Modèle de classification et distance dans le cas continue

Gérard Govaert

INRIA

UNITÉ DE RECHERCHE
INRIA-LORRAINE

Rapports de Recherche

N° 988

Programme 5

MODELE DE CLASSIFICATION ET DISTANCE DANS LE CAS CONTINUE

Gérard GOVAERT

Mars 1989

Institut National
de Recherche
en Informatique
et en Automatique

Domaine de Voluceau
Rocquencourt
BP 105
78153 Le Chesnay Cedex
France
Tél. (1) 39 63 55 11

MODELE DE CLASSIFICATION ET DISTANCE DANS LE CAS CONTINUE

CLUSTERING MODEL AND DISTANCE WITH CONTINUOUS DATA

Gérard GOVAERT

Université de Metz, Ile du Saulcy 57045 Metz
INRIA-Lorraine, BP 239 54506 Vandoeuvre les Nancy Cedex

Résumé

Les méthodes de classification se ramènent souvent à l'optimisation d'un critère numérique défini à partir d'une distance. Dans certain cas, il est possible de montrer que cela revient à estimer les paramètres d'un modèle probabiliste par une approche classification. Ainsi, il est bien connu que le critère d'inertie, très souvent utilisé en classification, correspond à l'hypothèse d'une population issue d'un mélange de lois gaussiennes.
Dans ce travail, nous étudions les liens qui existent entre ces deux approches lorsque les variables sont quantitatives. Pour ceci, nous définissons la notion de critère métrique et de critère probabiliste, nous montrons ensuite qu'un critère probabiliste peut toujours être considéré comme un critère métrique et établissons enfin les conditions pour que la réciproque soit vraie. Ces résultats sont alors appliqués à deux familles de critères métriques : les premiers sont définis à partir des distances quadratiques, les seconds, à partir de la distance L_1. Cette approche permet de préciser en particulier les différences entre la méthode des distances adaptatives et la méthode de reconnaissance de mélange dans le cas gaussien et de montrer que les critères utilisant la distance en valeur absolue correspondent à un mélange de lois exponentielles bilatérales.

Mots-clés : Classification, mélange de lois de probabilité, distance quadratique, distance L_1 .

Abstract

Clustering methods often come down to the optimization of a numeric criterion defined from a distance. It is possible to show that this problem is often equivalent to the estimation of a probabilistic model by a clustering approach. For instance, we know that the inertia criterion corresponds to the hypothesis of a population arising from a Gaussian mixture.
Here, we study the links between these two approaches when the data are quantitative. First, we show that a probabilistic criterion can always be considered as a metric criterion and then we establish that, under some conditions, the reciprocal property is true. Finally, these results are applied to the quadratic and to the city-block metrics. With this approach, we can explicit the differences between the adaptive distances method and the Gaussian mixture problem, and we can show that the criterion defined with the city block metric is associated to a bilateral exponential distributions mixture.

Keywords : Clustering, mixture model, quadratic distance, L_1 distance.

INTRODUCTION

De nombreuses méthodes de classification automatique se ramènent à l'optimisation d'un critère numérique lui-même souvent souvent défini à partir d'une distance. C'est le cas en particulier de toutes les méthodes des Nuées Dynamiques. Le choix de ce critère et de cette distance est alors la première et principale difficulté à résoudre.

Lorsqu'il est possible de trouver un modèle de mélange de lois de probabilité tel que l'estimation des paramètres du modèle par l'approche classification (Scott et Symons 1971, Schroeder 1976, Celeux 1988) conduise 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 montre que le critère d'inertie habituel correspond à l'hypothèse d'une population issue d'un mélange de lois gaussiennes ayant toutes la même matrice de variances de la forme a.I où a est un réel et I la matrice identité. De la même façon, on peut montrer (Govaert 1989) que le critère optimisé par l'algorithme MNDBIN pour les données binaires correspond à un mélange issu de lois de Bernoulli.

On s'intéresse dans ce travail aux tableaux de variables quantitatives (données continues). Dans les deux premiers paragraphes, nous définissons deux types de critères de classification et nous étudions dans quelles conditions ces critères peuvent être équivalents. Dans le troisième paragraphe, nous nous intéressons aux liens qui existent entre les lois de Gauss et les distances quadratiques. Nous expliquons en particulier les différences entre la méthode des distances adaptatives (Govaert 1975, Diday et Govaert 1977, Celeux 1988) et la méthode de reconnaissance de mélange dans le cas gaussien (Schroeder 1976). Enfin, dans le dernier paragraphe, on étudie les critères utilisant la distance en valeur absolue et on montre par exemple qu'un critère très simple défini à partir de cette distance correspond à un mélange de lois exponentielles bilatérales définies dans $\mathbf{R}^p$.

1. DEFINITION DES DEUX TYPES DE CRITERES

On suppose dans tout ce travail que les données initiales sont fournies sous la forme d'un tableau de n lignes et p colonnes contenant les valeurs prises par n individus pour p variables **quantitatives**. Nous envisageons ici deux types de critères : le premier, que nous appellerons **critère métrique**, utilise la notion de mesure de dissimilarité ; le second, que nous appellerons **critère probabiliste**, utilise la notion de mélange probabiliste. Nous définissons tout d'abord ces deux types de critères.

1.1. CRITERE METRIQUE

1.1.1. Définition

Dans cette approche, on représente le tableau de données sous la forme d'un ensemble Ω de n individus de $\mathbf{R}^p$. Chaque classe d'une partition va être représentée par un élément d'un ensemble L qui reste à préciser et qui sera appeler ensemble des "noyaux". Enfin, on se donne une fonction D de $\mathbf{R}^p \times$ L dans $\mathbf{R}^+$ qui mesurera la "dissimilarité" entre un élément de $\mathbf{R}^p$ et un noyau.
Le problème que l'on cherche à résoudre est de trouver la partition $P=(P_1,...,P_K)$ de Ω en K classes et un K-uple $(\lambda_1,...,\lambda_K)$ de noyaux (un par classe) minimisant le critère

$$\sum_{k=1}^{K} \sum_{x \in P_k} D(x,\lambda_k).$$

Ce critère qui dépend de la mesure de dissimilarité D sera appelé **critère métrique** et noté $\mathbf{CP(R^p,L,D)}$.

Remarque :
La méthode des Nuées Dynamiques (Diday et al 1980) propose une solution à ce problème en construisant, de manière itérative, une suite de couples partitions-noyaux faisant décroître le critère. Pour ceci, elle utilise les deux fonctions suivantes :

- une fonction de représentation g définie par $g(P) = g(P_1,...P_K) = (\lambda_1,...\lambda_K)$ où λ_k est

  l'élément de L minimisant $\sum_{x \in Pk} D(x,.)$

- une fonction d'affectation h définie par $h(\lambda) = h(\lambda_1,...\lambda_K) = (P_1,...P_K)$ où $P_k = \{x \in \Omega / D(x,\lambda_k) \leq D(x,\lambda_m)$ avec k<m en cas d'égalité$\}$.

On montre que l'algorithme ainsi défini converge sous certaines hypothèses et fournit à la convergence un optimum local du critère.

1.1.2. Critères métriques équivalents

Définition :
On dira que deux critères métriques sont équivalents si et seulement s'ils sont définis sur les mêmes ensembles $\boldsymbol{R}^p$ *et L et s'il existe une bijection* φ *de* $\boldsymbol{R}^p$ *strictement croissante vérifiant*

$$CM(\boldsymbol{R}^p,L,D_1) = \varphi \circ CM(\boldsymbol{R}^p,L,D_2)$$

où D_1 *et* D_2 *sont les mesures de dissimilarité associées aux deux critères.*

Remarque :
Toutes les solutions optimales correspondant à des critères équivalents sont identiques. De plus, des algorithmes de recherche d'optima locaux, comme l'algorithme des Nuées Dynamiques, fourniront pour des critères équivalents les mêmes résultats.

On peut facilement montrer que si l'on remplace D par une fonction linéaire croissante de D, on obtient un critère métrique équivalent :

Proposition 1:
$\forall\ \alpha \in \boldsymbol{R}^+$ *et* $\beta \in \boldsymbol{R}$, *les critères* $CM(\boldsymbol{R}^p,L,D)$ *et* $CM(\boldsymbol{R}^p,L,\alpha D+\beta)$ *sont équivalents.*

1.1.3. Contrainte sur les noyaux

On peut sans difficulté et en conservant le même critère modifier le problème posé en ajoutant une contrainte au k-uple de noyaux $(\lambda_1,...,\lambda_K)$ recherché. Par exemple, si le noyau est défini comme un couple (a,b), on peut imposer que le second terme du couple soit identique pour tous les noyaux du k-uple recherché :

$$\lambda = ((a,b_1),...,(a,b_K)).$$

1.2. CRITERE PROBABILISTE

1.2.1. Définition

On reprend ici la présentation de Celeux (1988).

Le tableau de données de départ de dimension (n,p) est maintenant considéré comme un échantillon Ω de taille n d'une variable aléatoire à valeurs dans $\mathbf{R}^p$ dont la loi de probabilité admet la densité

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

avec

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

où $f(.,\lambda)$ appartient à une famille **F** de fonctions de densité dépendant du paramètre λ et p_k est la probabilité qu'un point de l'échantillon suive la loi $f(.,\lambda_k)$. On notera L l'ensemble des paramètres.

Le problème posé est l'estimation du nombre K de composantes et des paramètres inconnus $\{p_k,\lambda_k / k=1,K\}$ au vu de l'échantillon.

Dans l'approche classification (Scott et Symons 1971, Schroeder 1976), on remplace le problème initial d'estimation par le problème suivant :

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

Il s'agit alors de maximiser un **critère de vraisemblance classifiante**

$$W(P,\lambda) = \sum_{k=1}^{K} \text{Log } L(P_k,\lambda_k)$$

où λ est le p-uplet $(\lambda_1,...\lambda_k)$ et $L(P_k,\lambda_k)$ est la vraisemblance du sous-échantillon P_k suivant la loi $f(.,\lambda_k)$.

Ce critère qui dépend de la famille **F** de fonctions de densité définies sur $\mathbf{R}^p$ sera appelé **critère probabiliste** et noté $\mathbf{CP(R^p,F)}$.

Remarque :
Comme pour le critère métrique, on peut maximiser ce critère en utilisant l'algorithme des Nuées Dynamiques qui construit à partir d'une partition P^o en K classes une suite de partitions en appliquant les deux fonctions suivantes :

- une fonction de représentation g définie par $g(P) = g(P_1,...P_K) = (\lambda_1,...\lambda_K)$ où λ_k est l'estimation du **maximum de vraisemblance** du paramètre de la densité associée au sous-échantillon P_k.

- une fonction d'affectation h définie par $h(\lambda) = h(\lambda_1,...\lambda_K) = (P_1,...P_K)$ où $P_k=\{x\in \Omega / f(x,\lambda_k)\geq f(x,\lambda_m)$ avec $k<m$ en cas d'égalité$\}$

On peut alors montrer que sous certaines hypothèses, cet algorithme est convergent. On obtient à la convergence une partition P et une estimation des paramètres λ_k. Les proportions p_k du mélange sont fournies par les fréquences des classes P_k.

1.2.2. Contraintes sur les paramètres

Nous avons vu précédemment qu'il était possible de modifier le problème en imposant des contraintes aux noyaux sans changer de critère métrique. De la même façon, nous pouvons dans le cas d'un critère probabiliste modifier le problème en imposant une contrainte aux paramètres des fonctions de densité associées aux classes d'une partition.

Par exemple, si la famille **F** est l'ensemble des lois gaussiennes sur $\mathbf{R}^p$, on peut imposer que toutes les lois gaussiennes associées aux classes d'une partition aient la même matrice de variances.

2. LIENS ENTRE LES DEUX TYPES DE CRITERES

Après avoir défini les deux types de critères, nous étudions maintenant comment ces deux notions peuvent se rejoindre. Pour ceci, nous définissons deux types de liens. Tout d'abord, nous montrons qu'à un critère probabiliste peut être associé de façon canonique un critère métrique que nous appelons **critère métrique associé** au critère probabiliste. La notion de critères équivalents, définis dans le cas des critères métriques, peut alors être étendue à l'ensemble des critères métriques et probabilistes. Ceci permet de définir le second type de liens. Nous étudions ensuite le problème suivant : un critère métrique donné est-il associé à un critère probabiliste? Cette propriété n'est pas vraie en général, mais nous établissons une condition nécessaire et suffisante pour qu'elle soit vérifiée. Enfin, nous montrons qu'il suffit d'une condition plus faible pour qu'un critère métrique donné soit simplement équivalent, et non plus associé, à un critère probabiliste.

2.1. CRITERE METRIQUE ASSOCIE A UN CRITERE PROBABILISTE

Proposition 2 :

$$CP(\mathbf{R}^p,F) = - CM(\mathbf{R}^p,L,D)$$

où L est l'ensemble de définition des paramètres de la famille F et D est définie par

$$\forall x \in \mathbf{R}^p,\ \forall \lambda \in L \qquad D(x,\lambda) = - Log\, f(x,\lambda).$$

Pour démontrer cette proposition, il suffit d'utiliser la définition des deux critères.

On appellera le critère métrique ainsi défini à partir d'un critère probabiliste **critère métrique associé.**

La maximisation d'un critère probabiliste est donc la même chose que la minimisation du critère métrique associé. Ce résultat permet donc de considérer que tous les critères probabilistes sont des critères métriques.

2.2. CRITERES PROBABILISTES ET METRIQUES EQUIVALENTS

Puisque tout critère probabiliste peut être considéré comme un critère métrique, on peut étendre la définition de critères métriques équivalents :

Définitions :

Deux critères probabilistes sont équivalents si leurs critères métriques associés sont équivalents.

Un critère probabiliste CP_1 et un critère métrique CM_2 sont équivalents si le critère métrique CM_1 associé à CP_1 est équivalent au critère métrique CM_2.

2.3. CONDITIONS POUR QU'UN CRITERE METRIQUE SOIT ASSOCIE A UN CRITERE PROBABILISTE

Nous venons de voir qu'à tout critère probabiliste pouvait être associé un critère métrique. Nous nous intéressons maintenant au problème réciproque. Cette propriété n'est pas vraie en général mais on peut établir la proposition suivante :

Proposition 3 :

Un critère métrique $CM(\mathbf{R}^p,L,D)$ est associé à un critère probabiliste si et seulement si $\forall \lambda \in L$ *la fonction* $x \rightarrow e^{-D(x,\lambda)}$ *est continue et vérifie* $\int_{\mathbf{R}^p} e^{-D(x,\lambda)}\, dx = 1.$

2.4. CRITERE PROBABILISTE EQUIVALENT A UN CRITERE METRIQUE

En utilisant la proposition 1, on peut obtenir une condition plus faible permettant de montrer qu'un critère métrique est équivalent (et non associé) à un critère probabiliste.

Proposition 4 :

Etant donné le critère métrique $CM(\mathbf{R}^p,L,D)$ *, s'il existe un réel* $r>1$ *tel que la quantité* $s = \int_{\mathbf{R}^p} r^{-D(x,\lambda)}\,dx$ *soit indépendante de* λ*, alors le critère probabiliste* $CP(\mathbf{R}^p,F)$ *où F est définie par les fonctions de densité f :*

$$f(x,\lambda) = \frac{1}{s} r^{-D(x,\lambda)}$$

est un critère équivalent.

Preuve :

En effet, le critère métrique associé à la famille proposée est définie par la fonction D' :

$$D'(x,\lambda) = -\,\mathrm{Log}\, f(x,\lambda) = -\,\mathrm{Log}\{\frac{1}{s}\, r^{-D(x,\lambda)}\} = s + r.D(x,\lambda)$$

La proposition 1 permet d'affirmer que les critères métriques associés à D et D' sont équivalents. D'où le résultat annoncé.

3. METRIQUES QUADRATIQUES ET LOIS GAUSSIENNES

Dans ce paragraphe, nous supposons toujours que l'ensemble à classifier Ω est inclus dans $\mathbf{R}^p$ et nous étudions un certain nombre de critères métriques tous issus de distances quadratiques définies sur $\mathbf{R}^p$ et, à chaque fois nous montrons que, sous certaines conditions, ces critères sont associés ou équivalents à des critères probabilistes où les lois de probabilité sont toujours gaussiennes. Nous retrouvons bien sûr des résultats connus, mais en outre cette approche permet d'aller plus loin et d'expliquer, par exemple, quels sont les liens qui existent entre la distance des distances adaptatives (Govaert 1975 et 1977) et la reconnaissance de mélange de lois gaussiennes (Schroeder 1976) qui conduisent à des algorithmes très voisins.

3.1. DISTANCE QUADRATIQUE FIXE ET IDENTIQUE POUR TOUTES LES CLASSES

On suppose ici que les noyaux sont des éléments de $\mathbf{R}^p$, c'est-à-dire que $L=\mathbf{R}^p$, et que la fonction D est définie à partir d'une matrice M définie symétrique positive fixée a priori

$$\forall x \text{ et } \lambda_\kappa \in \mathbf{R}^p \quad D(x,\lambda_\kappa) = \alpha.{}^t(x-\lambda_\kappa).M.(x-\lambda_\kappa) + \beta \qquad \text{où } \alpha\in\mathbf{R}^+ \text{ et } \beta\in\mathbf{R}.$$

Quelles que soient les valeurs α et β, les critères seront tous équivalents (proposition 1). On se limitera donc au critère le plus simple qui correspond à la fonction D :

$$\forall\, x \text{ et } \lambda \in \mathbf{R}^p \quad D(x,\lambda_\kappa) = {}^t(x-\lambda_\kappa).M.(x-\lambda_\kappa)$$

On peut montrer facilement que $\int_{\mathbf{R}^p} e^{-D(x,\lambda)} = \Pi^{p/2}.(\det M)^{-1/2}$. La proposition 4 permet donc de déduire que ce critère métrique est équivalent au critère probabiliste dont la famille F est définie par

$$\forall\ x \text{ et } \lambda_k \in \mathbf{R}^p \quad f(x,\lambda_k) = \Pi^{-p/2}.(\det M)^{1/2}.e^{-{}^t(x-\lambda_k).M.(x-\lambda_k)}$$

ce qui correspond à une loi gaussienne de centre λ_k et de matrice de variance $2.M^{-1}$.

Remarquons que l'on obtient un critère équivalent en prenant simplement M^{-1}. En effet, nous avons vu que l'on pouvait multiplier la matrice M par une constante positive et que l'on obtenait des critères métriques équivalents. Les critères probabilistes associés possèdent donc la même propriété : tous les critères probabilistes ayant comme fonctions de densité les lois gaussiennes de matrice de variances $\alpha.M$ où $\alpha>0$ sont équivalentes.

On retrouve, en prenant comme matrice M la matrice identité, que la minimisation du critère d'inertie revient à considérer que toutes les classes sont issues de lois gaussiennes sphériques ayant toutes la même matrice de variances.

3.2. DISTANCE QUADRATIQUE VARIABLE ET DEPENDANT DE CHAQUE CLASSE

On suppose maintenant que la métrique n'est plus fixée et qu'elle dépend de chaque classe. Nous étudions deux variantes.

3.2.1. Première variante

Pour ceci, nous prenons comme noyaux des couples (a,M) où a est un élément de $\mathbf{R}^p$ et M une matrice symétrique définie positive

$$D(x,(a_k,M_k)) = {}^t(x-a_k).M_k.(x-a_k)$$

Pour que ce critère soit associé à un critère probabiliste, on doit donc avoir (proposition 3) :

$$\int_{\mathbf{R}^p} e^{-D(x,(a_k,M_k))}.dx = 1$$

$$\pi^{p/2}.(\det M_k)^{-1/2} = 1$$

$$\det M_k = \pi^p$$

On retrouve exactement la méthode des distances adaptatives (Govaert 1975, Celeux 1988). Le seul changement porte sur la valeur de la constante que l'on avait fixée à 1. On obtient ainsi une justification a posteriori de cette méthode et en particulier du choix de la contrainte que l'on avait alors choisi arbitrairement : fixer le déterminant de la matrice . Elle correspond à la reconnaissance d'un mélange de lois gaussiennes dont les matrices de variances sont de déterminant constant.

3.2.2. Seconde variante

On suppose cette fois que $D(x,(a_k,M_k,\alpha_\kappa)) = \alpha_k + {}^t(x-a_k).M_k.(x-a_k).$

Pour que ce critère soit associé à un critère probabiliste, on doit donc avoir

$$\int_{R^p} e^{-D(x,(a_k,M_k,\alpha_k))}.dx = 1$$

$$e^{-\alpha_k} \int_{R^p} e^{-{}^t(x-a_k).M_k.(x-a_k)}\, dx = 1$$

$$e^{-\alpha_k}.\pi^{p/2}.(\det M)^{-1/2} = 1$$

$$\alpha_k = \frac{p}{2}.\text{Log}\,\pi - \frac{1}{2}.\text{Log}\,(\det M_k)$$

Il faut donc que α_κ dépende de M_k. Cela revient à prendre les mêmes noyaux que dans la première variante. On a donc

$$D(x,(a_k,M_k)) = \frac{p}{2}\,\text{Log}\,\pi - \frac{1}{2}\text{Log}\,\det(M_k)) + {}^t(x-a_k).M_k.(g-a_k)$$

L'introduction de la constante α_k a permis de supprimer la contrainte que l'on avait dans le premier cas. Cette fois, on n'impose plus au déterminant de la matrice M_k d'être constant pour que ce critère corresponde à un critère probabiliste. Le critère probabiliste correspondant est défini par les fonctions de densité :

$$f(x,(a_k,\Gamma_\kappa)) = \frac{1}{(2\Pi)^{p/2}.(\det \Gamma_\kappa)^{1/2}}\, e^{-\frac{1}{2}{}^t(x-g_k).\Gamma_k^{-1}.(x-g_k)}$$

où $\Gamma_k = 1/2\,.\,M_k^{-1}$.

Cette situation correspond au cas le plus général des lois gaussiennes. On sait alors (estimation du maximum de vraisemblance, Schroeder 1976) que les a_k sont les centres de gravité des classes et les Γ_k sont les $V_k = \frac{W_k}{\text{Card } P_k}$ avec $W_k = \sum_{x\in P_k} (x-g_k).{}^t(x-g_k)$

et on retrouve alors, à une relation linéaire près, la distance utilisée pour classer les points :

$$D(x,a_k) = \frac{p}{2} \text{Log}\, \pi - \frac{1}{2}\text{Log}\, \det\left(\frac{1}{2}.V_k^{-1}\right) + \frac{1}{2}\ {}^t(x-g_k).V_k^{-1}.(g-g_k)$$

La comparaison des résultats obtenus dans ce paragraphe permet de mieux comprendre les liens qui existent entre l'algorithme des Distances Adaptatives et celui de la reconnaissance de mélange gaussien. Le premier est un cas particulier du second : il correspond à l'introduction d'une contrainte sur le déterminant des matrices de variances qui doivent tous être constants. La méthode des Distances Adaptatives correspond donc elle aussi à un modèle de reconnaissance de mélange gaussien.

3.3. DISTANCE EUCLIDIENNE IDENTIQUE POUR TOUTES LES CLASSES

On suppose maintenant que la métrique, qui n'est toujours pas fixée, reste identique pour toutes les classes. Nous étudions, comme dans la paragraphe précédent, deux variantes.

3.3.1. Première variante

Pour ceci, nous prenons comme noyaux des couples (a,M) où a est un élément de $\mathbf{R}^p$ et M une matrice symétrique définie positive et nous imposons une contrainte supplémentaire : tous les noyaux doivent avoir la même matrice M.

$$D(x,(a_k,M)) = {}^t(x-a_k).M.(x-a_k)$$

Pour que ce critère soit associé à un critère probabiliste, on doit donc avoir

$$\int_{\mathbf{R}^p} e^{-D(x,(a_k,M))}.dx = 1$$

$$\pi^{p/2}.(\det M)^{-1/2} = 1$$

$$\det M = \pi^p$$

On retrouve une contrainte équivalente à celle que nous avions imposée dans la méthode des distances adaptatives avec distance unique (Diday, Govaert et Lemoine 1978). La fonction de densité correspondante est alors

$$f(x,a_k) = e^{-1/2\ {}^t(x-g_k).\Gamma^{-1}.(x-g_k)} \quad \text{en posant } \Gamma = 1/2\ M^{-1}$$

Ceci correspond bien à la loi de Gauss de matrice de variance Γ puisque $(2\pi)^{p/2}.(\det \Gamma)^{1/2} = 1$.

3.3.2. Seconde variante

Cette fois, on suppose $D(x,(a_k,M,\alpha)) = \alpha + {}^t(x-a_k).M.(x-a_k)$.

Pour que ce critère soit associé à un critère probabiliste, on doit donc avoir

$$\int_{R^p} e^{-D(x,(a_k,M,\alpha))} .dx = 1$$

$$e^{-\alpha}.\pi^{p/2}.(\det M)^{-1/2} = 1$$

$$\alpha = \frac{p}{2} \text{Log } \pi - \frac{1}{2} \text{Log det (M)}$$

Il faut donc que α dépende de M. Cela revient à prendre les mêmes noyaux que dans la première variante mais cette fois avec la fonction D suivante :

$$D(x,(a_k,M)) = \frac{p}{2} \text{Log } \pi - \frac{1}{2} \text{Log det (M)} + {}^t(x-a_k).M.(g-a_k)$$

Cette modification de la fonction D permet de supprimer la contrainte que l'on avait dans le premier cas. Cette fois, on n'impose plus au déterminant de M d'être constant pour que ce critère corresponde à un critère probabiliste. Le critère probabiliste correspondant est défini par les fonctions de densité :

$$f(x,(a_k,\Gamma)) = \frac{1}{(2\Pi)^{p/2}.(\det \Gamma)^{1/2}} e^{-1/2\ {}^t(x-a_k).\Gamma^{-1}.(x-a_k)}$$

où $\Gamma = 1/2\ M^{-1}$

Cette situation correspond au modèle avec des lois gaussiennes dont on impose que toutes les matrices de variance soient identiques (Celeux 1988). On peut montrer que les deux dernières hypothèses conduisent au même algorithme et sont donc équivalentes. L'ajout du terme additif α n'a donc eu aucun effet, ce qui n'avait pas été le cas de celui du terme additif α_k dans les hypothèses précédentes.

4. METRIQUES DE TYPE L_1

On peut remplacer dans le critère habituel de l'inertie la distance euclidienne au carré par la distance L_1 ou distance city-block. On montre alors que toutes les propriétés de convergence sont conservées et que le centre de gravité est remplacé par la notion de médiane.
Dans ce paragraphe, nous montrons que cette approche correspond elle aussi à un modèle de mélange de lois de probabilité. Nous étendons ces propriétés aux distances L_1 pondérées, ces pondérations pouvant être variables ou non suivant les classes. Ces distances avaient été proposées dans (Govaert 1975).

4.1. DISTANCE FIXE ET IDENTIQUE POUR TOUTES LES CLASSES

On suppose ici que les noyaux sont des éléments de $\mathbf{R}^p$, c'est-à-dire que $L=\mathbf{R}^p$, et que

$$D(x,\lambda_k) = \sum_{j=1}^{p} \alpha^j \, | x^j - \lambda_k^j |$$

où les α^j sont des constantes réelles positives.

Pour que ce critère métrique soit associé à un critère probabiliste, on doit vérifier

$$\int_{\mathbf{R}^p} e^{-D(x,\lambda_k)} .dx = 1$$

$$\int_{\mathbf{R}^p} \exp\left(-\sum_{j=1}^{p} \alpha^j . \, | x^j - \lambda_k^j | \right).dx = 1$$

$$\int_{\mathbf{R}^p} \prod_{j=1}^{p} \exp(-\alpha^j . \, | x^j - \lambda_k^j |).dx = 1$$

$$\prod_{j=1}^{p} \int_{-\infty}^{+\infty} \exp(-\alpha^j . \, | x^j - \lambda_k^j |).dx = 1$$

$$\prod_{j=1}^{p} \frac{2}{\alpha^j} = 1$$

$$\prod_{j=1}^{p} \alpha^j = 2^p$$

Il suffit donc d'imposer cette contrainte aux coefficients pour que le critère métrique défini avec la fonction

$$D(x,\lambda_k) = \sum_{j=1}^{p} \alpha^j . \, | x^j - \lambda_k^j |$$

soit associé à un critère probabiliste. La fonction de densité correspondante est

$$f(x,\lambda_k) = e^{-D(x,\lambda_k)}$$

$$= \prod_{j=1}^{p} \exp(-\alpha^j . \, | x^j - \lambda_k^j |)$$

$$= \prod_{j=1}^{p} \frac{\alpha^j}{2} \exp(-\alpha^j \, | x^j - \lambda_k^j |) \quad (\text{car } \prod_{j=1}^{p} \alpha^j = 2^p)$$

Cela revient à considérer que, pour chaque composant du mélange, les p variables sont indépendantes et que chacune d'elles est issues d'une **loi exponentielle bilatérale** $L(\lambda_k^j,\alpha^j)$.

Si les coefficients α^j ne vérifient pas la contrainte, on peut s'y ramener en divisant la fonction D par $1/2.(\prod_{j=1}^{p}\alpha^j)^{1/p}$. La proposition 1 permettant d'affirmer que les critères métriques obtenus avant et après la division sont équivalents, on peut en déduire que tous les critères définis à partir le la fonction

$$D(x,\lambda_k) = \sum_{j=1}^{p} \alpha^j |x^j - m_k^j|$$

sont équivalents à des critères probabilistes issus le lois exponentielles bilatérales. Par exemple, on peut examiner le cas particulier du critère métrique associé à la fonction

$$D(x,\lambda_k) = \sum_{j=1}^{p} |x^j - m_k^j| .$$

Nous avons ici $\prod_{j=1}^{p}\alpha^j = 1$. On prend donc le critère métrique équivalent suivant

$$D(x,\lambda_k) = 1/2. \sum_{j=1}^{p} |x^j - m_k^j| ,$$

puisque dans ce cas $1/2.(\prod_{j=1}^{p}\alpha^j)^{1/p} = 1/2.$

Ce critère métrique est le critère associé au critère probabiliste dont la fonction de densité est :

$$f(x,\lambda_k) = \prod_{j=1}^{p} 1/2 \,.\, \exp(-|x^j - \lambda_k^j|) .$$

Le critère métrique défini à partir de la distance L_1 habituelle revient donc à considérer chaque classe comme un échantillon d'un produit de p lois exponentielles bilatérales $L(\lambda_k^j,1)$.

4.2. DISTANCE L_1 VARIABLE ET DEPENDANT DE CHAQUE CLASSE

On suppose maintenant que la métrique n'est plus fixée et qu'elle dépend de chaque classe. Elle a la forme suivante :

$$D(x,(a_k,\alpha_k,\beta_k)) = \sum_{j=1}^{p} \alpha_k^j |x^j - a_k^j|) + \beta_\kappa$$

Pour que ce critère soit associé à un critère probabiliste, on doit vérifier

$$\int_{R^p} e^{-D(x,(ak,\alpha k,\beta k))} .dx = 1$$

$$e^{-\beta_k} \int_{R^p} \exp(- \sum_{j=1}^{p} \alpha_k^j \mid x^j - a_k^j \mid) \, dx = 1$$

$$e^{-\beta_k} \int_{R^p} \prod_{j=1}^{p} \exp(-\alpha_k^j \mid x^j - a_k^j \mid).dx = 1$$

$$e^{-\beta_k} \prod_{j=1}^{p} \int_{-\infty}^{+\infty} \exp(-\alpha_k^j \mid x^j - a_k^j \mid).dx = 1$$

$$e^{-\beta_k} \prod_{j=1}^{p} \frac{2}{\alpha_k^j} = 1$$

$$\beta_k = p.\mathrm{Log}2 - \sum_{j=1}^{p} \mathrm{Log}\, \alpha_k^j$$

D'où

$$D(x,(a_k,\alpha_k)) = p.\mathrm{Log}2 - \sum_{j=1}^{p} \mathrm{Log}\, \alpha_k^j + \sum_{j=1}^{p} \alpha_k^j \mid x^j - m_k^j \mid$$

Si on veut retrouver uniquement la distance L_1 pondérée, il faut imposer que la première partie de l'expression s'annule, c'est-à-dire $\sum_j \mathrm{Log}\beta_k^j = p \ \mathrm{Log}\ 2$ ou encore $\prod_j \beta_k^j = 2^p$.

On retrouve alors la méthode des distances adaptatives dans le cas L_1 (Govaert 1975) dans laquelle on imposait la même contrainte. On obtient ainsi une justification a posteriori de cette contrainte.

Si on n'impose pas cette contrainte, on aura alors comme dans le cas de la distance quadratique, une différence au niveau de la fonction d'affectation puisqu'on ajoutera un terme dépendant de la classe à cette distance.

CONCLUSION

Nous avons donc vu que la comparaison des critères métriques et probabilistes permettait d'apporter un éclairage nouveau de nombreuses méthodes de classification, de justifier a posteriori certaines contraintes imposées souvent pour des raisons techniques d'optimisation, de proposer de nouveaux critères, mais peut-être encore plus, cette comparaison permet d'expliquer l'intérêt et la souplesse de la Méthode de Nuées Dynamiques dont l'idée essentielle était l'utilisation de la notion de noyau associé

à chaque classe. Ce noyau correspond tout naturellement avec le critère probabiliste aux paramètres de la loi de probabilité associé à chaque classe. Il resterait à étudier si les autres critères définis dans la cadre de la Méthode des Nuées Dynamiques correspondent eux-aussi à des critères probabilistes.

Bibliographie

CELEUX G., DIDAY E. ,GOVAERT G., LECHEVALLIER Y., RALAMBON-DRAINY H., Classification automatique des données : Environnement statistique et informatique. Dunod 1989.

DIDAY E. et COLLABORATEURS. (1980), "Optimisation en Classification Automatique". INRIA. Rocquencourt .

DIDAY E. et GOVAERT G. (1977), "Classification avec distances adaptatives". RAIRO, V-11, n°4, pp. 329-349.

DIDAY E., GOVAERT G.et LEMOINE Y. (1978), "A new kind of representation in clustering". 4ème IJCPR. Kyoto. 1978.

CELEUX G. (1988), "Classification et modèles". R.S.A. Vol 36, n°3.

GOVAERT G. (1975), "Classification Adaptative". Thèse de 3ème cycle, Paris 6.

GOVAERT G. (1989), "Classification binaire et Modèle", Rapport de recherche INRIA.

SCHROEDER A. (1976), "Analyse d'un mélange de distribution de probabilité de même type". RSA, Vol 24. n°1.

SCOTT A., SYMONS M. (1971), "Clustering methods based on likelihood ratio criteria". Biometrics 27.

Imprimé en France
par
l'Institut National de Recherche en Informatique et en Automatique