RAIRO. RECHERCHE OPÉRATIONNELLE

Y. BENCHEIKH

Classification croisée et modèles

RAIRO. Recherche opérationnelle, tome 33, n° 4 (1999), p. 525-541

<http://www.numdam.org/item?id=RO_1999__33_4_525_0>

RAIRO Operations Research
RAIRO Oper. Res. **33** (1999) 525-541

CLASSIFICATION CROISÉE ET MODÈLES (*)

par Y. BENCHEIKH ([1])

Communiqué par Catherine ROUCAIROL

Résumé. – *Les liens existant entre les méthodes de classification automatique et les modèles de statistiques inférentielles ont surtout été étudiés lorsque les données mettent en jeu un seul ensemble. Nous nous proposons ici de le faire lorsque les données mettent en jeu deux ensembles. Nous nous sommes intéressés aux méthodes de classification croisée proposées par Govaert [6] ; nous montrons que ces méthodes, comme les méthodes de classification simple, peuvent être considérées, comme une approche classification d'un modèle de mélange. Nous introduisons la notion de mélange croisé à partir d'un exemple concret et nous définissons les notions de vraisemblance et de vraisemblance classifiante associées, nous étudions ensuite les liens qui existent entre les modèles de mélange croisé et les modèles de mélange simple et nous montrons que ces liens sont tout à fait analogues à ceux qui existent entre les méthodes de classification croisée et les méthodes de classification simple.*

Mots clés : Distance L_1, classification automatique, mélange de lois de probabilité, mélange croisé.

Abstract. – *The relations between automatic clustering methods and inferentiel statistical models have mostely been studied when the data involves only one set. We propose to study these relations in the case of data involving two sets. We shall look at cross clustering methods as suggested by Govaert [6]; we show that these methods, like the simple clustering methods, can be considered as a clustering approach of a mixture model. We introduce the notion of crossed mixture from a concret example and define the notions of likelihood and associated clustered likelihood. Then, we study the relations which exist between the crossed mixture models and simple models and we show that these relations are completely similar to those which exist between the crossed clustering methods and simple clustering methods.*

Keywords: L_1 distance, automatic clustering, mixture, cross mixture.

INTRODUCTION

L'une des principales difficultés pour les méthodes de classification automatique est souvent le choix de la métrique et du critère optimisés par ces méthodes. Ainsi lorsqu'il est possible de trouver un modèle de lois

(*) Reçu en janvier 1996.
([1]) Institut de Mathématiques, Université Férhat Abbas de Sétif, Sétif 19000, Algérie.

Recherche opérationnelle/Operations Research, 0399-0559/99/04/$ 7.00

de probabilités tel que l'estimation des paramètres du modèle par l'approche classification (Scott et Symons [10], Schroeder [9], Celeux [3], Govaert [7]) conduisent à l'optimisation d'un critère numérique de classification, on obtient un éclairage nouveau de ce critère et de la métrique sous-jacente permettant de les justifier ou éventuellement de les rejeter; par exemple Celeux [3] a donné une signification au critère d'inertie interclasse, utilisé pour la classification d'individus décrits par des variables quantitatives, pour le modèle de mélange gaussien; il a apporté une interprétation en termes probabilistes pour le critère d'information utilisé pour la classification d'individus décrits par des variables qualitatives, pour le modèle des classes latentes. De même Bencheikh [1] a donné une interprétation au critère du Khi2 utilisé pour la classification de tableaux disjonctifs complets. Dans le même cadre, Bock [2] montre que les critères classiques d'information s'interprètent comme des vraisemblances classifiantes de modèles log-linéaires et Govaert [7] montre que le critère optimisé par la méthode MNDBIN pour les données binaires correspond à un mélange de loi de Bernoulli.

Les liens qui existent entre les méthodes de classification automatique et les modèles de statistiques inférentielles ont surtout été étudiés lorsque les données mettent en jeu un seul ensemble. Nous proposons ici de le faire lorsque les données mettent en jeu deux ensembles; c'est le cas des méthodes de classification croisées proposées par Govaert [6].

Nous rappelons dans le premier paragraphe, le principe général de la classification croisée et l'un des algorithmes qui lui est associé Govaert [6]. Dans le second paragraphe nous introduisons la notion de mélange croisé et nous posons le problème de ces mélanges. Nous montrons dans le paragraphe trois comment la classification croisée peut être vue comme une solution à un problème d'estimation de paramètres d'un modèle de mélange croisé. Nous nous sommes intéressés dans le paragraphe quatre à l'étude des liens qui existent entre le modèle de mélange croisé et le modèle de mélange simple et nous avons montré que ces liens sont tout à fait analogues à ceux qui existent entre les méthodes de classification croisée et les méthodes de classification simple.

Nous terminons cet article par une application des résultats obtenus dans le paragraphe quatre sur deux types de fonctions de densités. Nous montrons alors que lorsque le nombre de composant du mélange d'un échantillon coincide avec sa taille, on retrouve exactement le modèle de la classification simple de Celeux [3] et Govaert [7].

1. LA CLASSIFICATION CROISÉE

1.1. Rappels et notations

Les données sont fournies sous la forme d'un tableau rectangulaire X de dimension (n, p) ou n est le nombre d'individus et p est le nombre de variables décrivant les n individus.

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

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

-- $\mathbf{P_k}$, l'ensemble des partitions de $\mathbf{I}$ en K classes, les éléments de $\mathbf{P_k}$ seront appelés K-partitions et notés $P = (P_1, ..., P_K)$.

-- $\mathbf{Q^m}$, l'ensemble des partitions de $\mathbf{J}$ en M classes, les éléments de $\mathbf{Q^m}$ seront appelés M-partitions et notés $Q = (Q^1, ..., Q^M)$.

-- $\mathbf{L}$, l'ensemble des noyaux, ces noyaux seront associés aux partitions des deux ensembles $\mathbf{I}$ et $\mathbf{J}$.

1.2. Le principe de la classification croisée

La position générale du problème de la classification croisée consiste à subdiviser la population des individus et la population des variables en un petit nombre de groupes ou classes homogènes dans un certain sens. Nous ne ferons pas de distinction entre les objets et les variables à classer en raison de l'identité de la position des problèmes et de la principale méthode d'étude; le problème à résoudre en classification croisée est alors le suivant: essayer de trouver une partition P de l'ensemble $\mathbf{I}$ des individus en K classes et une partition Q de l'ensemble $\mathbf{J}$ en M classes, tels qu'en réordonnant les lignes et les colonnes suivant les deux partitions, on obtient des classes homogènes. La recherche de ce meilleur couple de partitions (P, Q) correspond à l'optimisation d'un certain critère noté $\mathbf{W}(P, Q)$.

1.3. L'algorithme

Plusieurs algorithmes de classification croisée existent, on a retenu l'algorithme suivant développé par Govaert [6]; celui-ci utilise deux algorithmes voisins l'un de l'autre et tout deux basés sur le principe des Nuées Dynamiques (Diday [5]).

Cet algorithme se déroule en trois étapes:

(0) Partition initiale P° et Q° (quelconque).
$n = 0$

(1) On associe le noyau L^n minimisant $\mathbf{W}(P^n \times Q^n, L^n)$.

(2) Q^n et L^n fixés, P^{n+1} est la partition minimisant $\mathbf{W}(P^{n+1} \times Q^n, L^n)$.

(3) P^{n+1} et L^n fixés, Q^{n+1} est la partition minimisant $\mathbf{W}(P^{n+1} \times Q^{n+1}, L^n)$;
si $P^n = P^{n+1}$ et $Q^n = Q^{n+1}$ alors fin;
sinon $n = n + 1$ et aller en 1.

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 $K.M$ noyaux minimisant le critère: $\mathbf{W}((P \times Q), g(P \times Q))$.

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

La fonction d'affectation f

Cette fonction permet de déterminer, à Q et L fixés une partition P de l'ensemble $\mathbf{I}$ améliorant le critère $\mathbf{W}((P \times Q), L)$, en affectant chaque individu $i \in \mathbf{I}$ à la classe P_k du noyau de laquelle il est le plus proche.

La fonction d'affectation h

Cette fonction permet de déterminer, à P et L fixés, une partition Q de l'ensemble $\mathbf{J}$ améliorant le critère $\mathbf{W}(h(P \times Q), L)$, en affectant chaque individu $j \in \mathbf{J}$ à la classe Q^m du noyau de laquelle il est le plus proche.

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 [5]).

2. MODÈLE DE MÉLANGE « CROISÉ »

2.1. Exemple illustratif

Nous désignons par $\mathbf{P} = \{P_1, ..., P_n\}$ un ensemble de n produits que peut fabriquer un système, lequel est composé d'un ensemble $\mathbf{M} = \{m_1, ..., m_p\}$ de p machines; nous supposons que les échantillons $\mathbf{P}$ et $\mathbf{M}$ sont tous les deux formés respectivement de R et S sous-échantillons (appelés respectivement famille de produits et de machines).

Soit h une fonction permettant d'associer à chaque couple $(p_i, m_j) \in \mathbf{P} \times \mathbf{Q}$ une valeur $h(p_i, m_j)$ qui mesure la « ressemblance » entre les objets p_i et m_j. Par exemple si $h(p_i, m_j) \in \mathbf{R}$, $h(p_i, m_j)$ peut indiquer le nombre d'heures passées par la machine m_j pour la fabrication du produit p_i. Si maintenant $h(p_i, m_j) \in \{\mathbf{0}, \mathbf{1}\}$ la fonction h peut associer au couple (p_i, m_j) la valeur 1 si le produit p_i est fabriqué par la machine m_j, la valeur 0 sinon. La fonction h est définie sur l'ensemble $\mathbf{P} \times \mathbf{M}$ qui est le produit cartésien des deux ensembles $\mathbf{P}$ et $\mathbf{M}$. On remarque que l'ensemble $\mathbf{P} \times \mathbf{M}$ est lui aussi formé de $R.S$ sous-échantillons (appelés famille de produits-machines).

Supposons maintenant qu'un produit p_i soit issu d'un certain composant r (noté $\mathbf{CP}r$). De même, on suppose connu le composant s à laquelle la machine m_j est issue (on note cette famille par $\mathbf{CM}s$), la fonction h définie précédemment suit alors la densité de probabilité $f(h/\lambda_r^s)$ où λ_r^s est le paramètre caractérisant la densité de probabilité du (r.s)$^{\text{ème}}$ composant du mélange auquel le couple (p_i, m_j) appartient. Le modèle correspondant à cet exemple s'écrit alors :

$$f(h) = \sum_{r=1}^{R} \sum_{s=1}^{S} p_r^s f(h/\lambda_r^s) \tag{1}$$

$f(h/\lambda_r^s)$: représente la densité du (r.s)$^{\text{ème}}$ composant du mélange.

$f(h)$: représente la loi de probabilité résultante.

p_r^s : probabilité d'appartenance *a priori* à chacun des composants.

Si on reprend l'exemple précédent où $h(p_i, m_i) \in \mathbf{R}$, $f(h/\lambda_r^s)$ peut représenter la densité d'une loi gaussienne unidimensionnelle, le paramètre λ_r^s s'écrit :

$$\lambda_r^s = (\mu_r^s, \sigma_r^s)$$

où μ_r^s : espérance du composant (r.s),

σ_r^s : l'écart-type du composant (r.s).

Posons $V_r^s = (\sigma_r^s)^2$: variance du composant (r.s).

Le modèle (1) s'écrit alors :

$$f(h) = \sum_{r=1}^{R} \sum_{s=1}^{S} p_r^s (2\pi \cdot V_r^s)^{-\frac{1}{2}} \cdot \exp\left\{ -\frac{(h - \mu_r^s)^2}{2 \cdot V_r^s} \right\}. \tag{2}$$

Il s'agit de résoudre un problème classique d'estimation des paramètres. Nous avons retenu la méthode d'estimation du maximum de vraisemblance qui permet d'estimer les nombres R et S de composants du mélange et des paramètres inconnus :

$$(q_r^s = (p_r^s, (\mu_r^s, V_r^s))\,;\ r = 1, \ldots, R \text{ et } s = 1, \ldots, S)$$

au vu de l'échantillon $\mathbf{P} \times \mathbf{M}$.

Après résolution du problème on obtient :

$$\mu_r^s = \frac{1}{n_r \cdot q_s} \sum_{p_i \in CPr} \sum_{m_j \in CMs} h(p_i, m_j)$$

$$V_r^s = \frac{1}{n_r \cdot q_s} \sum_{p_i \in CPr} \sum_{m_j \in CMs} (h(p_i, m_j) - \mu_r^s)^2$$

$$p_r^s = \frac{n_r \cdot q_s}{n \cdot p}$$

μ_r^s : est la moyenne de la distribution $f(h/\lambda_r^s)$, V_r^s représente sa variance.

p_r^s : représente la probabilité d'apparition du composant $r.s$ dans le mélange.

n_r : représente le nombre de produits constituant le composant $\mathbf{CM}r$.

n_s : représente le nombre de machines constituant le composant $\mathbf{CM}s$.

2.2. Modèle général

L'exemple que nous venons de voir peut nous conduire à considérer que les tableaux de données habituels (contingence, binaire, qualitatif, quantitatif) qui sont tous décrits par deux ensembles que l'on note par $\mathbf{I}$ et $\mathbf{J}$ peuvent être replacés dans le même cadre que celui défini pour l'exemple, c'est-à-dire que l'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.

Dans la suite de ce travail, on suppose que l'ensemble $\mathbf{I}$ constitue un échantillon de taille $\mathbf{n}$ d'une population Ω, de même on suppose que l'ensemble $\mathbf{J}$ 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 $(n.p)$ d'une population $\Omega \times \Omega'$.

2.2.1. *Identification d'un mélange « croisé »*

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

$$f(x) = \sum_{k=1}^{K} \sum_{m=1}^{M} p_k^m f(x/\lambda_k^m) \tag{3}$$

$$\forall\, x \in \mathbf{R} \quad \forall\, k = 1, \ldots, K \qquad \text{et} \qquad m = 1, \ldots, M$$

$$0 \leq p_k^m \leq 1 \qquad \text{et} \qquad \sum_{k=1}^{K} \sum_{m=1}^{M} p_k^m = 1.$$

La formule (3) 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 et p_k^m est la probabilité d'apparition de l'observation $f(\,\cdot,\, \lambda_k^m)$ dans le mélange.

2.2.2. *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 *:*

$$(q_k^m = (p_k^m, \lambda_k^m)\,;\ k = 1, \ldots, K \text{ et } m = 1, \ldots, M)$$

au vu de l'échantillon $\mathbf{T} = \mathbf{I} \times \mathbf{J}$.

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

3. APPROCHE CLASSIFICATION

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

Problème 2

Rechercher une partition $P \times Q = \{P_k \times Q^m\,;\ k = 1, \ldots, M$ *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 [8] et la représentation de Celeux [3] 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 de critère de vraisemblance classifiante suivante :

$$\mathbf{VC}(P \times Q, L) = \sum_{k=1}^{K} \sum_{m=1}^{M} \operatorname{Log} L(P_k \times Q^m, \lambda_k^m) \tag{4}$$

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

On peut alors écrire :

$$L(P_k \times Q^m, \lambda_k^m) = \prod_{x \in P_k \times Q^m} f(x/\lambda_k^m). \tag{5}$$

Enfin le critère de **vraisemblance classifiante** s'écrit :

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

On peut remarquer que la résolution du problème (6) correspond exactement à la résolution d'un problème de classification croisé (Govaert [6]) rappelé au début de cet article. 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, et tout deux basés sur le principe des Nuées Dynamiques.

3.1. 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 $K.M$ 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.M noyaux minimisant le critère $\mathbf{VC}(P \times Q, g(P \times Q))$. On peut facilement voir que ces noyaux sont les

estimateurs du maximum de vraisemblance des paramètres associés aux sous-échantillons $\{P_k \times Q^m,\ k = 1, \ldots, K \text{ et } m = 1, \ldots, M\}$.

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

Une fonction d'affectation h 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 :

$$\mathbf{VC}(P \times Q, L) = \sum_{k=1}^{K} \sum_{i \in P_k} \operatorname{Log} F(x_i/\lambda_k) \tag{7}$$

où

$$F(x_i/\lambda_k) = \prod_{m=1}^{M} \left(\prod_{j \in Q^m} f(x_i^j/\lambda_k^m) \right). \tag{8}$$

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:

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

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:

$$\mathbf{VC}(P \times Q, L) = \sum_{m=1}^{M} \sum_{j \in Q^m} \operatorname{Log} F(x^j/\lambda^m) \tag{9}$$

où

$$F(x^j/\lambda^m) = \prod_{k=1}^{K} \left(\prod_{i \in P_k} f(x_i^j/\lambda_k^m) \right). \tag{10}$$

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

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

On peut montrer que sous certaines hypothèses (Govaert [6]), 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 $P_k \times Q^m$.

Nous proposons maintenant de développer sur l'échantillon $\mathbf{T} = \mathbf{I} \times \mathbf{J}$, l'approche mélange simple et de voir sous quelles conditions cette approche et l'approche mélange « croisé » coincident.

3.2. Position intermédiaire

Nous supposons toujours que l'échantillon $\mathbf{T} = \mathbf{I} \times \mathbf{J}$ est le produit cartésien des deux ensembles $\mathbf{I}$ et $\mathbf{J}$ de tailles respectives n et p. L'échantillon $\mathbf{T}$ de taille $(n.p)$ provient d'une variable aléatoire à valeurs dans R, dont la loi de probabilité admet cette fois-ci la fonction de densité suivante:

$$f(x) = \sum_{h=1}^{H} p_h f(x/\lambda_h)$$

avec

$$\forall\, h = 1, \ldots, H \qquad p_h \in]0,1[\qquad \text{et} \qquad \sum_{h=1}^{H} p_h = 1$$

H: nombre de composants du mélange T.

$f(\cdot, \lambda)$: appartient à une famille de fonctions de densité dépendant du paramètre λ, élément de $\mathbf{R}$, et p_h est la probabilité qu'un point de l'échantillon suive la loi $f(\cdot, \lambda_h)$. On appellera ces p_h les proportions du mélange.

Problème 1′

Le problème posé est l'estimation du nombre H de composants du mélange et des paramètres inconnus $\{p_h, \lambda_h / h = 1, \ldots, H\}$ au vu de l'échantillon.

Dans l'approche classification (Scott et Symons [10], Schroeder [8]), on remplace le problème 1′ d'estimation par le problème 2′ suivant:

Problème 2′

Rechercher une partition $R = (R_1, \ldots, R_H)$. H étant supposé connu, telle que chaque classe R_h soit assimilable à un sous-échantillon qui suit une loi $f(\cdot, \lambda_h)$.

Il s'agit alors de maximiser le critère de **vraisemblance classifiante** :

$$\mathbf{VC}(R, L) = \sum_{h=1}^{H} \operatorname{Log} L(R_h/\lambda_h)$$

où L est le H-uplet $(\lambda_1, \ldots, \lambda_H)$ et $L(R_h, \lambda_h)$ est la vraisemblance du sous-échantillon R_h suivant la loi $f(\cdot, \lambda_h)$: $L(R_h/\lambda_h) = \prod_{x \in R_h} f(x/\lambda_h)$.

Pour maximiser ce critère, on utilise la méthode des Nuées Dynamiques Diday [5] qui construit à partir d'une partition R^0 en H classes une suite de partitions-noyaux en minimisant à chaque étape le critère $\mathbf{VC}(R, L)$. On obtient à la convergence une partition R de l'échantillon $\mathbf{T}$ et une estimation des paramètres p_h ; cette partition ne correspond à aucune partition sur l'ensemble $\mathbf{I}$ ni sur l'ensemble $\mathbf{J}$. Ce problème ne correspond pas à un problème de classification croisée, là est la différence entre le modèle de mélange « croisé » proposé dans le paragraphe 2 et le modèle de mélange simple que l'on vient de développer. Ces deux modèles sont définis sur le même ensemble $\mathbf{T}$ formé de couples d'éléments. La différence essentielle entre ces deux approches vient du fait que si l'on a une partition P de l'ensemble $\mathbf{I}$ et une partition Q de l'ensemble $\mathbf{J}$, on a automatiquement une partition $P \times Q$ de l'ensemble $\mathbf{T} = \mathbf{I} \times \mathbf{J}$, mais la réciproque n'est pas vraie. Pour cette dernière raison on est amené à poser une condition sur la recherche de la partition R en H classes posé dans le problème $2'$. Cette condition s'exprime par la recherche séparée de deux partitions P et Q correspondant respectivement aux deux ensembles de départ $\mathbf{I}$ et $\mathbf{J}$ en K et M classes.

On peut alors écrire $R = P \times Q$ et $H = K.M$. Le problème $2'$ est donc remplacé par le problème 2. Ainsi nous nous retrouvons dans le cas de l'identification d'un mélange « croisé », et nous pouvons utiliser le même algorithme que celui développé dans le paragraphe 3.1.

Nous venons de voir comment la méthode de classification croisée (Govaert [6]) peut être vue comme une solution à un problème d'estimation de paramètres d'un modèle de mélange « croisé ». Ce problème est à rapprocher de celui de la méthode des Nuées Dynamiques (Diday [5]) qui a été utilisée par Schroeder [8] pour proposer une solution à un problème d'estimation de paramètres d'un modèle de mélange simple. Mais il serait intéressant dans la suite de ce travail d'établir s'il y en a, des liens entre le problème de mélange « croisé » et le problème de mélange simple ; nous savons que Govaert [6] s'est intéressé aux rapports qui existent entre les méthodes de classification croisée et les méthodes de classification simple et a remarqué que la comparaison des partitions obtenues par ces deux types de méthodes

était très délicate. Néanmoins l'algorithme qu'il a proposé lui permettait de passer d'un problème de classification croisée à un problème de classification simple. Nous allons essayer de suivre le même chemin que celui de Govaert [6], pour transformer le problème de mélange « croisé » en un problème de mélange simple. Nous verrons alors comment s'interprète le passage d'un problème de classification croisée à un problème de classification simple en termes de modèle probabiliste.

4. LIEN ENTRE LE MODÈLE DE MÉLANGE « CROISÉ » ET LE MODÈLE DE MÉLANGE SIMPLE

Pour éviter d'introduire de nouvelles notations pour indiquer les composants du mélange $\mathbf{I} \times \mathbf{J}$, qui risqueraient de rendre difficile la compréhension de ce paragraphe, nous avons tenu à garder la notation $P_k \times Q^m$ pour indiquer le composant du mélange ayant pour fonction de densité la loi $f(\cdot, \lambda_k^m)$. Rappelons que cette notation a été utilisée dans le paragraphe trois pour désigner une classe. Dans ce paragraphe, l'écriture $P_k \times Q^m$ indique plutôt le composant $k.m$ du mélange $\mathbf{I} \times \mathbf{J}$. De même P_k et Q_m indiquent respectivement les composants k du mélange $\mathbf{I}$ et m du mélange $\mathbf{J}$.

Rappelons rapidement le modèle associé à un mélange « croisé », $\mathbf{Z}$ étant une variable aléatoire définie sur le produit cartésien $\mathbf{I} \times \mathbf{J}$, dont la loi de probabilité f admet la fonction de densité suivante :

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

où encore $\forall\, (i,j) \in \mathbf{I} \times \mathbf{J}$

$$f(\mathbf{Z}(i,j)) = \sum_{k=1}^{K} \sum_{m=1}^{M} p_k^m f(\mathbf{Z}(i,j)/\lambda_k^m). \tag{11}$$

La valeur $\mathbf{Z}(i,j)$ ne peut provenir que de l'un des $K.M$ composants suivants $P_k \times Q^m$ pour $k = 1, \ldots, K$ et $m = 1, \ldots, M$; les p_k^m sont les proportions du mélange :

$$p_k^m = \frac{n_k \cdot q_m}{n \cdot p} = p(P_k \times Q^m).$$

n_k : représente le nombre d'éléments $i \in \mathbf{I}$ constituant le composant P_k,

q_m : représente le nombre d'éléments $j \in \mathbf{J}$ constituant le composant Q^m.

i) Supposons que l'on connaisse le composant $Q^{m(j)}$ auquel un élément $j \in \mathbf{J}$ appartient. Que devient alors le modèle donné par la formule (11) ?

Dans ce cas la valeur $\mathbf{Z}(i,j)$, sachant que $Q^{m(j)}$, ne peut provenir que de l'un des K composants suivants $\{P_k \times Q^m\}$ pour $k = 1, \ldots, K$. Le modèle (11) s'écrit alors :

$$f(\mathbf{Z}(i,j)/Q^{m(j)}) = \sum_{k=1}^{K} p_k^{m(j)} f(\mathbf{Z}(i,j)/\lambda_k^{m(j)}) \tag{12}$$

avec $p_k^{m(j)} = \frac{n_k \cdot q_{m(j)}}{n \cdot q_{m(j)}} = \frac{n_k}{n} = p_k$.

ii) On suppose maintenant que l'on connaît tout l'échantillon $\mathbf{J}$ de Ω' de taille p. Le composant dont chaque élément $j \in \mathbf{J}$ est issu est connu. Notons par $Q^{m(j)}$ ce composant. On peut définir une variable aléatoire ζ de dimension p telle que :

$$\zeta : \mathbf{I} \to \mathbf{R^p}$$
$$i \to \zeta(i) = (\zeta^1(i), \ldots, \zeta^p(i)) = (x_i^1, \ldots, x_i^p) = x_i.$$

On peut alors écrire $(\zeta^1(i), \ldots, \zeta^p(i)) = (\mathbf{Z}(i,1), \ldots, \mathbf{Z}(i,p))$.

Comme

$$f(\mathbf{Z}(i,j)/Q^{m(j)}) = \sum_{k=1}^{K} p_k^{m(j)} f(\mathbf{Z}(i,j)/\lambda_k^{m(j)}) \qquad \forall\, j = 1, \ldots, p$$

alors

$$f(x_i/Q) = f((\mathbf{Z}(i,1), \ldots, \mathbf{Z}(i,p))/(\lambda_k^{m(1)}, \ldots, \lambda_k^{m(p)}))$$
$$f(x_i/Q) = \sum_{k=1}^{K} p_k f((\mathbf{Z}(i,1), \ldots, \mathbf{Z}(i,p))/(\lambda_k^{m(1)}, \ldots, \lambda_k^{m(p)}))$$
$$= \sum_{k=1}^{K} p_k f(x_i/(\lambda_k, Q)). \tag{13}$$

En supposons qu'à l'intérieur de chaque classe, les p variables aléatoires $(\zeta^1, \ldots, \zeta^p)$ sont mutuellement indépendantes. On peut écrire :

$$f(x_i/(\lambda_k, Q)) = \prod_{j=1}^{p} f(x_i^j/\lambda_k^{m(j)}). \tag{14}$$

On considère deux cas:

Premier cas

Tous les composants auxquels les éléments j de l'échantillon **J** appartiennent sont distincts, dans ce cas, le nombre de composants du mélange **J** coïncide avec la taille de l'échantillon. On obtient:

$$\lambda_k^{m(j)} \neq \lambda_k^{m(j')} \qquad \forall\, j \neq j'.$$

On peut alors écrire:

$$\lambda_k^{m(j)} = \lambda_k^j \qquad \forall\, j = 1, \ldots, p.$$

La formule (13) s'écrit:

$$f(x_i) = \sum_{k=1}^{K} p_k f(x_i/\lambda_k) \tag{15}$$

et la formule (14) devient:

$$f(x_i/\lambda_k) = \prod_{j=1}^{p} f(x_i^j/\lambda_k^j). \tag{16}$$

Deuxième cas

Passons maintenant au cas le plus général où l'on suppose que plusieurs éléments de l'échantillon **J** peuvent provenir d'un même composant; soit M le nombre de composants dans lesquels sont répartis tout les éléments de l'échantillon **J**. On suppose toujours connu le composant auquel appartient chaque élément j de l'échantillon **J** que l'on note par $Q^{m(j)}$.

Posons:

$$\lambda_k^{m(j)} = \lambda_k^m \qquad \forall\, j \in Q^{m(j)} = Q^m.$$

Si on remplace maintenant dans l'expression (13) on obtient:

$$f(x_i/(\lambda_k, Q)) = \prod_{m=1}^{M} \prod_{j \in Q^m} f(x_i^j/\lambda_k^m). \tag{17}$$

Remarque

Que l'on soit dans le premier cas ou le deuxième cas, on se retrouve toujours avec une expression du modèle qui correspond à celle d'un mélange simple (13) et (15). Pour approfondir l'étude que l'on vient de faire, on se propose de l'appliquer à deux types de fonctions de densités.

5. APPLICATIONS

Dans ce paragraphe, nous nous proposons de voir ce que deviennent les expressions données par les formules (16) et (17) en les appliquant à deux types de familles de fonctions de densité ; on suppose dans tout ce paragraphe qu'on se trouve dans la situation (ii).

5.1. Lois gaussiennes unidimensionnelles

Premier cas

$$F(x/\lambda_k) = \prod_{j=1}^{p} f\left(x^j/\lambda_k^j\right) \qquad \text{avec} \quad \lambda_k^j = (\mu_k^j, V_k^j)$$

$$F(x/\lambda_k) = \prod_{j=1}^{p} \left(2\pi V_k^j\right)^{-\frac{1}{2}} \cdot \exp\left[-\frac{(x^j - \mu_k^j)^2}{2 \cdot V_k^j}\right]$$

$$F(x/\lambda_k) = (2\pi)^{-\frac{p}{2}} \cdot |V_k|^{-\frac{1}{2}} \cdot \exp\left[\frac{-t(x-\mu_k) \cdot V_k^{-1} \cdot (x-\mu_k)}{2}\right]$$

$\lambda_k = (\mu_k, V_k)$

μ_k : espérance du composant numéro k.

V_k : matrice de variance du composant numéro k.

$F(x/\lambda_k)$: représente la densité d'une loi gaussienne multidimensionnelle. On retrouve ainsi le modèle de mélange gaussien proposé par Celeux [3] pour la classification simple de tableaux décrits par des données quantitatives.

Deuxième cas

$$f(x/\lambda_k) = \prod_{m=1}^{M} \prod_{j \in Q^m} f(x^j/\lambda_k^m)$$

$$= \prod_{m=1}^{M} \prod_{j \in Q^m} (2\pi V_k^m)^{-\frac{1}{2}} \cdot \exp\left[-\frac{(x^j - \mu_k^m)^2}{2 \cdot V_k^m}\right]$$

$f(x/\lambda_k)$ représente aussi la densité d'une loi gaussienne multidimensionnelle. Par analogie au premier cas, on peut déduire que ce modèle correspond lui aussi à la classification simple de données quantitatives, dont les éléments $j \in \mathbf{J}$ sont répartis en M classes.

5.2. Lois de Bernoulli

La famille analysée est définie par:

$$p(x/\lambda) = \varepsilon^{|x-\lambda|} \cdot (1-\varepsilon)^{1-|x-\lambda|}$$

où $\varepsilon \in]0, \frac{1}{2}[$; $x \in \{0,1\}$ et $\lambda \in \{0,1\}$.
$p(x/\lambda)$ désigne la distribution d'une loi de Bernoulli de paramètre $(1-\varepsilon)$ ou de paramètre ε

Premier cas

$$p(x/\lambda_k) = \prod_{j=1}^{p} p(x^j/\lambda_k^j) = \prod_{j=1}^{p} \left\{ \varepsilon^{|x^j-\lambda_k^j|} \cdot (1-\varepsilon)^{1-|x^j-\lambda_k^j|} \right\}$$

$p(x/\lambda_k)$ représente alors le produit de p lois de Bernoulli. Ce modèle correspond exactement à celui proposé par Govaert [7] pour la classification simple de tableaux binaires.

Deuxième cas

$$p(x/\lambda_k) = \prod_{m=1}^{M} \prod_{j \in Q^m} p(x^j/\lambda_k^m)$$

$$= \prod_{m=1}^{M} \left\{ \varepsilon^{|x^m - q_m \cdot \lambda_k^m|} \cdot (1-\varepsilon)^{q_m - |x^m - q_m \cdot \lambda_k^m|} \right\}.$$

où $x^m = \sum_{j \in Q^m} x^j$ est le cardinal de Q^m.

Il est facile de vérifier que la distribution $p(x/\lambda_k)$ ainsi trouvée correspond au modèle de mélange simple associé à la classification croisée de données binaires en supposant connue et fixe la partition Q en colonnes de l'ensemble $\mathbf{J}$.

CONCLUSION

L'application faite ci-dessus nous conduit à conclure que lorsqu'on se retrouve dans le premier cas, on retrouve le même modèle que celui qui a été proposé pour la classification simple (Celeux [3] et Govaert [7]). Si maintenant on se trouve dans le deuxième cas, on montre que le modèle de mélange trouvé correspond aux méthodes de classification croisée en supposant connue une partition. Ainsi, la connaissance d'une

partition d'un ensemble s'interprète en termes de modèle par la connaissance des composants d'un échantillon, et nous venons de montrer par les deux applications faites ci-dessus que le lien qui existe entre les méthodes de classification simple et les modèles de mélange simple est le même que celui qui existe entre les méthodes de classification croisée en supposant connue une partition et les modèles de mélange croisé en supposant connus les composants d'un échantillon. De plus le lien que nous venons d'établir entre les méthodes de classification croisée et les modèles probabilistes nous permettent d'apporter un éclairage nouveau des méthodes de classification croisée très usitées, de justifier de manière rigoureuse des constatations faites de manière empirique et peut être encore plus de proposer de nouveaux critères pouvant améliorer la qualité de la partition.

RÉFÉRENCES

1. Y. Bencheikh, Classification automatique et modèles, Thèse Université de Metz, 1992.
2. H. H. Bock, Loglinear models and entropy clustering methods for qualitative data, Classification as Tool of Research, W. Gaul and M. Schader, Eds., 1986.
3. G. Celeux, Classification et modèles, *Rev. Statist. Appl.*, 1988, *36*, n°4, p. 43-58.
4. G. Celeux et G. Govaert, Clustering criteria for discrete data and latent class models, Rapport n° 1122 INRIA, 1989.
5. E. Diday, Nouvelles méthodes et nouveaux concepts en classification automatique et reconnaissance des formes, Thèse d'État, Université Paris 6, 1972.
6. G. Govaert, Classification croisée, Thèse de Doctorat d'État, Université Pierre et Marie Curie Paris 6, 1983.
7. G. Govaert, Classification binaire et modèle, Rapport de Recherche INRIA, n° 949, 1988
8. A. Schroeder, Reconnaissance des composants d'un mélange, Thèse de Doctorat de 3ème cycle. Université Paris 6, 1974.
9. A. Schroeder, Analyse d'un mélange de distributions de probabilité de même type, *R.S.A*, 1976, *24*, n° 1, p. 39-62.
10. A. Scott et Symons, Clustering methods based on likelihood ratio criteria, Biometrics, 1971, *27*, p. 387-397.