Classification basée sur des mélanges de modèles hiérarchiques bivariés

Vera Georgescu, Nicolas Desassis, Samuel Soubeyrand, André Kretzschmar, Rachid Senoussi

Classification basée sur des mélanges de modèles hiérarchiques bivariés

Vera Georgescu*[1], Nicolas Desassis[2], Samuel Soubeyrand[1], André Kretzschmar[1], Rachid Senoussi[1]

[1]*INRA, UR546 Biostatistique et Processus Spatiaux, Agroparc, F-84914 Avignon, France*
[2]*École Nationale supérieure des Mines de Paris, Centre de Géosciences, F-77300 Fontainebleau, France*

Résumé

Les approches probabilistes basées sur les modèles de mélange sont de plus en plus utilisées en classification automatique car elles fournissent un cadre formel pour résoudre des problèmes pratiques qui se posent en classification, tels que la détermination du nombre de classes, et permettent d'estimer l'incertitude associée à la classification. Les limites de ces méthodes résident principalement dans le choix de la loi de probabilité des composantes du mélange, qui dépend du type de données et va contraindre la forme des classes. Peu de modèles de mélange ont été étudiés dans le cas multivarié, et il est difficile d'adapter la méthode d'estimation d'une distribution à une autre. Nous proposons une généralisation des méthodes de classification basées sur des modèles de mélange qui :

- s'adapte rapidement à des données de type différents (continues, discrètes, binaires, surdispersées),
- permet d'obtenir des formes de classe et des structures de corrélation variées,
- permet de traiter des données multivariées comportant des observations de plusieurs types.

Pour cela nous considérons un modèle hiérarchique, dans lequel la couche cachée est issue d'un mélange de lois gaussiennes bivariées et la couche d'observation est obtenue par une distribution bivariée dont le choix dépend du type de données observées. L'estimation du modèle se fait par un algorithme MCEM.

Mots clés : Données multivariées, Modèle hiérarchique, Modèle Poisson log-normal, Monte Carlo EM.

Abstract

Clustering methods based on finite mixture models have proved to be very efficient in cluster analysis, and to provide nice statistical solutions to several practical issues encountered in this field, such as the selection of the number of clusters. Some

*e-mail : vera.georgescu@avignon.inra.fr

of the problems that still remain in these methods are closely related to the choice of the component distributions, which has to be made according to the type of data. There is still a limited choice of distributions that have been studied for multivariate data, and the estimation of such models is difficult to adapt from one distribution to another. We present a generalized model-based clustering method for bivariate data, which is easy to adapt to different types of data, can accomodate very different shapes of clusters (and correlations) and can cluster data with attributes of different types. Our method is based on a hierarchical model in which the hidden layer is a bivariate normal mixture and defines the clusters and the second layer depends on the type of data. The estimation algorithm is based on Monte Carlo EM.

Key words: Hierarchical model, Model-based clustering, Monte Carlo EM, Multivariate data, Poisson log-normal model.

1 Introduction

Les approches probabilistes basées sur les modèles de mélange sont de plus en plus utilisées en classification automatique. Elles fournissent plusieurs avantages par rapport aux approches heuristiques basées sur des critères métriques, notamment un cadre formel pour résoudre des problèmes pratiques qui se posent en classification, tels que la détermination du nombre de classes. De plus elles permettent d'estimer l'incertitude associée à la classification. Dans la classification à l'aide de modèles de mélange on suppose que les observations à classer sont issues de plusieurs populations (groupes), chaque population étant caractérisée par une distribution de probabilité. Le choix de la distribution des composantes du mélange se fait en fonction du type de données.

Pour obtenir une partition des données avec cette approche, les paramètres du mélange sont estimés et les observations attribuées à la classe la plus probable conditionnellement à ces paramètres. Des méthodes d'estimation basées sur le maximum de vraisemblance ont été mises au point pour plusieurs distributions (voir par exemple McLachlan et Peel, 2000), et les logiciels de classification correspondants sont disponibles (e.g. MIXMOD - Biernacki *et al.*, 2006, MClust - Fraley et Raftery, 2006). Pour les observations multivariées, les modèles utilisés sont essentiellement le mélange gaussien multivarié pour les données quantitatives continues (Fraley et Raftery, 2002) et le mélange multinomial multivarié pour les données qualitatives (Agresti, 2002).

Les limites des méthodes de classification à l'aide de modèles de mélange résident principalement dans le choix de la distribution, qui dépend du type de données et va contraindre la forme des classes.

- Dans le cas multivarié, peu de modèles de mélange ont été étudiés (gaussien et Student multivarié pour les données continues, Poisson multivarié pour les données de comptage, multinomial multivarié pour les variables qualitatives), et il est difficile d'adapter l'estimation de ces modèles d'une distribution à une autre,

- Les formes des classes sont parfois trop limitées par le choix de la distribution (la loi de Poisson multivarié ne permet pas de corrélation négative au sein d'une classe),
- Il n'existe pas encore de méthode, à notre connaissance, qui permette de classer des observations multivariées de type différent (e.g. continu vs. discret).

Nous proposons une généralisation des méthodes de classification à l'aide de modèles de mélange qui :

- s'adapte rapidement à des données de type différents (continues, binaires, discrètes, surdispersées),
- permet d'obtenir des formes de classe et des structures de corrélation variées,
- permet de traiter des données multivariées comportant des observations de plusieurs types.

Pour cela nous considérons un modèle hiérarchique à deux couches, dans lequel la couche cachée est issue d'un mélange de gaussiennes bivariées et la couche d'observation est obtenue par une distribution bivariée dont le choix dépend du type de données observées. Nous avons choisi le mélange de gaussiennes pour la 1$^{\text{ère}}$ couche car il a fait l'objet de nombreuses études et permet d'obtenir une grande diversité de formes de classes, ainsi que des corrélations négatives.

2 Méthode

2.1 Le Modèle

Soit $\mathcal{M}$ un mélange de K lois bivariées. La densité de la $k^{\text{ème}}$ composante de ce mélange est définie pour une variable aléatoire $\boldsymbol{Y}_{ik}$ par le modèle hiérarchique suivant :

$$\left\{ \begin{array}{rcl} \boldsymbol{\theta}_{ik} & \sim & \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \\ \boldsymbol{Y}_{ik}|\boldsymbol{\theta}_{ik} & \sim & \mathcal{L}_{biv}(g^{-1}(\boldsymbol{\theta}_{ik})) \end{array} \right.$$

où $\mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ est la loi gaussienne bivariée de moyenne $\boldsymbol{\mu}_k$, matrice de covariance $\boldsymbol{\Sigma}_k$, g est une fonction de lien et $\mathcal{L}_{biv}$ désgine une loi bivariée de paramètres $g^{-1}(\theta_{ik})$. Nous allons nous limiter au cas où la loi bivariée est formée de deux lois univariées indépendantes.

Le choix de $\mathcal{L}_{biv}$ et g^{-1} dépend du type de données et les deux variables ne sont pas nécessairement du même type. Des données de type différent sont obtenues en posant par exemple $Y_{1ik} = \theta_{1ik}$ pour avoir une variable continue et $Y_{2ik} \sim \mathcal{P}(e^{\theta_{2ik}})$ pour avoir une variable discrète. Nous pouvons utiliser des lois différentes pour les deux variables, par exemple une loi binomiale avec une fonction de lien logit et une loi de Poisson. On peut noter que le cas particulier où on n'a pas de mélange, et où la loi bivariée $\mathcal{L}_{biv}$ est formée de deux lois de Poisson indépendantes et g^{-1} est la fonction exponentielle, correspond au modèle Poisson log-normal proposé par Aitchison et Ho (1989).

2.2 Estimation par maximum de vraisemblance

Nous voulons répartir les n observations bivariées $\boldsymbol{Y}_i$ en K groupes. Le modèle précédent peut s'écrire comme un modèle hiérarchique à deux couches cachées :

$$\boldsymbol{Z} \rightarrow \boldsymbol{\theta} \rightarrow \boldsymbol{Y}$$

où $\boldsymbol{Y}$ est la variables observée, $\boldsymbol{Z}$ et $\boldsymbol{\theta}$ sont deux variables latentes. Z_{ik} est une variables binaire qui vaut 1 si l'observation i appartient à la classe k, 0 sinon. $\boldsymbol{Y}$ est indépendant de $\boldsymbol{Z}$ conditionnellement à $\boldsymbol{\theta}$. La variable latente $\boldsymbol{\theta}$ suit un mélange fini à K composantes de lois normales de paramètres $\boldsymbol{\Phi}_k = \{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}$ et de densité $f_{\boldsymbol{\theta}}$:

$$f_{\boldsymbol{\theta}}(\boldsymbol{\theta}; \boldsymbol{\Phi}) = \sum_{k=1}^{K} \pi_k f_k(\boldsymbol{\theta}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k).$$

La log-vraisemblance des données complètes peut s'écrire :

$$l_c(\boldsymbol{\Phi}; \boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{z}) = \log f_{\boldsymbol{Y}|\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{\theta}) + \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log(\pi_k f_k(\boldsymbol{\theta}_i)).$$

L'algorithme EM (*Expectation Maximization*) est utilisé pour obtenir les estimateurs par maximum de vraisemblance des paramètres du modèle et répartir les données dans les classes. Pour cela nous devons calculer dans l'étape E l'espérance de la vraisemblance conditionnellement aux observations :

$$Q(\boldsymbol{\Phi}, \boldsymbol{\Phi}^{(t)}) = \mathbb{E}_{\boldsymbol{\Phi}^{(t)}}[l_c(\boldsymbol{\Phi}; \boldsymbol{Y}, \boldsymbol{\theta}, \boldsymbol{Z})|\boldsymbol{Y} = \boldsymbol{y}] = \mathbb{E}_{\boldsymbol{\Phi}'}[p_{ik}^{\theta} \log(\pi_k f_{\boldsymbol{\theta}}(\boldsymbol{\theta}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k))|\boldsymbol{Y}_i = \boldsymbol{y}_i],$$

où $p_{ik}^{\theta} = P_{\boldsymbol{\Phi}'}(Z_{ik} = 1|\boldsymbol{\theta}_i)$ est la probabilité a posteriori que $\boldsymbol{\theta}_i$ soit issu de la composante k. Cette expression fait intervenir une intégrale double, donc on ne peut pas la calculer analytiquement.

Nous utilisons une variante de l'algorithme MCEM (Monte Carlo EM) proposé par Booth et Hobert (1999) en l'adaptant à notre cas. Une approximation de $Q(\boldsymbol{\Phi}, \boldsymbol{\Phi}^{(t)})$ est estimée par *importance sampling*, en utilisant comme loi d'importance une distribution de Student bivariée dont les paramètres sont calculés à l'aide d'une approximation de Laplace de la vraisemblance.

La taille N de l'échantillon d'importance est recalculée à chaque itération de façon à obtenir un compromis entre la rapidité des premières itérations et la précision finale de l'estimation.

L'étape M se résout analytiquement. Les estimateurs des paramètres $\boldsymbol{\Phi} = \{\pi, \boldsymbol{\mu}, \boldsymbol{\Sigma}\}$ sont obtenus par les expressions suivantes :

$$\hat{\pi}_k = \frac{\sum_{i=1}^{n} \sum_{j=1}^{N} w_{ij} p_{ik}^{\theta}}{nN}; \quad \hat{\boldsymbol{\mu}}_k = \frac{\sum_{i=1}^{n} \sum_{j=1}^{N} w_{ij} p_{ik}^{\theta} \theta_i^{(j)}}{\sum_{i=1}^{n} \sum_{j=1}^{N} w_{ij} p_{ik}^{\theta}}; \quad \hat{\boldsymbol{\Sigma}}_k = \frac{\sum_{i=1}^{n} \sum_{j=1}^{N} w_{ij} p_{ik}^{\theta} (\theta_i^{(j)} - \mu_k)(\theta_i^{(j)} - \mu_k)^T}{\sum_{i=1}^{n} \sum_{j=1}^{N} w_{ij} p_{ik}^{\theta}}.$$

3 Résultats

Des résultats de classification seront présentés sur données simulées et sur données réelles. Nous étudierons également les gammes de paramètres pour lesquelles les modèles sont estimables pour certains sous-modèles, notamment le mélange de modèles de Poisson log-normal.

4 Discussion

Nous avons proposé une méthode très générale de classification automatique basée sur un modèle hiérarchique de mélange de lois qui peut être utilisée pour des données de types différents. L'estimation des paramètres de ces lois se fait par un algorithme MCEM. L'adaptation de notre méthode d'estimation à différentes lois $f_{\boldsymbol{Y}|\boldsymbol{\theta}}$, donc à différents types de données, est peu coûteuse. La généralisation à plus de deux variables ne devrait pas poser de problèmes particuliers, même s'il faudra probablement diminuer le nombre de paramètres à estimer dans le mélange de lois normales multivariées (notamment reparamétrer les matrices de variance-covariance) afin d'obtenir des modèles plus parcimonieux, de la même manière que Fraley et Raftery (2002).

Pour rendre cette méthode de classification opérationnelle en pratique, il reste à choisir un critère de sélection du nombre de classes, ainsi qu'une méthode efficace d'initialisation automatique des partitions.

Nous prévoyons de diffuser cette méthode sous forme d'un package R.

References

[1] Agresti A. (2002). *Categorical Data Analysis, second edition.* Wiley, New York.

[2] Aitchison J. et Ho, C.H. (1989). The multivariate Poisson log-normal distribution. *Biometrika*, 76 (4), 643–653.

[3] Biernacki C., Celeux G., Govaert G. et Langrognet F. (2006) Model-Based Cluster and Discriminant Analysis with the MIXMOD Software. *Computational Statistics and Data Analysis*, 51 (2), 587–600.

[4] Booth J.G. et Hobert J.P. (1999). Maximizing Generalized Linear Mixed Model likelihoods with an automated Monte Carlo EM algorithm. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 61 (1), 265–285.

[5] Fraley C. et Raftery A.E. (2002). Model-Based Clustering, Discriminant Analysis, and Density Estimation. *Journal of the American Statistical Association*, 97 (458), 611–631.

[6] Fraley C. et Raftery A.E. (2006). MCLUST Version 3 for R: Normal Mixture Modeling and Model-Based Clustering. *Technical Report No. 504*.

[7] McLachlan G.J. et Peel D. (2000). *Finite Mixture Models*. Wiley, New York.