Un algorithme de classification automatique non paramétrique
A non parametric clustering algorithm[*]

Thierry Denœux et Gérard Govaert

Université de Technologie de Compiègne
U.R.A. CNRS 817 Heudiasyc
BP 529 F-60205 Compiègne cedex - France
Tél : 44 23 44 96 ; Télécopie : 44 23 44 77

Résumé : Nous présentons un nouvel algorithme de classification automatique reposant sur la recherche d'une partition stable pour une méthode de discrimination non paramétrique. Un lien formel entre cette approche et le modèle connexionniste de Hopfield est souligné, ce qui permet de mettre en évidence une fonction d'énergie dont la décroissance stricte à chaque itération garantit la stationnarité de la suite de partitions obtenue. Le lien entre cette méthode et le principe de maximisation de la vraisemblance classifiante est également étudié.

Abstract : *We present a new clustering algorithm based on the search for a stable partition with respect to a non parametric discrimination rule. A formal link between this method and the Hopfield connectionist model is demonstrated, allowing for the definition of an energy function whose decrease at each iteration guarantees the stationarity of the sequence of partitions obtained. A link between this approach and the principle of classification likelihood maximisation is also studied.*

Rubrique : Statistique.

[*]In *Comptes-Rendus de l'Académie des Sciences de Paris*, t. 324, Série I, 673–678, 1997.

1 Introduction

La classification automatique et la discrimination sont deux domaines voisins qui entretiennent des liens très étroits. Rappelons que la discrimination, ou classification en mode supervisé, a pour objectif de déterminer une fonction de décision permettant d'affecter, à partir d'un ensemble d'apprentissage, des individus à des classes connues. En revanche, la classification automatique, encore appelée classification en mode non supervisé, cherche à mettre en évidence une structure de classes dans un ensemble de données initialement non étiquetées.

La discrimination étant un domaine largement étudié, en particulier d'un point de vue statistique, il est intéressant de pouvoir déduire une méthode de classification d'une méthode de discrimination. Une approche, très naturelle mais peu utilisée, consiste à chercher une partition stable vis à vis d'une fonction de décision F, c'est-à-dire une partition telle que la classe de chaque individu ne change pas lorsqu'on lui applique la fonction F. Un exemple bien connnu d'application de cette démarche dans un cadre paramétrique est la méthode de classification des centres mobiles, qui correspond à la discrimination linéaire habituelle. Un intérêt supplémentaire de cette approche est la possibilité de prendre en compte de façon simple des situations hybrides mais très courantes (McLachlan, 1992) dans lesquelles les 2 problèmes de classification et de discrimination existent.

Nous présentons dans ce travail un algorithme original de classification automatique défini en utilisant la démarche précédente à partir d'une méthode de discrimination non-paramétrique.

2 Méthode de discrimination non paramétrique

On considère que l'on dispose d'un ensemble $\mathcal{X}$ de n individus $\mathbf{x}_1, \ldots, \mathbf{x}_n$ de $\mathbb{R}^d$ rangés dans M classes $\{P_1, \ldots, P_M\}$ et on cherche à classer un nouvel individu $\mathbf{x}$. Pour ceci, on applique la règle optimale de Bayes en maximisant la probabilité a posteriori (on considère ici que les coûts d'affectation sont 0 ou 1 suivant que l'affectation est correcte ou non). Les distributions f_q de chaque classe étant inconnues, on les estime par une méthode classique d'estimation non paramétrique des fonctions de densité : la méthode des noyaux. On obtient alors $\hat{f}_q(\mathbf{x}) = \frac{1}{n_q} \sum_{\mathbf{y} \in P_q} K(\mathbf{x} - \mathbf{y})$ où n_q est la taille de la classe P_q et K une fonction noyau, c'est-à-dire une fonction de densité sur $\mathbb{R}^d$.

En estimant les probabilités a priori des classes par leurs proportions dans l'ensemble d'apprentissage, les estimations des probabilités a posteriori d'appartenance de $\mathbf{x}$ aux classes q sont

$$\hat{p}(P_q/\mathbf{x}) = \frac{\hat{p}(P_q)\hat{f}_q(\mathbf{x})}{f(\mathbf{x})} = \frac{\frac{n_q}{n}\frac{1}{n_q}\sum_{\mathbf{y} \in P_q} K(\mathbf{x} - \mathbf{y})}{f(\mathbf{x})} = \frac{\sum_{\mathbf{y} \in P_q} K(\mathbf{x} - \mathbf{y})}{nf(\mathbf{x})}.$$

avec $f(\mathbf{x}) = \sum_{r=1}^{M} \hat{p}(P_r)\hat{f}_r(\mathbf{x})$. La fonction de classement se définit alors de manière simple : l'individu $\mathbf{x}$ est rangé dans la classe q qui maximise la quantité $\sum_{\mathbf{y} \in P_q} K(\mathbf{x} - \mathbf{y})$.

3 L'algorithme de classification NPCLUS

On considère toujours l'ensemble $\mathcal{X}$ de n individus $\mathbf{x}_1, \ldots, \mathbf{x}_n$ de $\mathbb{R}^d$, mais cette fois c'est la partition supposée inconnue qui est recherchée. On supposera en outre que le nombre de classes est fixé à M.

3.1 Description

L'application du principe de recherche d'une partition stable vis à vis de l'algorithme de discrimination que nous avons rappelé dans le paragraphe précédent nous conduit à proposer l'algorithme suivant.

Il s'agit d'un algorithme itératif construisant une suite de partitions P^0, P^1, ..., P^l définie par une partition initiale aléatoire P_0 et la fonction de récurrence $P^{i+1} = \varphi(P^i)$ où φ est la procédure suivante :

- tirage au hasard d'un ordre sur $\mathcal{X}$;
- examen dans cet ordre de chaque individu $\mathbf{x}$ et rangement dans la classe q qui maximise $\sum_{\mathbf{y} \in P_q, \mathbf{y} \neq \mathbf{x}} K(\mathbf{x} - \mathbf{y})$.

3.2 Étude de la convergence

Notre démonstration de la convergence de l'algorithme précédent s'inspire d'une analogie formelle avec le modèle connexionniste introduit par Hopfield (Hopfield, 1982). Ce modèle consiste en un réseau totalement interconnecté de neurones à états binaires. A chaque pas de temps, un neurone i est choisi aléatoirement et son état V_i est réévalué en fonction des états des autres neurones selon la règle suivante :

$$\begin{array}{lllll} V_i & \leftarrow & 1 & \text{si} & \sum_{j \neq i} W_{ij} V_j > 0 \\ V_i & \leftarrow & 0 & \text{sinon} & \end{array}$$

W étant la matrice des poids des connexions supposée symétrique. On montre que chaque changement d'état d'un neurone fait décroître la fonction d'énergie $E = -\frac{1}{2} \sum_{i \neq j} W_{ij} V_i V_j$. Dans la procédure de discrimination évoquée au paragraphe précédent, les vecteurs d'apprentissage peuvent être assimilés aux neurones du modèle de Hopfield, le poids de la connexion entre les vecteurs $\mathbf{x}$ et $\mathbf{y}$ étant $K(\mathbf{x} - \mathbf{y})$. La principale différence réside dans le fait qu'il y a maintenant en chaque nœud du réseau M états possibles correspondant aux M classes. Avec nos notations, la fonction d'énergie se généralise de la manière suivante :

$$E(P) = -\frac{1}{2} \sum_{\mathbf{x},\mathbf{y}} K(\mathbf{x} - \mathbf{y}) < \mathbf{V}(\mathbf{x}), \mathbf{V}(\mathbf{y}) >= -\frac{1}{2} \sum_{q=1}^{M} \sum_{\mathbf{x} \in P_q} \sum_{\mathbf{y} \in P_q} K(\mathbf{x} - \mathbf{y}) \quad (1)$$

où $< \cdot, \cdot >$ désigne le produit scalaire, et $\mathbf{V}(\mathbf{x})$ est un vecteur à M composantes tel que $V_q(\mathbf{x}) = 1$ si $\mathbf{x} \in P_q$, et $V_q(\mathbf{x}) = 0$ sinon. Il s'agit maintenant de montrer qu'à chaque étape de l'algorithme une modification de la partition correspond nécessairement à une diminution du critère. Pour cela, supposons que l'individu $\mathbf{x}$ passe de la classe q à la classe p. La variation du critère est :

$$\Delta E = -\sum_{\mathbf{y} \in P_p, \mathbf{y} \neq \mathbf{x}} K(\mathbf{x} - \mathbf{y}) + \sum_{\mathbf{y} \in P_q, \mathbf{y} \neq \mathbf{x}} K(\mathbf{x} - \mathbf{y})$$

où P_p et P_q sont les classes p et q avant la modification.

Puisque $\mathbf{x}$ est passé de la classe q à la classe p, on a

$$\sum_{\mathbf{y}\in P_p,\mathbf{y}\neq\mathbf{x}} K(\mathbf{x}-\mathbf{y}) < \sum_{\mathbf{y}\in P_q,\mathbf{y}\neq\mathbf{x}} K(\mathbf{x}-\mathbf{y})$$

et donc $\Delta E < 0$, ce qui démontre la stationnarité de la suite (P^l) définie par l'algorithme.

4 Lien avec la vraisemblance classifiante

En classification automatique, l'approche par modèle de mélange permet d'analyser de manière précise et d'interpréter d'un point de vue statistique certains critères métriques (trace ou déterminant de la matrice d'inertie intra-classe, etc.) dont les différentes variantes n'étaient pas toujours très claires dans les approches purement géométriques. Elle permet en outre de se donner un cadre précis pour traiter des problèmes comme le nombre de classes, l'existence d'une structure de classification ou encore la validation des partitions obtenues. Dans cette perspective, une des approches les plus fructueuses s'appuie sur la maximisation de la vraisemblance classifiante. Avant d'étudier les liens entre cette approche et l'algorithme NPCLUS, nous allons tout d'abord rappeler la définition de la vraisemblance classifiante.

4.1 Vraisemblance classifiante

La vraisemblance classifiante se définit à partir de la notion de mélange de lois de probabilités (Titterington, 1985 ; McLachlan, 1988). Ayant un ensemble de n individus $\mathbf{x}_1, \ldots, \mathbf{x}_n$ décrits par d variables, c'est-à-dire un échantillon de taille n dans $\mathbb{R}^d$, le modèle de mélange correspond à l'hypothèse que la population de référence est formée de M sous-populations $P_1, \ldots, P_M$ de densités $f_q(\mathbf{x})$ avec des proportions $p_1, \ldots, p_M$ ($p_q \in]0,1[$ et $\sum_{q=1}^{M} p_q = 1$). Les n individus constituent donc un échantillon $\mathbf{x}_1, \ldots, \mathbf{x}_n$ de réalisations indépendantes d'un vecteur aléatoire $\mathbf{X}$ de $\mathbb{R}^d$ de densité

$$f(\mathbf{x}, \theta) = \sum_{q=1}^{M} p_q f_q(\mathbf{x})$$

où f_q appartient à une famille de densités sur $\mathbb{R}^d$ et $\theta = (p_1, \ldots, p_M, f_1, \ldots, f_M)$.

Le problème est alors de déterminer θ, c'est-à-dire les proportions et les fonctions de densité, ce qui revient souvent à la recherche des paramètres de ces fonctions. L'algorithme EM fournit en général une bonne solution à ce problème.

En classification automatique, on cherche non seulement les fonctions f_1, ..., f_q et les proportions $p_1, \ldots, p_M$, mais aussi et surtout la partition P. Divers solutions sont envisageables. Il est possible, par exemple, après avoir résolu le problème précédent, d'affecter les individus aux classes suivant les probabilités a posteriori estimées à partir des $f_1, \ldots, f_M$ et des proportions $p_1, \ldots, p_q$.

Une autre approche (Symons, 1981 ; Celeux, 1995) consiste à ajouter comme paramètre supplémentaire à estimer la partition P elle-même. La vraisemblance est alors appelée vraisemblance classifiante et son logarithme est défini de la manière suivante :

$$L_C(P,\theta) = \sum_{q=1}^{M} \sum_{\mathbf{x}\in P_q} \ln p_q f_q(\mathbf{x}).$$

Le problème de la classification automatique est alors la recherche simultanée de la partition P et de θ maximisant la vraisemblance classifiante.

4.2 Relation entre les critères E et L_C

Si on associe à chaque classe q, comme fonction de densité, l'estimation obtenue par la méthode des noyaux $\frac{1}{n_q}\sum_{\mathbf{y}\in P_q} K(\mathbf{x}-\mathbf{y})$, on obtient

$$L_C(P,\theta) = \sum_{q=1}^{M} \sum_{\mathbf{x}\in P_q} \ln p_q f_q(\mathbf{x}) = \sum_{q=1}^{M} \sum_{\mathbf{x}\in P_q} \ln(\sum_{\mathbf{y}\in P_q} K(\mathbf{x}-\mathbf{y})) - n\ln n$$

Cette expression est, au signe près, formellement assez proche du critère $E(P)$ précédent (Équation 1). Une relation approximative entre les deux critères peut être mise en évidence par le raisonnement suivant. Posons $S_q(\mathbf{x}) = \sum_{\mathbf{y}\in P_q} K(\mathbf{x}-\mathbf{y})$. On a donc

$$E = -\frac{1}{2}\sum_{q=1}^{M} \sum_{\mathbf{x}\in P_q} S_q(\mathbf{x}) \quad \text{et} \quad L_C = \sum_{q=1}^{M} \sum_{\mathbf{x}\in P_q} \ln S_q(\mathbf{x}) - n\ln n.$$

Soit S la moyenne des $S_q(\mathbf{x})$: $S = \frac{1}{n}\sum_{q=1}^{M}\sum_{\mathbf{x}\in P_q} S_q(\mathbf{x})$. On a évidemment $E = -\frac{nS}{2}$. Par ailleurs, en faisant un développement limité au premier ordre de $\ln S_q(\mathbf{x})$ au voisinage de S, on obtient :

$$\ln S_q(\mathbf{x}) \approx \ln S + \frac{S_q(\mathbf{x})}{S} - 1$$

En remplaçant $\ln S_q(\mathbf{x})$ par son approximation dans l'expression de L_C, il vient :

$$L_C \approx \sum_{q=1}^{M} \sum_{\mathbf{x}\in P_q} \left(\ln S + \frac{S_q(\mathbf{x})}{S} - 1\right) - n\ln n = n\ln\frac{S}{n} = n\ln\frac{-2E}{n^2}$$

Cette relation approximative entre les deux critères s'avère expérimentalement bien vérifiée (cf. infra) et permet d'interpréter l'algorithme NPCLUS comme une méthode approchée de maximisation de la vraisemblance classifiante.

5 Un exemple

5.1 Implémentation de l'algorithme

Le noyau retenu est le noyau gaussien : il est défini de la façon suivante

$$K(\mathbf{x}-\mathbf{y}) = \frac{1}{(2\pi h^2)^{\frac{d}{2}}} \exp\{-\frac{1}{2h^2}||\mathbf{x}-\mathbf{y}||^2\}$$

où $\|\cdot\|$ est la norme euclidienne et h un paramètre de lissage pouvant être déterminé automatiquement à partir des données selon la méthode proposée par Duin (Duin, 1976). La partition initiale peut être choisie arbitrairement. Afin d'éviter de fixer un nombre de classes a priori, il est possible de choisir initialement autant de classes que d'individus dans l'ensemble d'apprentissage, chaque classe étant composée d'un seul individu.

5.2 Simulations

Afin d'illustrer le comportement de l'algorithme sur un exemple simple, nous considérons un échantillon de taille 100 composé de trois classes gaussiennes bidimensionnelles (Figure 1a). La Figure 1b montre un exemple de partition obtenue. On constate que l'algorithme trouve pour cet exemple le bon nombre de classes, et produit une partition correcte bien que les classes ne soient pas très séparées. Nous avons pu constater une grande stabilité de ce résultat au cours d'un grand nombre d'exécutions de l'algorithme. L'évolution des deux critères E et L_C, ainsi que de l'approximation $\tilde{L}_C = n \ln \frac{-2E}{n^2}$ est représentée sur la Figure 2. Conformément aux calculs exposés précédemment, on remarque que le critère de vraisemblance classifiante tend à augmenter au cours de l'apprentissage, de même que son approximation en fonction du critère E.

6 Conclusion

Un nouvel algorithme non paramétrique de classification automatique a été proposé. Cet algorithme, dont la convergence a été démontrée, repose sur le principe de l'estimation des densités de probabilité des classes par la méthode des noyaux, et peut être interprété comme une méthode approchée de maximisation de la vraisemblance classifiante, dans le cas non paramétrique. Un avantage de la méthode réside dans sa simplicité de mise en œuvre. En effet, le paramètre de lissage h associé aux fonctions noyaux peut être fixé automatiquement par une méthode de maximum de vraisemblance (Duin, 1976), et le nombre de classes de la partition initiale peut être pris égal au nombre d'individus : le nombre de classes n'a donc pas à être fixé à l'avance, à l'encontre de la plupart des méthodes de classification automatique. De nombreuses simulations réalisées sur différents jeux de données semblent montrer un bon comportement de la méthode lorsque les classes sont relativement bien séparées. Une comparaison systématique avec d'autres procédures de classification telles que l'algorithme CEM reste à effectuer.

Références

[1] G. CELEUX et G. GOVAERT. Gaussian parsimonious clustering models. *Pattern Recognition*, 28(5):781–793, 1995.

[2] R. P. DUIN. On the choice of smoothing parameters for Parzen estimators of probability density functions. *IEEE Transactions on Computers*, C-25:1175–1179, 1976.

[3] J. J. HOPFIELD. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79:2554–2558, 1982.

[4] G. J. MCLACHLAN. *Discriminant Analysis and Statistical Pattern Recognition.* Wiley, New York, 1992.

[5] G.J. MCLACHLAN et K.E. BASFORD. *Mixture Models, Inference and applications to clustering.* Marcel Dekker, New York, 1988.

[6] M. J. SYMONS. Clustering criteria and multivariate normal mixtures. *Biometrics*, 37:35–43, 1981.

[7] D.M. TITTERINGTON, A. F. M. SMITH, et U. E. MAKOV. *Statistical Analysis of Finite Mixture Distributions.* Wiley, New York, 1985.

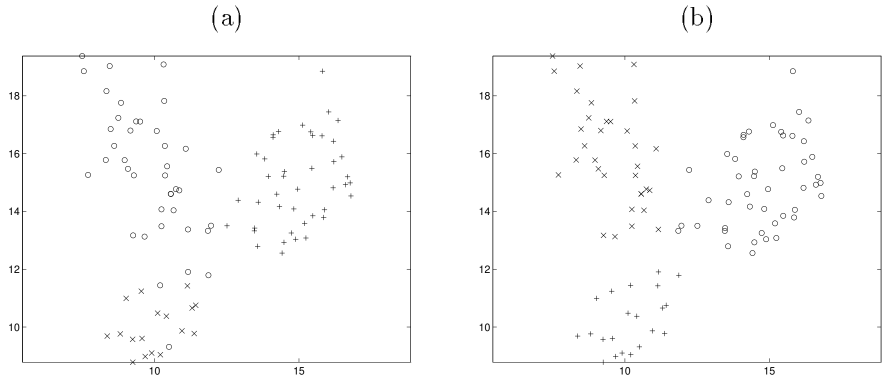

FIG. 1 - *Ensemble d'apprentissage (a) et partition obtenue (b).* (Learning set (a) and obtained partition (b).)

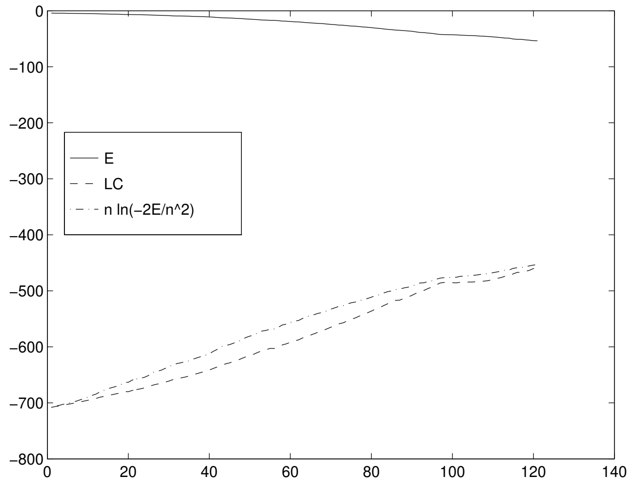

FIG. 2 - *Évolution au cours d'une exécution de l'algorithme des critères E, L_C et $\tilde{L}_C = n \ln \frac{-2E}{n^2}$.* (Evolution of criteria E, L_C et $\tilde{L}_C = n \ln \frac{-2E}{n^2}$ during one execution of the algorithm.)