Classification de données ordinales : modèles et algorithmes

François-Xavier Jollois, Mohamed Nadif

Classification de données ordinales : modèles et algorithmes

François-Xavier Jollois & Mohamed Nadif

CRIP5 EA 2517- UFR Math-Info, Université Paris Descartes,
45 rue des Saints-Pères, 75006 PARIS
{francois-xavier.jollois ;mohamed.nadif}@parisdescartes.fr

Résumé. La classification d'un ensemble d'objets décrits par un ensemble de variables ordinales est souvent abordée en considérant ces variables soit continues soit nominales. Dans les deux cas, cela représente souvent des inconvénients. Dans ce travail, nous traitons la classification des données ordinales sous l'approche modèle de mélange. Nous utilisons un modèle de mélange multinomial contraint respectant le caractère ordinal des modalités. L'estimation des paramètres est réalisée par la maximisation de la vraisemblance à l'aide de l'algorithme EM. Dans ce travail, nous considérons aussi une version stochastique et une version classifiante basée sur la maximisation d'une vraisemblance classifiante. Des modèles parcimonieux conduisant, sous l'approche classifiante, à des critères métriques sont décrits. Nous abordons également le problème du nombre de classes, de l'initialisation des algorithmes ainsi que la gestion des classes vides.

Abstract. Clustering of objects described by a set of attributes is often tackle by methods considering the attributes continuous or nominal. In the two cases, there is a lot of disadvantages. In this work, we treat the clustering of ordinal data under the mixture model approach. We use a constraint multinomial mixture model, taking into account the ordinal order of modalities. The estimation of parameters is performed by maximizing the likelihood by the EM algorithm. In this work, we consider a stochastic version and a hard version based on the maximization of the complete data likelihood. Parsimonious models leading, under the hard clustering approach, with metric criteria are described. We tackle also the problem of the number of classes, the initialization of the algorithms and the management of the empty classes.

1 Introduction

Les données ordinales sont présentes dans plusieurs situations : en particulier, elles sont utiles pour représenter les degrés de satisfaction ou les préférences d'individus vis-à-vis de services ou de produits. Plusieurs travaux ont déjà été effectués sur la mise en place de modèles probabilistes ou d'outils statistiques pour décrire les processus de classement ou de notation (Fligner and Verducci, 1993; Marden, 1995). Lorsque l'objectif

est de classifier un ensemble d'objets, l'utilisation des modèles de mélange est devenue une approche classique et puissante (Banfield and Raftery, 1993; Celeux and Govaert, 1995). L'algorithme EM (Dempster et al., 1977), composé de deux étapes : Estimation et Maximisation, est devenu quasiment incontournable dans ce contexte. L'utilisation des modèles de mélanges pour la classification de données ordinales a déjà été étudiée (D'Elia and Piccolo, 2005; Gouget, 2006). Les premiers considèrent ce problème, dans un contexte de comparaisons appariés, que la note donnée par un juge à un item est une réalisation d'une variable binomiale *décalée*. Gouget (2006) compare différentes approches à partir de modèles multinomiaux contraints : linéaire, polynomial et euclidien. Nous nous intéressons ici au modèle polynomial.

Dans ce travail, nous présentons dans la section 2, les modèles de mélange et l'algorithme EM. Dans la section 3, nous présentons le modèle polynomial adapté aux données ordinales et les différentes étapes de l'algorithme EM. Enfin, dans la section 4, nous présentons les travaux en cours que nous développerons lors de notre présentation.

2 Modèle de mélange et algorithme EM

Dans l'approche modèle de mélange, les objets $\mathbf{x}_1, \ldots, \mathbf{x}_n$ (décrits par d variables ordinales) à classifier sont supposés provenir d'un mélange de s densités dans des proportions inconnus $p_1, \ldots, p_s$. Chaque objet $\mathbf{x}_i$ est ainsi une réalisation d'une densité de probabilité (p.d.f.), décrite par $\varphi(\mathbf{x}_i; \boldsymbol{\theta}) = \sum_{k=1}^{s} p_k \varphi_k(\mathbf{x}_i; \boldsymbol{\alpha}_k)$ où $\varphi_k(\mathbf{x}_i; \boldsymbol{\alpha}_k)$ représente la densité de $\mathbf{x}_i$ de paramètre $\boldsymbol{\alpha}_k$. Le vecteur des paramètres à estimer θ est composé de $\mathbf{p} = (p_1, \ldots, p_k)$ et $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_k)$. La log-vraisemblance des données observées $\mathbf{x} = (\mathbf{x}_i, \ldots, \mathbf{x}_n)$ est donnée par $L(\mathbf{x}, \boldsymbol{\theta}) = \sum_{i=1}^{n} \log\left(\sum_{k=1}^{s} p_k \varphi_k(\mathbf{x}_i; \boldsymbol{\alpha}_k)\right)$. Dans la suite, nous allons aborder le problème de la classification sous l'approche estimation : les paramètres sont d'abord estimés, puis la partition en est déduite par la méthode du maximum a posteriori (MAP). L'estimation des paramètres du modèle passe par la maximisation de $L(\mathbf{x}, \boldsymbol{\theta})$. Une solution itérative pour la résolution de ce problème est l'algorithme EM (Dempster et al., 1977). Le principe de cet algorithme est de maximiser de manière itérative l'espérance de la log-vraisemblance complétée conditionnellement aux données $\mathbf{x}$ et la valeur du paramètre courant $\boldsymbol{\theta}^{(q)}$:

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(q)}) = \sum_{i=1}^{n} \sum_{k=1}^{s} t_{ik}^{(q)} \left(\log(p_k) + \log \varphi_k(\mathbf{x}_i; \boldsymbol{\alpha}_k)\right)$$

où $t_{ik}^{(q)} \propto p_k^{(q)} \varphi_k(\mathbf{x}_i; \boldsymbol{\alpha}_k^{(q)})$ est la probabilité conditionnelle a posteriori. Chaque itération de EM a deux étapes :

- **Estimation** : On calcule $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(q)})$, notons que dans le contexte modèle de mélange, cette étape se réduit aux calculs des $t_{ik}^{(q)}$.
- **Maximisation** : On cherche le paramètre $\boldsymbol{\theta}^{(q+1)}$ qui maximise $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(q)})$.

3 Données ordinales

Lorsqu'on ne tient pas compte de l'ordre des modalités, on considère que les données sont constitués d'un échantillon $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$, où $\mathbf{x}_i = (x_i^{je}; j = 1, \ldots, d; e = 1, \ldots, c_j)$, avec c_j le nombre de modalités de la variable j et $x_i^{je} = 1$ si l'individu i prend la modalité e pour la variable j, et 0 sinon. On utilise alors le modèle des classes latentes (Lazarfeld and Henry, 1968), dont le principe de base est la supposition d'une variable qualitative latente à c_j modalités dans les données. Dans ce modèle, les associations entre chaque paire de variables disparaissent, si la variable latente est constante. C'est le modèle basique de l'analyse des classes latentes, avec l'hypothèse fondamentale d'indépendance locale. Cette hypothèse est couramment choisie quand les données sont de type qualitatif ou binaire (Celeux and Govaert, 1992; Cheeseman and Stutz, 1996). Ainsi, la densité d'une observation $\mathbf{x}_i$ peut se décrire comme suit :

$$\varphi_k(\mathbf{x}_i; \boldsymbol{\alpha}_k) = \prod_{j=1}^{d} \prod_{e=1}^{c^j} \left(\alpha_k^{je}\right)^{x_i^{je}}, \quad \text{avec} \quad \sum_{e=1}^{c^j} \alpha_k^{je} = 1$$

L'hypothèse d'indépendance locale permet d'estimer les paramètres séparément. Cette hypothèse simplifie grandement les calculs, principalement quand le nombre de variables est grand. Bien que cette affirmation est toujours fausse dans la pratique, l'indépendance locale est généralement très performante pour la classification. Ce modèle noté $\lfloor p_k, \alpha_k^{je} \rfloor$, conduit malheureusement à des classes dont les centres ne sont pas de même nature que les données initiales. En imposant une contrainte sur les probabilités associées aux modalités, on peut considérer le modèle noté $\lfloor p_k, \varepsilon_k^j \rfloor$ (Nadif and Marchetti, 1993) conduisant à des classes formées par des modalités notées a_k^j. Dans ce modèle, nous considérons que la probabilité α_k^{je} associée à une modalité e de j est égale à $1 - \varepsilon_k^j$ si $e = a_k^j$ et égale à $\frac{\varepsilon_k^j}{c_j - 1}$ sinon. Malheureusement, ces deux modèles ne prennent pas en compte l'aspect ordinal des données. Pour ce faire, il est nécessaire d'imposer des contraintes d'ordre sur les probabilités des modalités : pour chaque classe k et chaque variable j, les probabilités α_k^{je} de la modalité e vont en décroissant à partir de la modalité centrale a_k^j.

En premier lieu, le plus simple est de considérer cette décroissance constante entre deux modalités. Dans le cas de variables avec un nombre de modalités important, ce modèle s'apparente au modèle multinomial $\lfloor p_k, \varepsilon_k^j \rfloor$. Une solution pour éviter cet écueil est de considérer que la croissance n'est pas linéaire, mais polynomial de degré q. Ainsi, dans ce modèle (Gouget, 2006), noté ici $\lfloor p_k, \alpha_k^{je}, q \rfloor$, la probabilité d'une modalité e pour la variable j dans la classe k peut s'exprimer par :

$$\alpha_k^{je} = \begin{cases} p_{a_k^j} & \text{si } e = a_k^j \\ \dfrac{(1 - p_{a_k^j})(1 + \max(a_k^j - 1; c_j - a_k^j) - |e - a_k^j|)^q}{\sum_{u \neq a_k^j} (1 + \max(a_k^j - 1; c_j - a_k^j) - |u - a_k^j|)^q} & \text{sinon} \end{cases}$$

Notons que dans ce modèle, on peut introduire une paramétrisation supplémentaire en considérant que q est variable et dépend à la fois de la classe k et de la variable j, dans ce cas le modèle est noté $\lfloor p_k, \alpha_k^{je}, q_k^j \rfloor$. Nous nous concentrons, dans ce travail, sur le modèle simple $\lfloor p_k, \alpha_k^{je}, q \rfloor$. Pour l'estimation des paramètres du modèle, l'algorithme EM maximisera itérativement $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(q)})$ qui prend la forme suivante

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(q)}) = \sum_{k=1}^{s}\sum_{i=1}^{n} t_{ik}\log(p_k) + \sum_{i=1}^{n} t_{ik}\sum_{j=1}^{d}\sum_{e=1}^{c^j} x_i^{je}\log(\alpha_k^{je}).$$

Il est nécessaire de s'assurer que la probabilité associée à la modalité a_k^j choisie est supérieure à toutes les autres probabilités des modalités $e \neq a_k^j$. Ainsi, nous définissons le seuil $\beta_q^{c_j}$ comme étant la probabilité telle que, pour tout $e = 1, \ldots, c^j$, $e \neq a_k^j$, avec

$$\frac{(1-\beta_q^{c_j})(1+\max(a_k^j-1; c_j-a_k^j)-|e-a_k^j|)^q}{\sum_{u\neq a_k^j}(1+\max(a_k^j-1; c_j-a_k^j)-|u-a_k^j|)^q} \leq \beta_q^{c_j}$$

Ce seuil ne dépend que du nombre de modalités c_j et du paramètre q.

4 Conclusion

Dans l'approche EM dite aussi approche *estimation*, en utilisant le principe du maximum a posteriori nous pouvons construire des classes. Dans ce travail, nous considérerons aussi deux autres approches : une classifiante et une stochastique.

Nous insisterons sur l'approche classiante qui est basée sur la maximisation d'une vraisemblance classifiante réalisée par une version classifiante de EM. Rappelons que dans cette approche, lorsque les proportions sont supposées égales les modèles de mélange conduisent à des algorithmes de type nuées dynamiques. Dans notre cas, en imposant des contraintes sur les paramètres α_k^{je} nous proposerons différents modèles parcimonieux et décrirons les critères ainsi les différentes mesures de dissimilarités utilisées.

Enfin, dans nos expériences numériques à partir de données simulées et réelles, nous aborderons les problèmes d'initialisation, la gestion des classes vides et le choix du nombre de classes.

Bibliographie

J. D. Banfield and A. E. Raftery. Model-based gaussian and non-gaussian clustering. *Biometrics*, 49 :803–821, 1993.

G. Celeux and G. Govaert. A classification em algorithm for clustering and two stochastic versions. *Computational Statistics & Data Analysis*, 14 :315–332, 1992.

G. Celeux and G. Govaert. Gaussian parcimonious clustering methods. *Pattern Recognition*, 28 :781–793, 1995.

P. Cheeseman and J. Stutz. Bayesian classification (autoclass) : Theory and results. In U. Fayyad, G. Piatetsky-Shapiro, and R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, pages 61–83. AAAI Press, 1996.

Angelia D'Elia and Domenico Piccolo. A mixture model for preferences data analysis. *Computational Statistics & Data Analysis*, 49 :917–934, 2005.

A. Dempster, N. Laird, and D. Rubin. Mixture densities, maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statitical Society*, 39(1) : 1–38, 1977.

M.A. Fligner and J.S. Verducci. *Probability models and statistical analysis of ranking data*. Springer, New-York, 1993.

Cyril Gouget. *Utilisation des modèles de mélange pour la classification automatique de données ordinales*. PhD thesis, Université de Technologie de Compiègne, December 2006.

P.F. Lazarfeld and N.W. Henry. *Latent Structure Analysis*. Houghton Mifflin, Boston, 1968.

J.L. Marden. *Analyzing and modeling rank data*. Chapman & Hall, London, 1995.

Mohamed Nadif and Franck Marchetti. Classification de donnés qualitatives et modèles. *Revue de Statistique Appliquée*, XLI(1) :55–69, 1993.