. Cette dernière tient compte des partitions durant le déroulement de l'algorithme et par conséquent une meilleure classification est prévue. Une étude analogue à été faite pour la méthode CROBIN (Classification croisée sur données binaires) par Bencheikh [3] et a permis de mettre la méthode utilisant des distances non adaptatives à défaut. Nous proposons dans le paragraphe suivant de comparer les méthodes correspondantes respectivement aux critères (8) et (10) en les appliquant sur deux types de données.

4. Application et comparaison des méthodes

Le modèle proposé dans ce travail pour la méthode CROEUC, a non seulement permis de justifier le choix du critère utilisé par celle-ci, mais de proposer une nouvelle méthode de classification croisée sur des données continues utilisant des distances adaptatives que nous appellerons CROEUC adaptatif. Ce dernier est composé de deux variantes : la variante 1 associée au critère défini par la formule (8) (CROEUC), et la variante 2 associée au critère défini par la formule (10).

Stratégie utilisée

Pour réduire l'influence de la partition d'initialisation des algorithmes sur les résultats, nous avons opté pour la démarche suivante : on applique la variante 1 de l'algorithme CROEUC adaptatif en demandant 20 tirages, puis on applique la variante 2 en l'initialisant par le meilleur couple de partitions obtenu par la variante 1. On compare les résultats obtenus par rapport au couple de partition d'initialisation dans le cas de données réelles, et au couple de partitions ayant servi à la simulation dans le cas de données simulées.

4.1. Données réelles

On a appliqué l'algorithme CROEUC adaptatif sur les données nommées : «Poissons d'Amiard» traitées dans Cailliez et Pages [5] en demandant trois classes en lignes et deux classes en colonnes.

Il s'agit de 24 mulets (sorte de rougets) qui ont été répartis dans trois aquariums radio-contaminés de façon identique. À ces trois aquariums correspondent des durées de contact avec le polluant radio-actif différentes :

– Le premier contient les poissons numérotés de 1 à 8

– Le second contient les poissons numérotés de 9 à 17

– Le troisième contient les poissons numérotés de 18 à 24

(Le poisson 17 est mort en cours d'expérience)

16 caractères sont mesurés. On peut séparer ces caractères en deux groupes.

Groupe 1 : *Caractéristiques de radio-activité*

Caractère 1 : radio-activité des yeux

Caractère 2 : radio-activité des branchies

Caractère 3 : radio-activité des opercules

Caractère 4 : radio-activité des nageoires

Caractère 5 : radio-activité du foie

Caractère 6 : radio-activité du tube digestif

Caractère 7 : radio-activité des reins
Caractère 8 : radio-activité des écailles
Caractère 9 : radio-activité des muscles

Groupe 2 : *Caractéristiques de taille*

Caractère 10 : poids
Caractère 11 : longueur
Caractère 12 : longueur standard
Caractère 13 : largeur de la tête
Caractère 14 : largeur
Caractère 15 : largeur du museau
Caractère 16 : diamètre des yeux

Tableau de données des poissons d'Amiart

	1	2	3	4	5	6	7	8	9	0	11	12	13	14	15	16
1	10	65	65	107	7	76	16	142	1	132	214	197	54	47	18	11
2	9	43	39	67	29	113	10	99	2	122	220	198	49	44	16	10
3	6	47	71	95	11	192	9	121	2	129	220	198	49	45	17	11
4	7	70	40	66	8	310	10	90	2	133	225	199	52	48	15	11
5	8	59	67	100	14	289	4	244	1	57	168	149	37	37	9	9
6	8	46	55	112	17	115	8	153	1	59	178	160	38	35	11	9
7	7	47	36	87	16	100	4	162	1	59	176	156	40	36	11	9
8	11	79	46	95	20	106	10	141	4	47	176	165	39	31	10	8
9	13	80	64	155	42	192	9	169	3	72	182	164	40	39	12	10
10	21	150	115	145	49	229	9	233	5	79	200	179	45	38	12	9
11	12	91	84	138	22	590	9	220	2	80	185	163	43	41	12	11
12	14	120	76	125	21	309	9	617	5	72	175	158	40	39	13	10
13	14	142	86	135	34	523	9	211	11	75	189	169	42	39	18	10
14	23	92	80	132	49	459	9	197	2	52	164	147	36	35	12	9
15	13	85	64	124	20	318	9	191	4	86	195	175	41	39	16	10
16	14	106	67	110	31	115	9	248	6	87	210	170	46	40	17	10
18	32	224	260	314	36	107	13	461	3	72	181	164	41	36	13	9
19	22	162	218	318	25	884	5	590	2	63	175	160	38	35	12	9
20	31	195	208	350	73	109	5	809	11	49	170	154	39	33	12	8
21	15	127	119	197	23	99	7	157	2	107	204	185	47	45	15	11
22	22	160	256	282	12	102	11	690	3	83	190	176	42	44	14	9
23	24	162	231	308	51	1031	17	558	2	82	194	168	42	39	14	10
24	19	64	163	229	16	109	8	345	1	91	190	172	44	42	13	11

Sur le tableau de données des poissons d'Amiart ci-dessus, nous avons appliqué les deux variantes de l'algorithme CROEUC adaptatif en suivant la stratégie définie au début du paragraphe 4.

Le couple de partitions obtenu par la variante 2 de l'algorithme CROEUC adaptatif, compte 4 et 3 éléments classés différemment respectivement en lignes et en colonnes par rapport au couple de partitions optimal obtenu par la variante 1.

La comparaison de ces partitions est délicate dans le cas de données réelles, pour cette raison nous proposons de nous appuyer sur des modèles probabilistes pour simuler des données sur lesquelles on a pu approfondir notre comparaison.

4.2. Données simulées

Nous avons simulé un tableau de données quantitatives de 20 lignes et 20 colonnes. Les paramètres ayant servi à la simulation de ce tableau sont :

$K = 2\ M = 2$

$n_1 = \text{Card}\,(P_1) = 10,\ n_2 = \text{Card}\,(P_2) = 10$

$q_1 = \text{Card}\,(Q^1) = 10,\ q_2 = \text{Card}\,(Q^2) = 10$

Les moyennes $((\mu_k^m); k = 1,2; m = 1,2)$: $(\mu_1^1 = 10, \mu_1^2 = 30, \mu_2^1 = 20, \mu_2^2 = 40)$

Les écart-types $((\sigma_k^m); k = 1,2; m = 1,2)$: $(\sigma_1^1 = 2, \sigma_1^2 = 2.5, \sigma_2^1 = 3, \sigma_2^2 = 4)$

Sur ce tableau de données nous avons appliqué les deux variantes de l'algorithme CROEUC adaptatif en suivant la stratégie définie au début du paragraphe 4.

Nous avons alors remarqué que le couple de partition obtenu par la variante 1 diffère de 4 éléments en lignes et 3 éléments en colonnes par rapport à la partition ayant servi à la simulation ; par contre la variante 2 donne exactement la même partition qui a servi à la simulation. Ce résultat nous semble important dans le sens où la variante 2 a permis de « reconnaître » les données qui ont été simulées suivant le même modèle probabiliste utilisé par la méthode correspondant au critère (10).

Conclusion

Nous venons de voir dans ce papier, comment les algorithmes de partitionnements simultanés utilisant un critère d'inertie peuvent se présenter comme des méthodes pour identifier un mélange gaussien à l'aide d'une classification. Suivant différentes hypothèses sur les variances des différents composants, on retrouve à chaque fois le critère d'inertie utilisé pour la classification de données quantitatives. Nous montrons aussi que dans le cas où les variances sont toutes les mêmes pour tous les composants, cela nous conduit au même critère qui correspond à la version la plus simple et la plus utilisée de la méthode CROEUC . La comparaison des résultats obtenus en particulier dans le troisième cas où les variances de chaque composant du mélange sont différentes entre elles, permet de mieux comprendre les liens qui existent entre l'algorithme des distances adaptatives et celui de la reconnaissance de mélange gaussien. La méthode des distances adaptatives correspond donc elle aussi à

un modèle de reconnaissance de mélange gaussien. De plus nous proposons dans ce papier un nouvel algorithme de classification croisée sur données quantitatives utilisant des distances adaptatives, ce dernier qui sera intégré au logiciel SICLA (Système interactif de classification automatique, INRIA), présente un avantage certain, du moins sur l'exemple des données simulées.

Références

[1] BENCHEIKH Y. (1992), *Classification automatique et modèles*, Thèse Université de Metz.

[2] BENCHEIKH Y. (1999), Classification croisée et modèles, *Rairo operations research*, vol 33.4 p 525-541.

[3] BENCHEIKH Y. (2002), Classification croisée et distance L1 adaptative, *Rev. Statistique Appliquée*, L(3), 53-72.

[4] BZIOUI M. (1999), *Classification croisée et modèles*, Thèse, Université de Metz.

[5] CAILLIEZ F., PAGES J.P. (1976), Introduction à l'analyse des données, *SMASH*.

[6] CELEUX G. (1988), Classification et modèles, *Rev statist. Appl*, n °4, p 43-58, 1988.

[7] DIDAY E. (1972), *Nouvelles méthodes et nouveaux concepts en classification automatique et reconnaissance des formes*, Thèse d'État, Université Paris 6.

[8] GOVAERT G. (1975), *Classification avec distance adaptative*, Thèse de Doctorat de 3^{eme} cycle, Paris 6.

[9] GOVAERT G. (1983), *Classification croisée*, Thèse de Doctorat d'État, Université Pierre et Marie Curie, Paris 6.

[10] GOVAERT G. (1990), Classification binaire et modèles, *Rev.Statistique Appliquée*, XXXVIII (1). 67-81.

[11] GOVAERT G. (1995), Simultaneous clustering of rows and columns, *Control and cybernetics*, vol 24, n °4.

[12] HARTIGAN J.-A. (1975), Statistical theory in clustering, *Journal of classification*, 2, 63-76.

[13] SCHROEDER A. (1974), *Reconnaissance des composants d'un mélange*, Thèse de Doctorat de 3^{eme} cycle, Université de Paris 6.

[14] SCOTT A. et SYMONS A. (1971), Clustering methods based on likelihood ratio criteria, *Biometrics* 27, 387-397.