

Clustering in feature space for interesting pattern identification of categorical data

Marina Marino¹, Francesco Palumbo², Cristina Tortora³

¹ Dip. di Ingegneria Agraria e Agronomia del Territorio, Univ. di Napoli Federico II

² Dip. di Istituzioni Economiche e Finanziarie, Università di Macerata

³ Dip. di Matematica e Statistica, Università di Napoli Federico II

e-mail: marina.marino@unina.it, palumbo@unimc.it, cristina.tortora@unina.it

Abstract Standard clustering methods fail when data are characterized by non-linear associations. A suitable solution consists in mapping data in a higher dimensional feature space where clusters are separable. The aim of the present contribution is to propose a new technique in this context to identify interesting patterns in large datasets.

Keywords: kernel method, machine learning, Support vector clustering.

1. Introduction

Cluster Analysis is a multivariate analysis technique that seeks to organize information about variables so that homogeneous groups, or “clusters”, can be formed. 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.

In the case of continuous variables, clustering problem consists in minimising the sum of the variances within clusters; if variables are linearly independent this is equivalent to minimise the sum of the squared Euclidean distances within classes. When there is linear association between variables, suitable transformations of the original variables or proper distance measures permit to get satisfactory solutions (Saporta 1990). However when data are characterised by non-linear association the actual cluster structure remains invisible to these approaches.

Categorical data clustering and classification present well known issues. Categorical variables are characterised by non-linear association. 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 in presence of categorical data: *i*) to transform categorical variables into continuous ones and then to perform clustering on the transformed variables; *ii*) to adopt 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 is nonlinear association 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 factor coordinates into a higher dimensional space; *iii*) clusters identification in the high dimensional space; *iv*) clusters visualisation in the factorial space.

2. Support Vector Clustering on MCA Factors

The core of the proposed approach consists of steps *i* and *ii* indicated at the end of the previous section. This section aims at motivating the synergic advantage of this mixed strategy.

When the number of variables is large, projecting data into a higher dimensional space is a self-defeating and computationally unfeasible task. In order to carry only the significant association structure in the analysis, dealing with continuous variables, some authors propose to perform a Principal Component Analysis on the raw data, and then to project the first components in a higher dimensional feature space (Ben-Hur et al., 2002). In the case of categorical variables, the dimensionality depends on the whole number of categories, this implies an even more dramatic problem of sparseness. Moreover, as modalities are a finite number, the association between variables is non-linear.

Multiple Correspondence Analysis (MCA) on the raw data matrix permits to reduce the number of variables, and to combine the categorical variables into continuous variables that preserve the non-linear association structure. Let us indicate with $\mathbf{Y}$ the $n \times q$ coordinates matrix of n points into the orthogonal space spanned by the first q MCA factors. For sake of brevity we do not go into the MCA details. Interested readers are referred to Greenacre's book (2006). Mapping the first factorial coordinates into a feature space permits to cluster data via a Support Vector Clustering approach.

Support Vector Clustering (SVC) is a non parametric cluster method based on support vector machine that maps data points from the original variable space to a higher dimensional feature space through a proper kernel function (Muller et al., 2002).

A feature space is an abstract t -dimensional space where each pattern sample is represented as a point. Given an observations $\times$ variables data matrix $\mathbf{X}$ with general term x_{ij} and $i = 1, 2, \dots, n$ and $j = 1, 2, \dots, p$, any generic row or column vector of $\mathbf{X}$ can be represented into a feature space using a non linear mapping function. Formally, the generic column (row) vector $\mathbf{x}_j$ ($\mathbf{x}'_i$) of $\mathbf{X}$ is mapped into a higher dimensional space F through a function

$$\varphi(\mathbf{x}_j) = (\phi_1(\mathbf{x}_j), \phi_2(\mathbf{x}_j), \dots, \phi_t(\mathbf{x}_j)),$$

with $t > p$ ($t > n$ in the case of row vectors) and $t \in \mathbb{N}$.

The solution of the problem implies the identification of the minimal radius hypersphere that includes the images of all data points: $r^2 \geq \|\varphi(\mathbf{x}_j) - \mathbf{a}\|^2 \quad \forall j$, where $\mathbf{a}$ is the center of the sphere and $\|\cdot\|$ denotes the Euclidean norm. To avoid that only the most far point determines the solution, slack variables $\xi_j \geq 0$ can be added:

$$r^2 + \xi_j \geq \|\varphi(\mathbf{x}_j) - \mathbf{a}\|^2 \quad \forall j.$$

This problem can be solved by introducing the Lagrangian:

$$L(r, \mathbf{a}, \xi_j) = r^2 - \sum_j (r^2 + \xi_j - \|\varphi(\mathbf{x}_j) - \mathbf{a}\|^2) \beta_j - \sum_j \xi_j \mu_j + C \sum_j \xi_j, \quad (1)$$

where $\beta_j \geq 0$ and $\mu_j \geq 0$ are Lagrange multipliers, C is a constant and $C \sum_j \xi_j$ is a penalty term. To solve the minimization problem we set to zero the derivative of L with respect to r , $\mathbf{a}$ and ξ_j and we get the following solutions:

$$\sum_j \beta_j = 1$$

$$\begin{aligned}\mathbf{a} &= \sum_j \beta_j \varphi(\mathbf{x}_j) \\ \beta_j &= C - \mu_j\end{aligned}$$

We remind that Karush-Kuhn-Tucker complementary condition implies:

$$\begin{aligned}\xi_j \mu_j &= 0 \\ (r^2 + \xi_j - \|\varphi(\mathbf{x}_j) - \mathbf{a}\|^2) \beta_j &= 0\end{aligned}$$

The Lagrangian is a function of r , $\mathbf{a}$ and μ_j . Turning the Lagrangian into the more simple Wolfe dual form, which is a function of the variables β_j , we obtain

$$W = \sum_j \varphi(\mathbf{x}_j)^2 \beta_j - \sum_{j,j'} \beta_j \beta_{j'} \varphi(\mathbf{x}_j) \cdot \varphi(\mathbf{x}_{j'}), \quad (2)$$

with the constraints $0 \leq \beta_j \leq C \quad \forall \{j, j'\}$.

It is worth noticing that in 2 the function $\varphi(\cdot)$ only appear in products. The dot products $\varphi(\mathbf{x}_j) \cdot \varphi(\mathbf{x}_{j'})$ can be computed using an appropriate kernel function $K(\mathbf{x}_j, \mathbf{x}_{j'})$. The Lagrangian W is now written as: $W = \sum_j K(\mathbf{x}_j, \mathbf{x}_j) \beta_j - \sum_{j,j'} \beta_j \beta_{j'} K(\mathbf{x}_j, \mathbf{x}_{j'})$. The SVC problem requires the choice of a kernel function and a respective suitable parameter α . Points, whose distance from the surface of the hypersphere is less then ξ , are called *support vectors* and they define a partition of the feature space.

The choice of the kernel function remains a still open issue (Shave-Taylor and Cristianini 2004). There are several proposal in the recent literature: *Gaussian* and *polynomial* are among the most largely used functions. In the present work we adopt a polynomial kernel function; the choice was based on the empirical comparison of the results (Abe 2005). Most important for the final clustering result is the choice of the parameter α , in fact this parameter affects the number of clusters.

To simplify the notation, we indicate with $K^*(\cdot)$ the parametrised kernel function: then in our specific context the problem consists in maximising the following quantity with respect to β

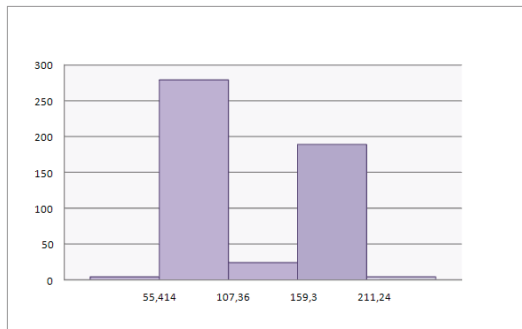
$$W = \sum_m K^*(\mathbf{y}_m, \mathbf{y}_m) \beta_m - \sum_{m,m'} \beta_j \beta_{m'} K^*(\mathbf{y}_m, \mathbf{y}_{m'}), \quad (3)$$

where $\mathbf{y}_m$ represents the generic coordinate obtained via MCA, and $1 \leq m \leq q$.

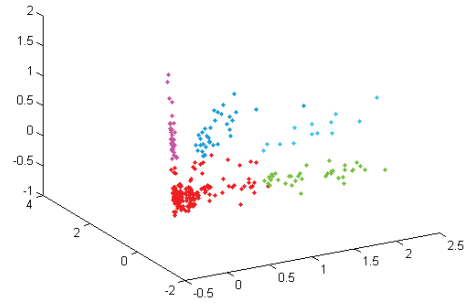
The last clustering phase consists in assigning the points projected in the feature space to the classes. It is worth reminding that the analytic form of $\{\phi_1(\mathbf{x}), \phi_2(\mathbf{x}), \dots, \phi_t(\mathbf{x})\}$ is unknown, so that computing points coordinates in the feature space is an unfeasible task. Alternative approaches permit to define the points memberships without computing all coordinates. In this paper, in order to assign points to clusters we use the *cone cluster labeling algorithm* (Lee and Daniels 2006).

3. Empirical evidence and concluding remarks

The method has been applied on the ‘epub’ dataset that is distributed with ‘arules’ dataset on cran-R. The dataset contains the documents download history from the electronic publication platform of the Vienna University of Economics and Business Administration. Data consists of 465 binary attributes and 3975 sequences, attributes with less then



[a]



[b]

Figure 1: **a)** Within variance on 500 iteration of *k*-means algorithm. **b)** Cluster obtained with SVC algorithm.

five occurrences have been discarded, so that the final dataset is composed by 1024 observations and 283 publications. In the first step we used an MCA algorithm in order to reduce the number of variables. We can see that, starting from the fourth factor the growth of the explained variability is minimal. For this reason we computed the coordinates of the variables on the first three MCA factors that explain the 40% of the variability. In the second step we used the *k*-means method to cluster the coordinates on factorial axes, but the results did not converge into global minimum. The quality of the classification has been computed by within variance, we reiterate the algorithm 500 times. In the best case the within variance is 54.18, it happened in the 0.6% of cases.

Using SVC algorithm the clustering converge into global minimum. We chose a polynomial kernel because it is better performing this dataset. The choice of the parameter influences the number of classes: with a parameter $\alpha = 3$ we obtained 5 classes. Within variance is 52.74, lower than using *k*-means method, so SVC performs better with this type of data.

References

- Abe S. (2005) *Support vector machine for pattern classification*, Springer.
- Ben-hur A., Horn D., Siegelmann H. T., Vapnik, V. (2001) *Support vector clustering*, Journal of machine learning research.
- Lenca P., Patrick M., Benoît V., Stéphane L. (2008) *On selecting interestingness measures for association rules: User oriented description and multiple criteria decision aid*, European Journal of Operational Research.
- 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.
- Muller K. R., Mika S., Ratsch G., Tsuda K., Scholkopf B. (2001) *An introduction to Kernel-based learning algorithms*, IEEE transaction on neural networks 12.
- Saporta G. (1990) *Simultaneous analysis of qualitative and quantitative data*, Società italiana di Statistica, CEDAM.
- Shawe-Taylor J., Cristianini N. (2004) *Kernel methods for pattern analysis*, Cambridge University press.