Classification d'un tableau de contingence et modèle probabiliste

Gérard Govaert*, Mohamed Nadif**

*Heudiasyc, UMR CNRS 6599, Université de Technologie de Compiègne,
BP 20529, 60205 Compiègne Cedex, France
gerard.govaert@utc.fr
**IUT de Metz, LITA, Université de Metz,
Ile du Saulcy, 57045 Metz Cedex, France
nadif@iut.univ-metz.fr

Résumé. Les modèles de mélange, qui supposent que l'échantillon est formé de sous-populations caractérisées par une distribution de probabilité, constitue un support théorique intéressant pour étudier la classification automatique. On peut ainsi montrer que l'algorithme des k-*means* peut être vu comme une version classifiante de l'algorithme d'estimation EM dans un cas particulièrement simple de mélange de lois normales.

Lorsque l'on cherche à classifier les lignes (ou les colonnes) d'un tableau de contingence, il est possible d'utiliser une variante de l'algorithme des k-*means*, appelé Mndki2, en s'appuyant sur la notion de profil et sur la distance du khi-2. On obtient ainsi une méthode simple et efficace pouvant s'utiliser conjointement à l'analyse factorielle des correspondances qui s'appuie sur la même représentation des données.

Malheureusement et contrairement à l'algorithme des k-*means* classique, les liens qui existent entre les modèles de mélange et la classification ne s'appliquent pas directement à cette situation. Dans ce travail, nous montrons que l'algorithme Mndki2 peut être associé, à une approximation près, à un modèle de mélange de lois multinomiales.

1 Introduction

Les modèles de mélange, qui supposent que l'échantillon est formé de sous-populations caractérisées par une distribution de probabilité, sont des modèles très souples permettant de prendre en compte des situations variées comme la présence de populations hétérogènes ou d'éléments atypiques. Grâce à l'algorithme d'estimation EM, particulièrement adapté à cette situation, les modèles de mélange ont fait l'objet de nombreux développements en statistique et en particulier en classification automatique. On peut ainsi montrer que l'algorithme des k-*means* peut être vu comme une version classifiante de l'algorithme EM, appelé CEM, dans un cas particulièrement simple de mélange de lois normales. Dans ce travail, on étudie comment ces propriétés peuvent être étendues aux tableaux de contingence.

Rappelons qu'un tableau de contingence est obtenu à partir du croisement de 2 variables qualitatives ; par exemple, si on note I et J les ensembles de r et s modalités de chaque variable, chaque élément x_{ij} de la matrice de données contiendra le nombre

d'éléments prenant respectivement les modalités i et j pour chacune des 2 variables qualitatives. Les tableaux de contingence sont quelquefois directement obtenus à partir de la saisie des données ; on retrouve, par exemple, ce type de tableaux en analyse de données textuelles où la matrice de données comptabilise le nombre d'occurrences d'un ensemble de mots et d'un ensemble de documents. On parle alors quelquefois de tableau d'occurrences. En outre, la plupart des méthodes d'analyse de tels tableaux s'appliquent généralement aussi à des tableaux possédant des propriétés équivalentes (Benzécri 1973) comme les tableaux de variables quantitatives avec des variables homogènes et positives (quantités de même nature comme des longueurs ou des poids) ou même les tableaux binaires.

La section 2 sera consacrée à l'algorithme Mndki2, algorithme permettant de classifier les lignes, ou les colonnes, d'un tableau de contingence. Dans la section 3, nous définirons un modèle de mélange de lois de probabilité adapté à un tel type de données et la section 4 sera consacrée à l'application de l'algorithme d'estimation EM à ce modèle. La version classifiante de cet algorithme EM, baptisée Cemki2, sera étudiée dans la section 6. On montrera en particulier qu'on obtient ainsi un algorithme maximisant un critère approximant celui utilisé par Mndki2 et fournissant des résultats quasi-identiques.

Notations Dans tout ce texte, on notera $\mathbf{x} = (x_{ij})$ le tableau de contingence construit sur les deux ensembles I et J ayant respectivement r et s éléments, $n = \sum_{i,j} x_{ij}$ la somme des éléments du tableau et $x_{i.} = \sum_j x_{ij}$ et $x_{.j} = \sum_i x_{ij}$ ses marges. On utilisera aussi le tableau des fréquences relatives $f_{ij} = x_{ij}/n$, $f_{i.} = \sum_j f_{ij}$ et $f_{.j} = \sum_i f_{ij}$ ses marges et les profils en ligne $f_J^i = (f_{i1}/f_{i.}, \ldots, f_{ir}/f_{i.})$. Une partition en g classes de l'ensemble I sera notée $\mathbf{z} = (z_1, \ldots, z_r)$, où $z_i \in \{1, \ldots, g\}$ indique la classe de l'objet i. $z_i = k$ lui-même pourra aussi être représenté par le vecteur $(z_{i1}, \ldots, z_{ig})$ avec $z_{ik} = 1$ si $i \in z_k$ et $z_{ik} = 0$ sinon. Dans ce dernier cas, la classification sera représentée par une matrice $\mathbf{z}$ de n vecteurs de $\mathbb{R}^g$ vérifiant $z_{ik} \in \{0, 1\}$ et $\sum_k z_{ik} = 1$. Ainsi, la classe k correspond à l'ensemble des objets i tel que $z_i = k$ ou encore $z_{ik} = 1$ et $z_{i\ell} = 0\ \forall \ell \neq k$. Par ailleurs, pour simplifier la présentation, les sommes et les produits portant sur les lignes, les colonnes ou les classes seront indicés respectivement par les lettres i, j et k sans indiquer les bornes de variation qui seront donc implicites. Ainsi, la somme $\sum_i$ portera sur tous les lignes i allant de 1 à r.

2 Algorithme Mndki2

2.1 L'objectif

Pour mesurer l'information apportée par un tableau de contingence, c'est-à-dire les liens existant entre deux ensembles I et J mis en correspondance dans le tableau de données, il existe plusieurs mesures dont l'une des plus courantes est le χ^2 de contingence. Ce critère, utilisé par exemple dans l'analyse factorielle des correspondances, est

défini de la manière suivante

$$\chi^2(I,J) = \sum_{i,j} \frac{(x_{ij} - \frac{x_{i.}x_{.j}}{n})^2}{\frac{x_{i.}x_{.j}}{n}} = n \sum_{i,j} \frac{(f_{ij} - f_{i.}f_{.j})^2}{f_{i.}f_{.j}}.$$

Cette quantité représente l'écart entre les fréquences théoriques $f_{i.}f_{.j}$ que l'on aurait s'il y avait indépendance entre les deux ensembles I et J et les fréquences observées f_{ij} : une grande valeur correspondra à une forte dépendance alors qu'une valeur nulle correspondra à une indépendance entre I et J.

Cette mesure d'information peut également être utilisée pour évaluer la qualité d'une partition $\mathbf{z}$ de l'ensemble I : pour ceci, on associera à cette partition $\mathbf{z}$ le χ^2 du tableau de contingence à g lignes et r colonnes obtenu à partir du tableau initial en faisant la somme des éléments de chaque classe : $x_{kj} = \sum_{i|z_i=k} x_{ij} \quad \forall k,j$ et qui sera noté $\chi^2(\mathbf{z},J)$. On établira plus loin la relation $\chi^2(I,J) \geq \chi^2(\mathbf{z},J)$ qui montre que le regroupement des valeurs du tableau de contingence conduit nécessairement à une perte d'information. L'objectif de la classification sera donc de trouver la partition $\mathbf{z}$ qui minimise cette perte, c'est-à-dire qui maximise le critère $\chi^2(\mathbf{z},J)$.

Remarquons que dans le cas idéal, où les profils en ligne sont égaux à l'intérieur de chaque classe, alors $\chi^2(\mathbf{z},J) = \chi^2(I,J)$ et il n'y a donc pas de perte d'information. Par ailleurs, le problème que l'on vient de définir n'a de sens que pour un nombre fixé de classes sinon la partition optimale est simplement la partition où chaque élément de I forme une classe.

2.2 L'algorithme

L'algorithme Mndki2 repose sur la représentation géométrique d'un tableau de contingence utilisée dans l'analyse factorielle des correspondances. Cette représentation est justifiée pour plusieurs raisons, en particulier pour les rôles analogues dévolus à chacune des deux dimensions du tableau analysé. Dans cette représentation, à l'ensemble des lignes est associé dans $\mathbb{R}^s$ le nuage $\mathcal{N}(I)$ des r vecteurs des profils f_J^i munis des masses $f_{i.}$. La métrique utilisée dans cet espace est la métrique quadratique définie par la matrice diagonale $\mathrm{diag}(\frac{1}{f_{.1}}, \ldots, \frac{1}{f_{.s}})$ appelée métrique du χ^2 et notée $d_{\chi 2}$. Avec cette représentation, la relation classique de décomposition de l'inertie en une somme d'inertie intraclasse et d'inertie interclasse s'écrit simplement $\chi^2(I,J) = n.W(\mathbf{z}) + \chi^2(\mathbf{z},J)$ où $W(\mathbf{z}) = \sum_k \sum_{i|z_i=k} f_{i.} d^2_{\chi 2}(f_J^i, g_k)$ avec g_k centre de gravité de la classe k.

En conséquence, puisque la quantité $\chi^2(I,J)$ ne dépend pas de la partition $\mathbf{z}$, la recherche de la partition maximisant le critère $\chi^2(\mathbf{z},J)$ est équivalente à la recherche de la partition minimisant le critère $W(\mathbf{z})$. Pour minimiser ce critère d'inertie d'inertie intraclasse, il est alors possible d'appliquer la méthode des k-*means* sur le nuage des profils avec la métrique du χ^2. On obtient ainsi un algorithme itératif, appelé Mndki2, maximisant localement le critère $\chi^2(\mathbf{z},J)$.

Malheureusement, ce critère, contrairement à celui des k-*means*, ne vérifie pas les conditions sous lesquelles un critère métrique de classification est équivalent à un critère de vraisemblance classifiante associé à un modèle de mélange (Govaert 1989). Il est toutefois possible de montrer que ce critère est lié, au moins approximativement, au modèle de mélange de lois multinomiales.

3 Le modèle de mélange de lois multinomiales

Le modèle de mélange proposé consiste à considérer que chaque ligne $\boldsymbol{x}_i$ du tableau de contingence est générée suivant le mécanisme suivant :

- la marge $x_{i.}$ est simulée suivant une loi discrète ψ à valeur entière quelconque (Poisson, binomiale négative,...) ;
- la classe k est tirée au hasard suivant les probabilités $\pi_1, \ldots, \pi_g$;
- le vecteur $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ir})$ est simulé suivant la distribution multinomiale de paramètres $x_{i.}, \alpha_{k1}, \ldots, \alpha_{kr}$.

Plus formellement, si on note $\boldsymbol{\theta} = (\pi_1, \ldots, \pi_g, \alpha_{11}, \ldots, \alpha_{gr})$ le paramètre du modèle et φ est la densité de la loi multinomiale, la densité du modèle s'écrit

$$\begin{aligned} f(\mathbf{x};\boldsymbol{\theta}) &= \prod_i f(\boldsymbol{x}_i;\boldsymbol{\theta}) = \prod_i \left(\psi(x_{i.}) \sum_k \pi_k \varphi(\boldsymbol{x}_i; x_{i.}, \alpha_{k1}, \ldots, \alpha_{ks}) \right) \\ &= \prod_i \left(\psi(x_{i.}) \sum_k \pi_k \frac{x_{i.}!}{x_{i1}! \ldots x_{is}!} \alpha_{k1}^{x_{i1}} \ldots \alpha_{ks}^{x_{is}} \right) = A \prod_i \sum_k \pi_k \alpha_{k1}^{x_{i1}} \ldots \alpha_{ks}^{x_{is}} \end{aligned}$$

où $A = \prod_i \psi(x_{i.}) \prod_i \frac{x_{i.}!}{x_{i1}! \ldots x_{is}!}$ est un terme qui ne dépend pas du paramètre $\boldsymbol{\theta}$. On notera $L(\boldsymbol{\theta};\mathbf{x}) = \sum_i \log \sum_k \pi_k \alpha_{k1}^{x_{i1}} \ldots \alpha_{ks}^{x_{is}}$ la log-vraisemblance associée et $L(\boldsymbol{\theta};\mathbf{x},\mathbf{z}) = \sum_{i,k} z_{ik} \left(\ln \pi_k + \sum_j x_{ij} \log \alpha_{kj} \right)$ la log-vraisemblance des données complétées.

Le problème généralement posé est alors l'estimation du paramètre $\boldsymbol{\theta}$ à partir de l'échantillon. Il s'agit d'un problème classique d'estimation statistique. L'utilisation en classification automatique de ce modèle de mélange conduit en réalité à un autre problème : retrouver le composant dont est issu chaque élément de l'échantillon. Nous verrons plus loin comment utiliser le modèle de mélange pour atteindre cet objectif.

4 Estimation des paramètres

L'estimation du paramètre $\boldsymbol{\theta}$ peut être obtenue en maximisant la log-vraisemblance $L(\boldsymbol{\theta};\mathbf{x})$ à l'aide de l'algorithme EM. Sous certaines conditions de régularité, il a été établi que l'algorithme EM assure une convergence vers un maximum local de la vraisemblance. Il a un bon comportement pratique mais peut être toutefois assez lent dans certaines situations ; c'est le cas, par exemple, si les classes sont très mélangées. Cet algorithme, proposé par Dempster, Laird et Rubin (1977) dans un papier célèbre, souvent simple à mettre en place est devenu populaire et a fait l'objet de nombreux travaux que l'on pourra trouver dans l'ouvrage très complet de McLachlan et Krishnan (1997).

Partant d'un paramètre initial $\boldsymbol{\theta}^{(0)}$, une itération de l'algorithme EM consiste à maximiser l'espérance de la log-vraisemblance des données complétées conditionnellement à l'estimation courante $\boldsymbol{\theta}^{(c)}$ et aux données $\mathbf{x}$ qui s'écrit

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(c)}) = \sum_{i,k} t_{ik}^{(c)} \left(\log \pi_k + \sum_j x_{ij} \log \alpha_{kj} \right)$$

où $t_{ik}^{(c)}$ est la probabilité d'appartenance de $\boldsymbol{x}_i$ à la classe k conditionnellement à $\boldsymbol{\theta}^{(c)}$. Chaque itération se décompose en 2 étapes : l'étape E calcule les probabilités $t_{ik}^{(c)}$ $t_{ik}^{(c)} = \frac{\pi_k \alpha_{k1}^{x_{i1}} ... \alpha_{kr}^{x_{ir}}}{\sum_\ell \pi_\ell \alpha_{\ell 1}^{x_{i1}} ... \alpha_{\ell r}^{x_{ir}}}$; l'étape M détermine le paramètre $\boldsymbol{\theta}$ maximisant $Q(\theta, \theta^{(c)})$. On montre facilement que cette maximisation conduit pour notre modèle aux relations $\pi_k^{(c+1)} = \frac{\sum_i t_{ik}^{(c)}}{n}$ et $\alpha_{kj}^{(c+1)} = \frac{\sum_i t_{ik}^{(c)} x_{ij}}{\sum_i t_{ik}^{(c)} x_{i.}}$.

5 Algorithme Cemki2

L'utilisation du modèle de mélange pour obtenir une partition des données initiales peut alors se faire en rangeant chaque individu dans la classe maximisant la probabilité *a posteriori* t_{ik} calculée à partir des paramètres estimés. Une autre solution consiste à rechercher une partition de l'échantillon de telle sorte que chaque classe k soit assimilable à un sous-échantillon issue de la loi $f(., \alpha_k)$. Il s'agit donc d'estimer simultanément les paramètres du modèle et la partition recherchée en maximisant la vraisemblance complétée $L(\boldsymbol{\theta}; \mathbf{x}, \mathbf{z})$ définie précédemment.

Cette maximisation peut être obtenue par l'algorithme CEM (classification EM) (Celeux et Govaert 1992), version classifiante de l'algorithme EM obtenue en lui ajoutant une étape de classification. Chaque itération se décompose maintenant en 3 étapes : l'étape E calcule les $t_{ik}^{(c)}$ comme dans l'algorithme EM ; l'étape C détermine la partition $\mathbf{z}^{(c+1)}$ en rangeant chaque $\mathbf{x}_i$ dans la classe maximisant $t_{ik}^{(c)}$; l'étape M maximise la vraisemblance conditionnellement aux $z_{ik}^{(c+1)}$: les estimations du maximum de vraisemblance des π_k et des α_k sont obtenues en utilisant les classes de la partition $\mathbf{z}^{(c+1)}$ comme sous-échantillons. On montre facilement que cette maximisation conduit à $\pi_k^{(c+1)} = \frac{n_k}{n}$ et $\alpha_{kj}^{(c+1)} = \frac{x_{kj}}{x_{k.}}$ où n_k est le cardinal de la classe k de la partition $\mathbf{z}^{(c+1)}$, $x_{kj} = \sum_{i,k} z_{ik}^{(c+1)} x_{ij}$ et $x_{k.} = \sum_j x_{kj}$.

Après la maximisation en $\boldsymbol{\theta}$ et en notant $f_{kj} = \frac{x_{kj}}{n}$, le critère s'exprime

$$\begin{aligned} L(\boldsymbol{\theta}; \mathbf{x}, \mathbf{z}) &= \sum_k n_k \ln \pi_k + \sum_{k,j} x_{kj} \log \frac{x_{kj}}{x_{k.}} \\ &= \sum_k n_k \ln \pi_k + n \sum_{k,j} f_{kj} \log \frac{f_{kj}}{f_{k.} f_{.j}} + n \sum_j f_{.j} \log f_{.j}. \end{aligned}$$

En utilisant l'approximation $2x \log x \approx x^2 - 1$, cette quantité peut être approximée par

$$\sum_k n_k \ln \pi_k + \frac{n}{2} \sum_{k,j} \frac{(f_{kj} - f_{k.} f_{.j})^2}{f_{k.} f_{.j}} + n \sum_j f_{.j} \log f_{.j}$$

On retrouve alors à une constante près et lorsque les proportions sont fixées, le critère maximisé par l'algorithme Mndki2 : la maximisation de la vraisemblance classifiante est donc approximativement équivalente à la maximisation du χ^2 de contingence et utiliser ce critère revient à supposer que les données sont issues d'un modèle de mélange de lois multinomiales. En pratique, les deux algorithmes Mndki2 et Cemki2 fournissent généralement les mêmes partitions.

6 Conclusion

L'interprétation probabiliste de l'algorithme Mndki2 constitue un support intéressant pour traiter différentes situations qui, sinon, auraient nécessité le développement de méthodes ad hoc : par exemple, elle permet de prendre en compte des situations où les classes sont très mélangées en appliquant l'algorithme EM, de prendre en compte des classes de proportions très différentes alors que le critère du χ^2 suppose implicitement des proportions égales, de traiter le problème du nombre de classes en s'appuyant sur les outils statistiques de choix de modèle comme l'utilisation des critères de complexité BIC ou ICL. Par ailleurs, l'interprétation géométrique du modèle de mélange de lois multinomiales permet en utilisant la représentation factorielle de l'analyse des correspondances de visualiser les résultats fournis par les algorithmes EM ou CEM appliqués au modèle de mélange.

Références

Benzécri, J.-P. (1973) L'analyse des données, tome 2 : l'analyse des correspondances, Dunod, Paris.

Celeux, G. et Govaert, G. (1992), A classification EM algorithm for clustering and two stochastic versions, Computational Statistics and Data Analysis, 14(3), pp 315-332.

Dempster, A.P., Laird, N.M., et Rubin, D.B. (1977), Maximum likelihood from incomplete data via the EM algorithm (with discussion), Journal of the Royal Statistical Society, B 39, pp 1-38.

Govaert, G. (1989), Clustering model and metric with continuous data, In Diday, Y., editor, Data Analysis, Learning Symbolic and Numeric Knowledge, Nova Science Publishers, New York, pp 95-102.

McLachlan, G.J. et Krishnan, K. (1997), The EM Algorithm, Wiley, New York.

Summary

Mixture models, which suppose that the sample is made of subpopulations characterized by a probability distribution, constitutes an interesting theoretical support to study clustering. One can thus show that the k-means algorithm can be seen like a classifying version of the EM algorithm in a particularly simple case of mixture of normal distributions. When one seeks to classify the lines (or the columns) of a contingency table, it is possible to use an alternative of the k-means algorithm, called Mndki2, based on the concept of profile and the khi-2 distance. We obtain a simple and effective method which can be used jointly with the factorial correspondence analysis based on the same representation of the data. Unfortunately and contrary to the traditional k-means algorithm, the links which exist between mixture models and clustering does not apply directly to this situation. In this work, we show that the Mndki2 algorithm can be associated, with an approximation, with a mixture model of multinomial distributions.