

Robust Classification of E-government Data ^(*)

Andrea Cerioli and Marco Riani

Dipartimento di Economia, Sezione di Statistica e Informatica, Università di Parma

e-mail: andrea.cerioli@unipr.it

Riassunto: In questo lavoro si considera il problema della classificazione non supervisionata (*cluster analysis*) in presenza di variabili qualitative. Tale problema, che è tipico nel contesto della valutazione dei servizi, è affrontato attraverso l'approccio robusto denominato *forward search*.

Keywords: cluster analysis, dissimilarity, e-government, forward search, random starts.

1. Introduction

In spite of its practical relevance, clustering of discrete multivariate observations has received relatively little attention within the classification literature. The traditional approach has been to compute suitable measures of pairwise dissimilarity, such as the simple matching coefficient, and then to use these measures as input for hierarchical clustering algorithms. Hierarchical agglomeration plays an important role also in the clustering algorithm of Friedman and Meulman (2004), which can cope with categorical information. The main problem with hierarchical algorithms is that they rapidly become computationally unacceptable and provide results that are difficult to represent as the number of objects grows. Therefore, they may be inappropriate for the analysis of large or even moderate data sets, like those encountered in marketing research or service quality evaluation. The *k*-modes algorithm of Huang (1998) and Chaturvedi *et al.* (2001) is a notable exception trying to combine the efficiency of the standard *k*-means paradigm for interval data with the need to take categorical information into account. This is accomplished by running a *k*-means type algorithm with simple matching dissimilarities instead of Euclidean distances and cluster modes instead of means. However, similarly to *k*-means, the results from *k*-modes can be quite sensitive to the choice of the starting solution and even to the order of the observations in the data set (Huang, 1998). An additional shortcoming is that cluster modes may not be uniquely defined at some steps of the iterative procedure, thus leading to indeterminacy in the clustering solution.

The goal of this paper is to address the issue of clustering data sets with categorical information through the robust forward search approach. Our method is computationally affordable and provides an assessment of the impact of each observation on the fitted clustering solution. It also helps to shed light on the actual number of clusters in the data, a critical issue with most, if not all, partitioning techniques. The suggested method is sketched in §2. It is then shown in action in §3, in the context of evaluation of e-government services.

(*) Work partially supported by MIUR Grants “Nuovi metodi statistici multivariati di classificazione e riduzione dimensionale per la valutazione e la customer satisfaction nei servizi”, and “Stima parametrica e non parametrica e previsione delle dinamiche di momenti condizionali di serie storiche”.

2. The forward search for clustering binary data

The forward search is a powerful general method for detecting unidentified subsets and multiple masked outliers and for determining their effect on models fitted to the data. The search for multivariate data, including cluster analysis with quantitative variables, is given book length treatment by Atkinson *et al.* (2004). The basic idea is to start from a small, robustly chosen, subset of the data and to fit subsets of increasing size, in such a way that outliers and subsets of data not following the general structure are clearly revealed by diagnostic monitoring. Searches from more than one starting point are often needed to reveal the clustering structure. Atkinson *et al.* (2006) demonstrate the usefulness of starting the search from randomly selected subsets that avoid any preliminary data analysis in the case of several continuous populations. In this paper we provide evidence of multiple random starts from subsets of discrete multivariate observations.

To gain focus, we restrict attention to the case of v binary variables $X_1, X_2, \dots, X_v$, the situation to be analyzed in §3. The more general case of v nominal variables can be dealt with a similar approach, but will not be considered in this paper. Let $S = \{u_1, u_2, \dots, u_n\}$ be a set of n units. Unit u_i is represented as $[x_{i1}, x_{i2}, \dots, x_{iv}]'$, where $x_{ij} \in \{0, 1\}$ is the observed category of variable X_j in unit u_i . We evaluate the dissimilarity between pairs of units through

$$d(u_i, u_l) = \sum_{j=1}^v w_j (x_{ij} - x_{lj})^2, \quad i, l = 1, \dots, n, \quad (1)$$

where $w_j \geq 0$ is the weight given to variable X_j in the classification process. Popular choices are either equal variable weighting, i.e. $w_j = 1$ for $j = 1, \dots, v$, or

$$w_j = \{\bar{x}_j(1 - \bar{x}_j)\}^{-1}, \quad (2)$$

where $\bar{x}_j = \sum_{i=1}^n x_{ij}/n$ is the proportion of units in S for which $X_j = 1$. Given the binary nature of $X_1, X_2, \dots, X_v$, $d(u_i, u_l)$ is the (weighted) simple matching coefficient between u_i and u_l . Similarly, for $0 < \bar{x}_j < 1$, (2) is the inverse of the within-sample variance of X_j , a scaling measure adopted by Friedman and Meulman (2004) among others.

The (sample) mode is a common measure of central tendency with categorical data. The sample mode of the binary variable X_j is defined as $\tilde{x}_j = 1$ if $\bar{x}_j > 0.5$, and $\tilde{x}_j = 0$ if $\bar{x}_j < 0.5$. The mode is undefined if $\bar{x}_j = 0.5$. Dealing with binary data, we also consider the sample mean $\bar{x}_j$ as a measure of central tendency of X_j . The sample mean estimates $E(X_{ij})$, for a Bernoulli random variable X_{ij} , and has the important advantage of being properly defined for each $n \geq 1$. We write $\hat{\mu} = [\hat{\mu}_1, \dots, \hat{\mu}_v]'$ for the multivariate centroid of $X_1, X_2, \dots, X_v$, where $\hat{\mu}_j$ stands for either $\tilde{x}_j$ or $\bar{x}_j$ depending on the context. Following (1), the dissimilarity between unit u_i and $\hat{\mu}$ is defined as

$$d_i = d(u_i, \hat{\mu}) = \sum_{j=1}^v w_j (x_{ij} - \hat{\mu}_j)^2.$$

In the forward search the multivariate centroid $\hat{\mu}$ is repeatedly computed based on a subset of m observations, $S(m)$ say, yielding the vector $\hat{\mu}(m) = [\hat{\mu}_1(m), \dots, \hat{\mu}_v(m)]'$. From this subset we obtain n dissimilarities

$$d_i(m) = d(u_i, \hat{\mu}(m)) = \sum_{j=1}^v w_j (x_{ij} - \hat{\mu}_j(m))^2 \quad i = 1, \dots, n. \quad (3)$$

We start with a subset of m_0 observations which grows in size during the search. When subset $S(m)$ is used in fitting, we order the dissimilarities (3) and take the units corresponding to the $m + 1$ smallest as the new subset $S(m + 1)$. To detect clusters, we look at forward plots of quantities derived from the dissimilarities (3). One of the most useful plots is that of the minimum dissimilarity amongst units not in the subset

$$d_{\min}(m) = \min d_i(m) \quad i \notin S(m). \quad (4)$$

If unit u_i does not come from the same population as the other m units in $S(m)$, the dissimilarity (4) will be large compared to the maximum dissimilarity within $S(m)$.

Simulation results provide evidence that the forward search approach outlined above performs well in recovering known clusters of discrete multivariate observations. We do not show these results here. However, we use them as a guidance for the interpretation of forward plots in the tricky real life application of the next section.

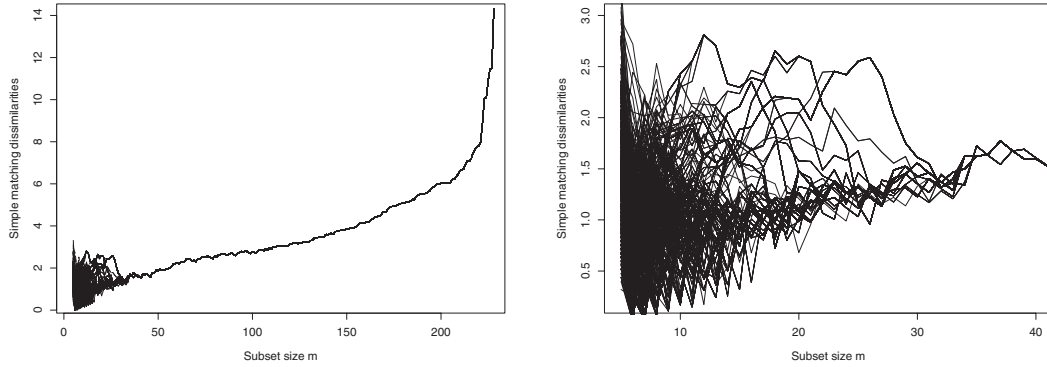
3. Application to e-government data

Our motivating example is the classification of the municipalities of Emilia-Romagna according to their activity on the domain of e-government. This classification can help policy-makers to evaluate the impact of e-government policies, following the significant investment of the past years. The data we analyze come from a questionnaire sent to all the municipalities of Emilia-Romagna in October 2004. The questionnaire aim was to measure the implementation of e-government technologies in a wide range of activities, from labour organization to on-line services for citizens. The data set consists of 229 municipalities, i.e. those filling the questionnaire and having a web site. Almost all the information was collected on a categorical scale. Here we focus on 34 binary indicators describing implementation of advanced e-government services in citizen-oriented activities, such as availability of interactive on-line services and e-democracy facilities. The number of implemented services ranges from 22 (in the case of Bologna) to only 1. A more detailed description of the data, together with further references and some preliminary analyses, can be found in Cerioli *et al.* (2005).

The left-hand panel of Figure 1 shows the results of 1000 forward searches from randomly selected starting subsets with $m_0 = 5$. For each search we have plotted the statistic $d_{\min}(m)$ defined in (4), with $\hat{\mu}_j = \hat{x}_j$ and $w_j = 1$. The first feature of the plot is that, from m around 40, all searches follow the same trajectory, regardless of the starting point. The end of the search shows a few possible outliers, without evidence of masking. Bologna and Modena, the two most remote units, are those with the highest number of e-government services. Other municipalities with good e-government facilities enter towards the end of the search, but without evidence of clustering. This means that these units are fairly well separated in the space spanned by the indicators, although they are far from the majority of the municipalities of Emilia-Romagna.

The plot also leads to identification of a few clusters giving rise to the peaks visible in some trajectories for $m < 40$. The peaks are perhaps best seen in the right-hand panel of Figure 1. They are typically originated by searches started in subsets of units with similar features. As there are not many such searches, we may anticipate a structure with some small and not very well separated clusters. For instance, consider the peak at $m = 12$. It is originated by a bunch of 25 searches including municipalities with a relatively large, but

Figure 1: E-government data: Forward plots of $d_{\min}(m)$, with $\hat{\mu}_j = \bar{x}_j$ and $w_j = 1$, for 1000 searches with random starting points (left-hand panel); a zoom taken for $m \leq 40$ (right-hand panel).



not extreme, number of e-government services. At $m = 12$, observations not belonging to this group have large dissimilarities from $\hat{\mu}(m)$, thus explaining the high value of $d_{\min}(m)$. The slow decay of the trajectory originating from this subset confirms a smooth transition from the cluster to the bulk of the data. Cluster membership is computed by identifying the units in the subsets just before the peaks. The searches that do not give rise to such peaks either contain observations from more than one group, or come from an unstructured population. Therefore, they do not provide evidence of clustering.

To better understand the cluster structure, further forward plots could be displayed. For instance, we monitor how individual distances (3) and proportions $\bar{x}_j(m)$ evolve for selected searches, such as those originating the peak at $m = 12$. These pictures (not shown) reinforce the idea of a large unstructured population to which most municipalities of Emilia-Romagna belong, plus a few relatively small and not well separated clusters added to such a “noisy” background, and of some “outstanding” but isolated municipalities.

References

- Atkinson A. C., Riani M., Cerioli A. (2004) *Exploring Multivariate Data with the Forward Search*, Springer, New York.
- Atkinson A. C., Riani M., Cerioli A. (2006) Random start forward searches with envelopes for detecting clusters in multivariate data, to appear in the series *Studies in Classification, Data Analysis, and Knowledge Organization*, Springer, Berlin.
- Cerioli A., Milioli M. A., Morlini I., Zani S. (2005) L’ICT nella pubblica amministrazione: un’applicazione ai comuni dell’Emilia-Romagna, in: *Atti della Riunione Scientifica su Valutazione e Customer Satisfaction per la Qualità dei Servizi*, Facoltà di Scienze Statistiche, Università di Roma, 65–68.
- Chaturvedi A., Green P. E., Carroll J. D. (2001) K-modes clustering, *Journal of Classification*, 18, 35–55.
- Friedman J. H., Meulman J. J. (2004) Clustering objects on subsets of attributes, *Journal of the Royal Statistical Society, B*, 66, 815–849.
- Huang Z. (1998) Extensions to the k-means algorithm for clustering large data sets with categorical values, *Data Mining and Knowledge Discovery*, 2, 283–304.