Etude comparee de classifications sur matrices tres creuses et de grandes dimensions

Gettler Summa Mireille[1] & Palumbo Francesco[2] & Tortora Cristina[3]

[1] *Université Paris-Dauphine, CEREMADE , Place du Maréchal De Lattre De Tassigny 75775 PARIS CEDEX 16 - FRANCE*

[2] *Dip. Scienze Relazionali "Gustavo Iacono" Université de Naples Federico II, via Porta di Massa 1, 80133 Naples Italie*

[3] *Dip. Matematica e Statistica Université de Naples Federico II, Complesso Monte Sant Angelo via Cinthia 26, 80126 Naples Italie*

Résumé

Les méthodes de classification non supervisée ont pour but de révéler une structure entre des éléments, selon les associations qu'on peut y détecter par leurs valeurs sur un ensemble de variables. Le processus permet selon un critère prédéfini de regrouper en classes les unités qui sont homogènes à l'intérieur d'une même classe, et le mieux séparées possibles de ceux d'une classe distincte. Lorsque l'on s'intéresse à des grands ensembles d'éléments, il est nécessaire d'en réduire la dimensionnalité avant le processus de classification. Lorsque les associations considérées entre les variables sont linéaires, grâce à des transformations adéquates des variables initiales ou bien à des bon choix de mesures des liens, on obtient des solutions satisfaisantes pour les problèmes de partitionnement (Saporta 1990). Mais lorsque les variables présentent plutôt des liens non linéaires, les mêmes approches sont inopérantes. Les classifications supervisées et non supervisées de variables qualitatives soulèvent dans ce sens de nombreux problèmes ; de telles variables peuvent en effet être combinées de façon à se limiter à un sous espace de l'espace de départ, mais les associations restent en général non linéaires. En fait le point délicat est que lorsque l'on procède à un recodage binaire de l'ensemble des modalités des variables, on obtient le plus souvent des matrices très creuses. Il y a classiquement deux manières de contourner la situation; l'une est de transformer les variables qualitatives en variables continues, puis de faire la classification sur les valeurs de ces dernières pour récupérer une structure sur les éléments de l'étude, l'autre consiste à mettre en œuvre des mesures d'appariement non métriques (Lenca et al., 2008). Il faut noter que ces mesures sont d'autant moins intéressantes que le nombre de variables est important. Notre travail s'attache à classifier de façon non supervisée des variables qualitatives dans le contexte général suivant: il n'y a pas de liens linéaires entre les variables et elles sont en grand nombre. Nous proposons une approche en plusieurs étapes dont voici les principales :

Analyse Factorielle pour réduire la dimension de l'espace de travail (par exemple une Analyse Factorielle des Correspondances Multiples), redéploiement des coordonnées des premiers axes factoriels dans un espace de plus grande dimension (comme dans le cas des approches par vecteurs de support) (Abe 2005), construction des classes dans ce dernier nouvel espace, enfin visualisation par projection des classes obtenues dans l'espace des facteurs. On appliquera cette approche aux données 'epub' du 'CRAN-R', et nous nous intéresserons sur cet exemple à la comparaison entre l'approche par le détour des vecteurs de support, (Ben-Hur 2001, Lee 2006) et celle classique d'un arbre hiérarchique. On choisit de classifier les 283 variables binaires connues pour 1108 individus.

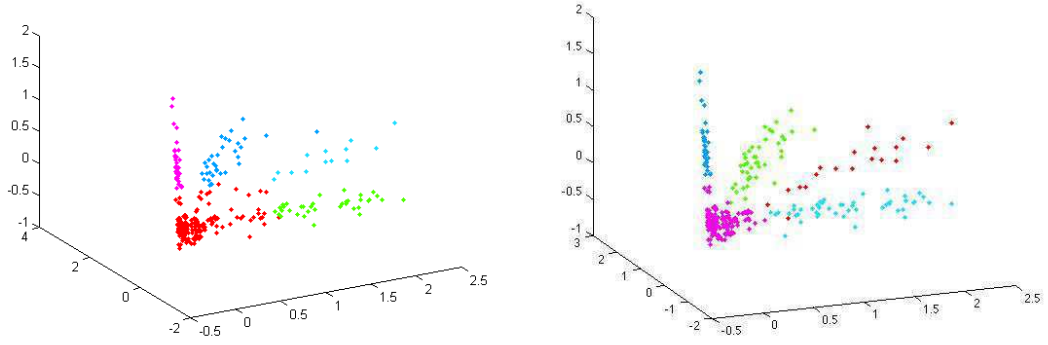

Figure 1: Classes obtenues par SVC et par CAH dans l'espace des trois premiers facteurs.

	SVC	CAH
CATANOVA	$566,5$	$562,57$
variance intra classes	$52,74$	$52,53$

Table 1: Qualité de la classification

Les graphiques représentent les coordonnées des variables du jeu des données 'epub' sur le premiers trois axes de l'ACM. On a choisi trois axes parce que ils représentent 60% environ de la variabilité totale, après correction de Benzécri (Le Roux 2010). On peut considérer que les autres facteurs représentent du bruit du fond. Avec l'ACM on a réduit la dimension du problème, initialement de 1108, et on n'a pas perdu les associations non linéaires entre les variables. On applique ensuite deux algorithmes, SVC et CAH sur ces coordonnées. On utilise le noyau polynomial pour SVC (Muller 2001, Shawe-Taylor 2004) et l'indice d'aggrégation de Ward pour CAH. Les graphiques obtenus par SVC (à gauche) et par CAH (à droite) classifient les données en 5 groupes, les classes trouvées sont de formes allongées et de fait difficiles à identifier par les méthodes classiques comme les k-means. Pour comparer les deux méthodes on a calculé la variance intra classes et

l'index C du CATANOVA (Singh 1993), les valeurs sont reportées dans le tableau 1. On peut constater d'une part que les deux méthodes classifient bien les données, d'autre part que la variance intra classes est légèrement meilleure pour la CAH, mais que le valeur du C est légèrement meilleure pour SVC.

Abstract

Cluster Analysis is a multivariate analysis technique that organizes information about variables so that homogeneous groups, or "clusters", can be formed and well separated one from the other. In other words, clustering algorithms aim at finding homogeneous groups with respect to their association structure among variables. Proximity measures or distances can be properly used to separate homogenous groups.
Dealing with large dataset it is necessary to reduce the dimensionality of the problem before applying clustering algorithms. When there is a linear association between variables, suitable transformations of the original variables or proper distance measures leads to satisfactory solutions (Saporta 1990). However when data are characterized by non-linear associations the actual cluster structure remains invisible to these approaches.

Categorical data clustering and classification present well known issues. Categorical data can be combined in order to define a limited subspace of the global data space: this type of data are thus characterized by non-linear associations. Moreover when dealing with variables having different number of categories, the usually adopted complete binary coding leads to very sparse binary data matrices. There are two main strategies to cope with the clustering of categorical data: *i*) transforming categorical variables into continuous ones and then performing a clustering process on the transformed variables; *ii*) adopting non-metric matching measures (Lenca et *al.*, 2008). It is worth noticing that matching measures become less effective as the number of variables increases.

This paper focuses the attention on the cluster analysis for categorical data under the following general hypotheses: there are nonlinear associations between variables and the number of variables is quite large. In this framework we propose a clustering approach based on a multistep strategy: *i*) Factor Analysis on the raw data matrix; *ii*) projection of the first factors' coordinates into a higher dimensional space; *iii*) clusters identification in the high dimensional space; *iv*) clusters visualisation in the factorial space.

We compare the performance of our method with the one of a hierarchical clustering algorithm, on the 'epub' dataset that belongs to 'CRAN.R'. We have chosen to classify the 283 variables whitch describe 1108 units.

The graphic in figure 1 represents the coordinates of the 'epub' dataset variables on the first factorial MCA's axes. We have chosen 3 factorial axes because they explain 60% of the variability, using Benzecrì (Le Roux 2010) correction. Starting from the fourth factor down to the last we may consider they represent ground noise. MCA factors preserve non-linear relations between data and reduce the dimensionality of the problem. After we have applied SVC and HAC on the factorial axes. We have chosen a polynomial kernel

for SVC and a Ward linkage for HAC. Figure 1 shows 5 clusters obtained with SVC (right) and HAC (left). Clusters are of arbitrary form, classical clustering algorithms, like k-means, wouldn't be able to find this type of clusters. In order to evaluate the better performing procedure we computed within variance and CATANOVA index (Singh 1993), results are shown in table 1. We obtain a good clustering structure using both methods, within variance is lower using HAC but index C is better using SVC.

Mots clés: Apprentissage et classification, Analyse des données - data mining

Bibliographie

[1] Abe S. (2005), Support vector machine for pattern classification, Springer.
[2] Ben-hur A., Horn D., Siegelmann H. T., Vapnik, V. (2001), Support vector clustering, Journal of machine learning research.
[3] Lee S.H., Danels K.M. (2006), Cone cluster labeling for support vector clustering, Proceeding of the sixth SIAM international conference on data mining, Bethesda.
[4] Lenca P., Patrick M., Benoòt V., Stèphane L. (2008), On selecting interestingness measures for association rules: User oriented description and multiple criteria decision aid, European Journal of Operational Research,184, 610–626.
[5] Le Roux B., Rouanet H., (2010), Multiple correspondence analysis Series: Qunatitative Applications in the Social Sciences. SAGE.
[6] Muller K. R., Mika S., Ratsch G., Tsuda K., Scholkopf B. (2001), An introduction to Kernel-based learning algorithms, IEEE transiction on neural networks 12.
[7] Saporta G. (1990), Simultaneous analysis of qualitative and quantitative data, Società italiana di Statistica, CEDAM.
[8] Shawe-Taylor J., Cristianini N. (2004), Kernel methods for pattern analysis, Cambridge University press.
[9] Singh, B. (1993), On the Analysis of Variance Method for Nominal Data,The Indian Journal of Statistics, Series B