Réalisation simultanée d'une réduction de la dimension et d'une classification multi-objectifs

Vincent Vandewalle

Université Lille 2, EA 2694 & Inria,
25-27 rue du Maréchal Foch,
59100 Roubaix
vincent.vandewalle@univ-lille2.fr

Résumé. En classification non supervisée à base de modèles pour les données quantitatives, on suppose en général qu'une seule variable de classe explique l'hétérogénéité des données. Cependant, quand les variables proviennent de différentes sources il est souvent irréaliste de supposer que cette hétérogénéité peut être expliquée par seulement une variable. Si une telle hypothèse est faite, elle peut conduire à l'estimation d'un grand nombre de groupes, ce qui peut être difficile à interpréter. On propose ici un modèle de mélange multi-objectifs, il suppose l'existence de plusieurs variables latentes de classification, chacune d'entre-elles expliquant l'hétérogénéité des données selon une projection classifiante particulière. L'estimation des paramètres du modèle est réalisée par un algorithme de type EM. Les résultats obtenus sont des projections des données sur des composantes classifiantes, ce qui permet une interprétation synthétique des principales classifications présentes dans les données. Le comportement du modèle proposé est illustré sur un jeu de données réelles.

Mots-clés. classification non supervisée, réduction de dimension, modèles de mélange, algorithme EM

Abstract. In model based clustering of quantitative data it is often supposed that only one clustering variable explains the heterogeneity of all the others variables. However, when variables come from different sources, it is often unrealistic to suppose that the heterogeneity of the data can only be explained by one variable. If such an assumption is made, this could lead to a high number of clusters which could be difficult to interpret. A model based multi-objective clustering is proposed, it assumes the existence of several latent clustering variables, each one explaining the heterogeneity of the data on some clustering projection. In order to estimate the parameters of the model an EM algorithm is proposed. The obtained results are projections of the data on some principal clustering components allowing some synthetic interpretation of the principal clusters raised by the data. The behavior of the model is illustrated on a real dataset.

Keywords. unsupervised clustering, dimension réduction, mixture models, EM algorithm

1 Introduction

En analyse exploratoire, le statisticien utilise souvent la classification non supervisée et la visualisation afin d'améliorer sa connaissances des données. Dans la visualisation il recherche des composantes principales expliquant certaines caractéristiques des données, par exemple, l'axe de plus forte variance. En classification non supervisée, l'objectif est de trouver des groupes expliquant l'hétérogénéité des données. En pratique, ces deux approches sont souvent utilisées ensembles. Par exemple, le praticien peut commencer par effectuer une ACP, puis effectuer un algorithme de $k-means$ basé sur un nombre limité de composantes principales. Toutefois, la fusion de ces deux approches d'un point de vue rigoureux peut être difficile. Par exemple, comment choisir simultanément le nombre de composants et le nombre de classes ?

Un bon moyen de fusionner rigoureusement la classification non supervisée et la visualisation consiste à utiliser des modèles de mélange (McLachlan et Peel (2004)). Récemment Bouveyron et Brunet (2012) ont proposé l'algorithme Fisher-EM qui effectue simultanément la classification non supervisée et la réduction de dimension. Ceci est fait par une version modifiée de l'algorithme EM. Cette approche permet d'appliquer la même projection sur toutes les données. Pour certains modèles, l'algorithme Fisher-EM peut être utilisé pour obtenir le maximum de vraisemblance. Un des principaux avantages de l'approche proposée est qu'elle combine naturellement classification non supervisée et visualisation, contrairement aux visualisations qui ne se concentrent pas sur le point de vue de la classification comme l'ACP, ou aux approches de classification qui ne prennent pas en compte l'aspect visualisation.

Le problème abordé dans cette communication se situe entre la réduction de dimension et la classification non supervisée. Ici nous recherchons des composantes principales qui se réfèrent à un point de vue de classification non supervisée comme dans Bouveyron et Brunet (2012). Mais, ici on suppose que les données peuvent contenir plusieurs variables latentes classifiantes, ce qui est à notre connaissance un problème rarement abordé en classification non supervisée. Il en résulte des composantes principales classifiantes qui permettent au praticien de voir ses données sous un nouvel angle.

Dans un premier temps nous présenterons le modèle de classification multi-objectifs. Dans un second temps nous présenterons l'algorithme d'estimation des paramètres. Enfin nous présenterons une illustration de notre méthode sur un jeu de données réelles.

2 Présentation du modèle de mélange pour la classification multi-objectifs

On suppose qu'on dispose de n données quantitatives en dimension d, la donnée i est notée $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{id})^{\mathrm{T}}$, avec x_{ij} la valeur de la variable j pour l'individu i. Le jeu de données complet est noté $\mathbf{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^{\mathrm{T}}$. Nous supposons de plus que nous disposons

de H variables de classe (contrairement à une seule habituellement) $\boldsymbol{z}_i^1, \ldots, \boldsymbol{z}_i^H$ avec K_1, ..., K_H modalités. On notera $z_{ik}^h = 1$ si la variable de classe h prend la modalité k pour l'individu i et $z_{ik}^h = 0$ sinon.

Le modèle générateur de ces données est le suivant :

1. On suppose que $\boldsymbol{z}_i^1, \ldots, \boldsymbol{z}_i^H$ sont indépendantes avec $p(z_{ik}^h = 1)$ noté π_k^h.
2. On note par $\boldsymbol{y}_i^h \in \mathbb{R}^{p_h}$ la variable classifiante reliée à la variable de classe $\boldsymbol{z}_i^h$:

$$\boldsymbol{y}_i^h | z_{ik}^h = 1 \sim \mathcal{N}_{p_h}(\boldsymbol{\nu}_k^h, I_{p_h})$$

   où $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ est la loi normale en dimension p d'espérance $\boldsymbol{\mu}$ et de variance $\boldsymbol{\Sigma}$.
3. On note par $\boldsymbol{u}_i$ le vecteur des variables non classifiantes dont on suppose que

$$\boldsymbol{u}_i \sim \mathcal{N}_{d-p_\bullet}(\boldsymbol{\gamma}, I_{d-p_\bullet}),$$

   avec $p_\bullet = \sum_{h=1}^H p_h$.
4. Enfin on définit $\boldsymbol{x}_i$ par :

$$\boldsymbol{x}_i = \begin{pmatrix} \boldsymbol{V}_1 \\ \vdots \\ \boldsymbol{V}_H \\ \boldsymbol{R} \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{y}_i^1 \\ \vdots \\ \boldsymbol{y}_i^H \\ \boldsymbol{u}_i \end{pmatrix}.$$

La figure 1 illustre le modèle dans le cas où $H = 2$.

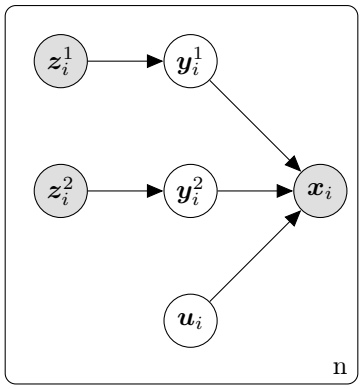

FIGURE 1: Réseau bayésien associé au modèle dans le cas où $H = 2$

Dans le cas supervisé, les données dont on dispose sont les $\boldsymbol{x}_i$, $\boldsymbol{z}_i^1$, ..., $\boldsymbol{z}_i^H$, et les variables manquantes sont $\boldsymbol{y}_i^1$, ..., $\boldsymbol{y}_i^H$ et $\boldsymbol{u}_i$. Bien sûr si on connaissait $\boldsymbol{V}_1$, ..., $\boldsymbol{V}_H$ et $\boldsymbol{R}$ leur calcul à partir de $\boldsymbol{x}_i$ serait direct. Dans le cas non supervisé les variables $\boldsymbol{z}_i^1$, ..., $\boldsymbol{z}_i^H$ sont elles aussi manquantes et on n'observe plus que les $\boldsymbol{x}_i$.

Les paramètres du modèle noté $\boldsymbol{\theta}$ sont $\boldsymbol{\theta} = (\boldsymbol{V}_1, \ldots, \boldsymbol{V}_H, \boldsymbol{R}, \boldsymbol{\gamma}, \boldsymbol{\nu}_1^1, ..., \boldsymbol{\nu}_{K_H}^H, \pi_1^1, \ldots, \pi_{K_1}^1, \ldots, \pi_{K_H}^H)$. Si pour tout h on a $p_h \leq K_h - 1$, alors le modèle proposé est identifiable à une transformation orthonormale prés des matrices $\boldsymbol{V}_1$, ..., $\boldsymbol{V}_H$ et $\boldsymbol{R}$. Dans le cas non supervisée, le

modèle est défini à une permutation près des variables classifiantes et à une permutation près des classes.

3 Estimation des paramètres

Les paramètres du modèle sont estimés par maximum de vraisemblance. Dans un premier temps on présente l'estimation dans le cas supervisé, et dans un second temps l'estimation dans le cas non supervisé.

3.1 Cas supervisé

La vraisemblance s'écrit :

$$\begin{aligned}\ell(\boldsymbol{\theta}; \mathbf{x}, \mathbf{z}) &= n \log \left| \det \begin{pmatrix} \boldsymbol{V}_1 \\ \vdots \\ \boldsymbol{V}_H \\ \boldsymbol{R} \end{pmatrix} \right| - \sum_{i=1}^{n} \sum_{h=1}^{H} \sum_{k=1}^{K_h} z_{ik}^h \|\boldsymbol{V}_h^{\mathrm{T}} \boldsymbol{x}_i - \boldsymbol{\nu}_k^h\|^2 \\ &\quad + \sum_{i=1}^{n} \sum_{h=1}^{H} \sum_{k=1}^{K_h} z_{ik}^h \log(\pi_k^h) - \sum_{i=1}^{n} \|\boldsymbol{R}^{\mathrm{T}} \boldsymbol{x}_i - \boldsymbol{\gamma}\|^2 - \frac{n}{2} \log(2\pi). \end{aligned} \tag{1}$$

La vraisemblance ne peut pas être maximisée directement. Cependant dans les cas où $H = 1$ le problème se réduit au problème de l'analyse discriminante linéaire sous une contrainte de rang sur la matrice des centres dont la solution est explicite (voir Campbell (1984)).

Ici nous proposons une approche d'optimisation alternée. Supposons tous les paramètres fixés à l'exception de $\boldsymbol{V}_h$, $\boldsymbol{R}$, $\boldsymbol{\nu}_1^h$, …, $\boldsymbol{\nu}_{K_h}^h$, π_1^h, …, $\pi_{K_h}^h$ et $\boldsymbol{\gamma}$. A l'itération $q+1$, si on contraint $\boldsymbol{V}_h^{(q+1)}$ and $\boldsymbol{R}^{(q+1)}$ à être des combinaison linéaires de $\boldsymbol{V}_h^{(q)}$ et $\boldsymbol{R}^{(q)}$, alors on se ramène au problème de l'analyse discriminante linéaire sous une contrainte de rang sur la matrice des centres.

Soit $\boldsymbol{M} \in \mathcal{M}_{d-p_\bullet+p_h, d-p_\bullet+p_h}(\mathbb{R})$ la matrice permettant le calcul de $\boldsymbol{V}_h^{(q+1)}$ et $\boldsymbol{R}^{(q+1)}$ à partir de $\boldsymbol{V}_h^{(q)}$ et $\boldsymbol{R}^{(q)}$:

$$\begin{pmatrix} \boldsymbol{V}_h^{(q+1)} \\ \boldsymbol{R}^{(q+1)} \end{pmatrix} = \boldsymbol{M} \begin{pmatrix} \boldsymbol{V}_h^{(q)} \\ \boldsymbol{R}^{(q)} \end{pmatrix} = \begin{pmatrix} \boldsymbol{M}_1 \\ \boldsymbol{M}_2 \end{pmatrix} \begin{pmatrix} \boldsymbol{V}_h^{(q)} \\ \boldsymbol{R}^{(q)} \end{pmatrix},$$

où $\boldsymbol{M}_1$ est la sous-matrice contenant les p_h premières lignes de $\boldsymbol{M}$ et $\boldsymbol{M}_2$ la matrice contenant les $d - p_\bullet$ dernières lignes de $\boldsymbol{M}$.

En notant $\begin{pmatrix} \boldsymbol{y}_i^{h(q)} \\ \boldsymbol{u}_i{}^{(q)} \end{pmatrix} = \begin{pmatrix} \boldsymbol{V}_h^{(q)} \\ \boldsymbol{R}^{(q)} \end{pmatrix} \boldsymbol{x}_i$, la maximisation de la vraisemblance en $\boldsymbol{M}$, $\boldsymbol{\nu}_1^h$, …, $\boldsymbol{\nu}_{K_h}^h$, π_1^h, …, $\pi_{K_h}^h$ et $\boldsymbol{\gamma}$ tous les autres paramètres étant fixés par ailleurs se réduit à une analyse linéaire discriminante sous contrainte sur les centres des données $(\boldsymbol{y}_i^{h(q)^{\mathrm{T}}}, \boldsymbol{u}_i{}^{(q)^{\mathrm{T}}})$.

Afin d'optimiser la vraisemblance sur l'ensemble des paramètres, nous répétons cette approche pour chacune des dimensions classifiantes jusqu'à convergence de l'algorithme. La procédure permet d'améliorer la vraisemblance à chaque étape, et la convergence est obtenue au bout de quelques itérations en pratique. Cependant, puisque la log-vraisemblance n'est pas concave nous n'avons pas de garantie de trouver l'optimum global et il est recommandé d'initialiser l'algorithme à partir de plusieurs valeurs différentes.

Ce modèle peut être utilisé dans le cadre supervisé pour la visualisation de données mixtes. Cependant notre motivation principale dans cette présentation est le cadre non supervisé.

3.2 Cas non supervisé

Ici on se place dans le cas où $\boldsymbol{z}_i^1, \ldots, \boldsymbol{z}_i^H$ sont inconnus. Par conséquent nous allons utiliser l'algorithme EM pour « reconstituer »les variables de classes manquantes. L'algorithme reste ici très similaire au cas supervisé à la différence que les données sont maintenant pondérées par $t_{ik}^{h\,(q+1)} = p(z_{ik}^h = 1|\boldsymbol{x}_i; \boldsymbol{\theta}^{(q)})$ au lieu de z_{ik}^h.

L'algorithme est le suivant :

— Jusqu'à convergence
  — Pour $h \in \{1, \ldots, H\}$
    — **Etape E :** calculer

$$t_{ik}^{h\,(q+1)} = \frac{\pi_k^h \phi_{p_h}(\boldsymbol{y}_i^{h(q)}; \boldsymbol{\nu}_k^{h(q)}, \boldsymbol{I}_{p_h})}{\sum_{k'=1}^{K} \pi_{k'}^h \phi_{p_h}(\boldsymbol{y}_i^{h(q)}; \boldsymbol{\nu}_{k'}^{h\,(q)}, \boldsymbol{I}_{p_h})}$$

avec $\phi_p(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ la densité de probabilité de la loi normale p-variée d'espérance $\boldsymbol{\mu}$ et de matrice de variance covariance $\boldsymbol{\Sigma}$.

    — **Etape M :** calculer $\pi_1^{h(q+1)}, \ldots, \pi_1^{H(q+1)}$, $\boldsymbol{V}_h^{(q+1)}$, $\boldsymbol{R}^{(q+1)}$, $\boldsymbol{\gamma}^{(q+1)}$ et $\boldsymbol{\nu}_1^{h(q+1)}$, $\ldots$, $\boldsymbol{\nu}_{K_h}^{h\,(q+1)}$ de manière similaire au cas supervisé en utilisant les poids $t_{ik}^{h\,(q+1)}$

A proprement parler l'algorithme présenté ici est un algorithme EM généralisé (GEM) puisque qu'à chaque étape il ne maximise pas l'espérance de la vraisemblance complétée, mais se contente de la faire croître. Comme tout algorithme EM il est sensible à l'initialisation. Celle-ci peut par exemple être effectuée à partir de projections aléatoires, puis par un *clustering* sur chacune de celles-ci. Comme pour tout algorithme, la vitesse de convergence de l'algorithme peut être plutôt lente en pratique.

4 Illustration sur les données crabes

On considère ici les données crabes issue de Campbell & Mahon (1974). Ces données représentent les mesures de 5 variables morphologiques sur 200 crabes. Parmi ces crabes on

a 50 mâles orange, 50 mâles bleus, 50 femelles orange, 50 femelles bleues. En classification non supervisée on aimerait retrouver une variables de classe représentant le sexe des crabes, une autre représentant la sous-espèce des crabes. De plus le modèle nous permet aussi d'obtenir le couple d'axe associé à ces deux variables de classes.

Les données sur les trois premiers axes de l'ACP normée sont présentés figure 2. On voit que le premier axe de l'ACP ne parvient pas bien à discriminer les classes, tandis que les axes 2 et 3 permettent respectivement de bien discriminer le sexe et la sous-espèce.

Dans l'approche proposée on ajuste le modèle non-supervisé avec avec $H = 2$ variables de classes comportant chacune 2 classes, et responsable chacune d'une variable classifiante en dimension 1 ($p_1 = p_2 = 1$). On représente figure 3 la visualisation obtenue en utilisant le modèle proposé. On retrouve clairement sur un premier axe la séparation selon le sexe et sur un second axe la séparation selon la sous-espèce. Ainsi le modèle permet de trouver simultanément des axes classifiants et de visualiser les données et les classes obtenues.

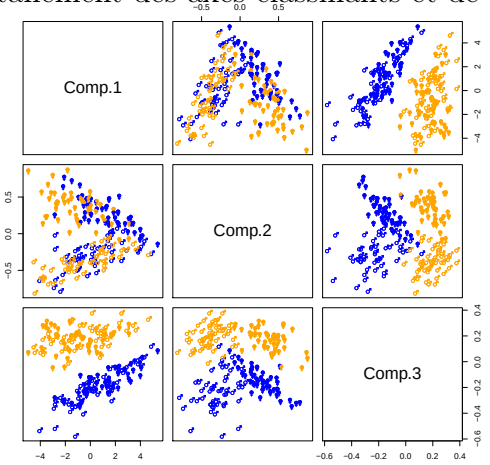

FIGURE 2: Données crabes sur les trois premiers axes de l'ACP

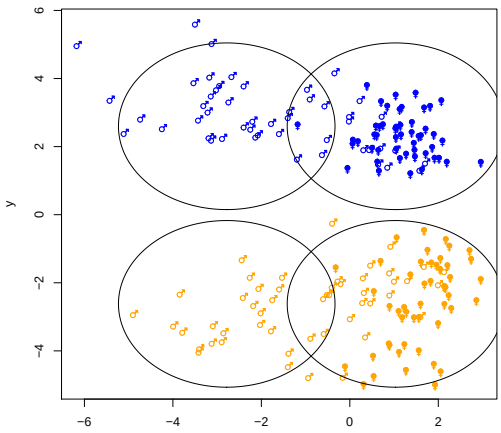

FIGURE 3: Données crabes sur les deux composantes classifiantes

Bibliographie

[1] McLachlan, G. and Peel, D. (2004), *Finite mixture models*, John Wiley & Sons.
[2] Bouveyron, C. and Brunet, C. (2012), Simultaneous model-based clustering and visualization in the fisher discriminative subspace, *Statistics and Computing*, 22(1) :301–324.
[3] Campbell, N. A. (1984), Canonical variate analysis - a general model formulation, *Australian Journal of Statistics*, 26(1) :86–96.
[4] Campbell, N. A., & Mahon, R. J. (1974), A multivariate study of variation in two species of rock crab of the genus Leptograpsus,*Australian Journal of Zoology*, 22(3), 417-425.