Issues on Clustering and Data Gridding

Jukka Heikkonen, Domenico Perrotta, Marco Riani, and Francesca Torti

Abstract This contribution addresses clustering issues in presence of densely populated data points with high degree of overlapping. In order to avoid the disturbing effects of high dense areas we suggest a technique that selects a point in each cell of a grid defined along the Principal Component axes of the data. The selected sub-sample removes the high density areas while preserving the general structure of the data. Once the clustering on the gridded data is produced, it is easy to classify the rest of the data with reliable and stable results. The good performance of the approach is shown on a complex dataset coming from international trade data.

1 Introduction

In this paper we address clustering issues in presence of data consisting of an unknown number of groups with high degree of overlapping and presenting both high and low density regions which invalidate the hypothesis of ellipticity.

J. Heikkonen (✉)
Department of Information Technology, University of Turku, Turku, Finland
e-mail: jukka.heikkonen@utu.fi

D. Perrotta
EC Joint Research Centre, Ispra site, Ispra, Italy
e-mail: domenico.perrotta@ec.europa.eu

M. Riani
University of Parma, Parma, Italy
e-mail: mriani@unipr.it

F. Torti
University of Milano Bicocca, Milan, Italy
e-mail: francesca.torti@unimib.it

A. Giusti et al. (eds.), *Classification and Data Mining*, Studies in Classification, Data Analysis, and Knowledge Organization, DOI 10.1007/978-3-642-28894-4_5,

To find the clusters on these data, the model complexity selection issue becomes difficult. Typically model complexity selection is based on the maximum likelihood formulation of the model with respect of the data and an additional cost function that penalises too complex models, i.e. the ones having too many parameters needed to capture the main characteristics of the data (Schwarz, 1978; Rissanen, 1986; Bishop, 2006). When the model complexity selection is formulated as a probabilistic problem, in addition to well known disturbing effects of noise and outliers, the presence of dense and dispersed groups of points causes additional challenges. Because of the likelihood formulation, the dense groups dictate the parameter values of the model and the groups with less points may not be properly detected. This has some similarity to sampling theory where the goal is to have representative samples from the population of interest (Cochran, 1977). Often some elements are hard to get or very costly to be obtained and one has to select the correct sampling strategy to obtain representative population statistics. For instance, in stratified sampling, the population is first divided by some meaningful rules into as homogeneous groups as possible. These groups (strata) should be mutually exclusive meaning that one element should be assigned only to one and only one group (stratum). When properly used, stratified sampling reduces sampling error, as is its goal.

In our clustering case we are interested in recognizing also those clusters that only consist of few data points. In order to achieve this goal, we propose a sampling approach that tries to avoid the disturbing effects of the dense populated data points through a data gridding technique based on Principal Component Analysis (PCA). This technique consists in defining a grid along the Principal Component (PC) axes of the data and selecting one point in each cell of the grid. Our goal is to have through the balanced data points correct model complexity for the given data and to avoid the domination of dense populated data points over the dispersed ones. The performance of the proposed data gridding technique (Sect. 2) is evaluated with two versions of Gaussian Mixture Models (GMMs) and the Forward Search (FS) on data coming from the international trade of the European Union (EU) (Sect. 3). We show that the gridded data preserve the general structure of the original dataset, which is then well captured by three clustering methods. We will see that once the clustering on the gridded data is produced, it is easier to classify the rest of the data with results which are more reliable and stable than those obtained on the original data.

To illustrate the procedure and the above problems we use the dataset in the left panel of Fig. 1. The variables are the volume and value of the monthly imports of a fishery product in the EU in a period of 3 years. One of the 27 EU Members States is producing trade flows which deviate from the main dense cluster. This situation, with a concentration of points towards the origin of the axes where the clusters intersect, is typical of international trade data. Perrotta and Torti (2009) made an exploratory analysis of this dataset and Riani et al. (2008) treated the case as regression problem. In this paper the emphasis is on inferring automatically the number and the shape of the groups. The dataset is included in the MATLAB FSDA toolbox Riani et al. (2012), which can be downloaded at http://www.riani.it/MATLAB.htm.

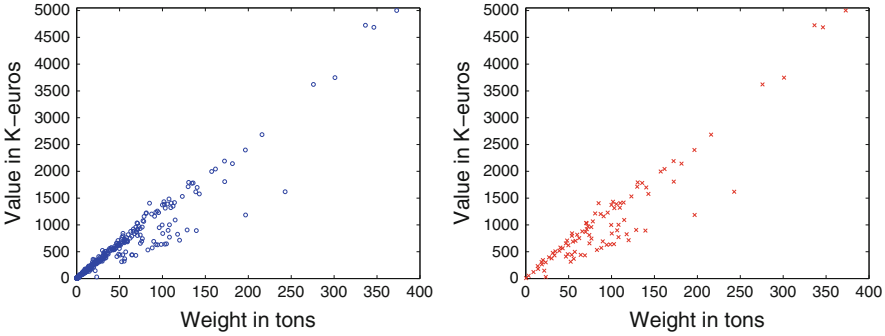

Fig. 1 Fishery data (*left plot*, 677 observations) and gridded data with 80 cells along the first principle component axis (*right plot*, 95 observations)

2 Gridding Approach

In our gridding approach the original variables are normalized to zero mean and unit variances to avoid the dominance of the scaling of variables in PCA. After scaling when the PC axes are defined, the data points are projected to this domain to have their PCs. Taking the maximum and minimum coordinates of the PCs we can define a grid of a predefined number of cells along each PC axis. In our case, when we have 2-dimensional data, the grid is also 2-dimensional and defined by the 2 PC axes. The same approach can be extended to higher dimensions. Note that the grid cells do not necessarily have equal width in all PC axes directions and a single cell can cover zero, one or more data points of the original data. Especially where the data are densely populated, the grid cells corresponding to a dense group of points include multiple original values. For each cell the goal is to search for one representative point from the original data. This is done by taking the median of points belonging to the cell and finding the closest point to the calculated median. As a result we obtain either none or one point for each cell of the grid. With the new reduced subset we can perform a desired analysis, for example estimating the Gaussian Mixture Model and the proper number of clusters over the balanced dataset.

The right panel of Fig. 1 shows the result of the gridding approach when 80 cells in each PC axis direction is applied to the original data of 677 observations. As can be observed, the gridded result represents rather well the original data and the number of points in dense and dispersed groups is better balanced.

3 Example Results

The GMM models used are Model-Based Clustering/Normal Mixture Modeling (Fraley, 1998) and its robust version Robust Trimmed Clustering (Garcia-Escudero et al., 2008). For the runs we used their well known R implementations MCLUST

and TCLUST. Both methods are based on a finite mixture of distributions where each mixture component corresponds to a different group. A common reference model for the components is the multivariate Gaussian distribution. In MCLUST the standard approach for estimating the mixture consists of using the EM algorithm and the BIC to select the number of components. Each observation is assigned to the cluster to which it is most likely to belong. The TCLUST approach is defined through the search of k centers $m_1, \ldots, m_k$ and k shape matrices $U_1, \ldots, U_k$ solving the double minimization problem:

$$\arg\min_{\mathbf{Y}} \min_{\substack{m_1,\ldots,m_k \\ U_1,\ldots,U_k}} \sum_{j=1,\ldots,k} (x_i - m_j)' U_j^{-1} (x_i - m_j) \qquad i = 1, \ldots, n \quad (1)$$

where $\mathbf{Y}$ ranges on the class of subsets of size $[n(1-\alpha)]$ within the sample $\{x_1, \ldots, x_n\}$. Note that in this approach we allow for a proportion α of observations, hopefully the most outlying ones, to be left unassigned. In order to chose k, the authors suggest using the so called Classification trimmed Likelihood curves (Garcia-Escudero et al., 2011).

The third method that we consider is based on the Forward Search of Atkinson et al. (2004). This approach was originally introduced for detecting subsets and masked outliers and for estimating their effect on the models fitted to the data. This method produces a sequence of subsets of increasing size through a dynamic process that leaves outliers in the last subsets. By monitoring the trajectory of the values of the minimum Mahalanobis distance among observations outside the current subset, it is possible to detect towards the end of the search peaks that correspond to the presence of outliers. Besides, by monitoring the same statistic for searches initialised from many different randomly chosen subsets, it is possible to reveal the presence of multiple populations as separated peaks that can occur at any position along the search depending on the size and structure of the groups (Atkinson and Riani, 2007). However, problems may occur in presence of high density areas. For example, the left panel of Fig. 1 shows that more than 50 % of the data are concentrated near the origin of the axes and, thus, the random start trajectories of minimum Mahalanobis distance degenerate into the same search path in the very first steps of the FS, as shown in Fig. 2. This behaviour, which is caused by the dense population near the origin of the axes, makes the information produced by the random start FS difficult or even impossible to interpret. The detection of the first cluster is shown in the zoom of Fig. 1, where the envelopes based on about 130 size sub-sample are exceeded by the Minimum Mahalanobis distance trajectories (Riani et al., 2009). The same procedure can be repeated iteratively for the data not yet assigned to an homogeneous subgroup. However, in this case this procedure results in an excessive number of subgroups.

In presence of highly dense areas, similar difficulties arise with other classical statistical clustering methods such as K-means clustering or GMMs which, however, compared to the FS have less or even no instruments to accurately identify the parts

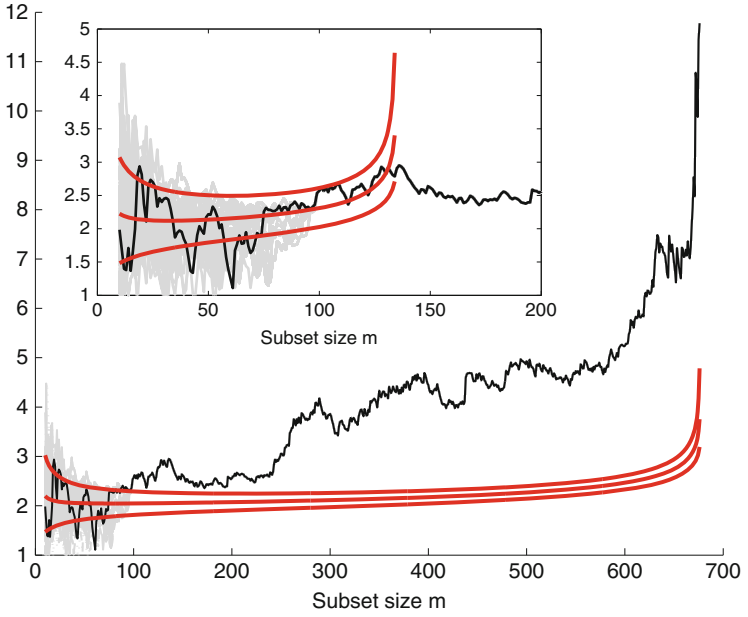

Fig. 2 Fishery dataset: 200 FS random starts and zoom on the initial part of the search (with superimposed envelopes)

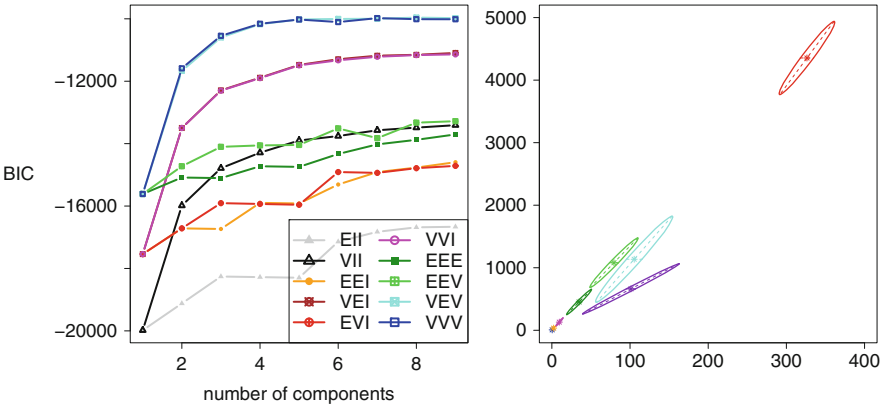

Fig. 3 Fishery dataset: BIC selection of the best MCLUST model (*left panel*) and ellipses associated with MCLUST classification (*right panel*)

of the data that are causing these issues. The left panel of Fig. 3 shows the BIC trajectories as a function of the number of groups k when using MCLUST. The highest value of BIC is obtained for $k = 8$. However, judging from the right-hand panel of the figure, which shows the ellipses associated to the eight components,

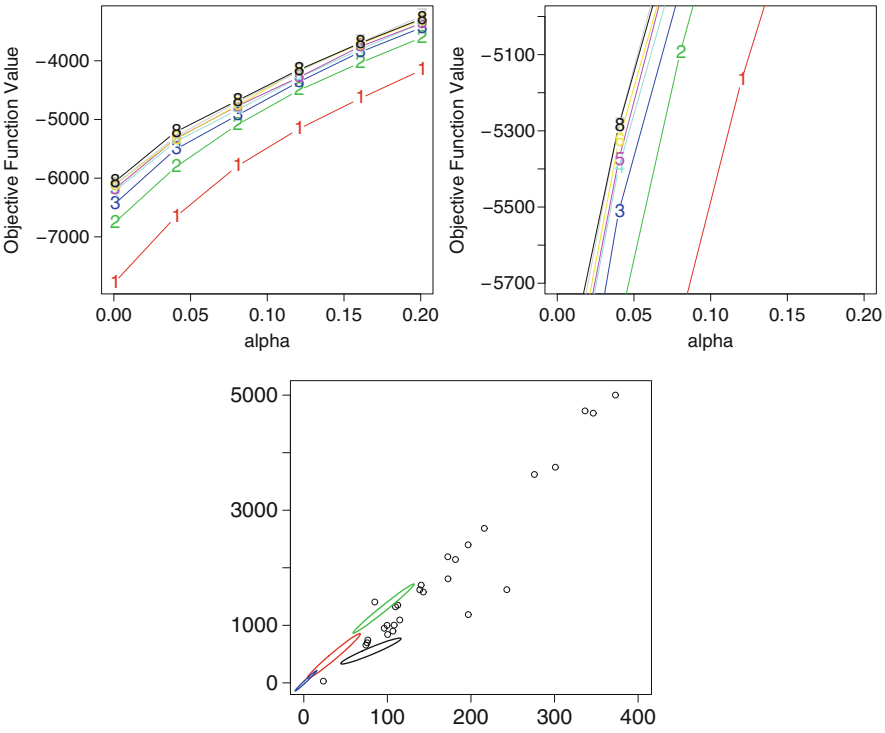

Fig. 4 Fishery dataset: Classification Trimmed Likelihood Curves for TCLUST (*upper left*), a zoom on an area of the classification plot (*upper right*) and ellipses associated with the TCLUST classification for $k = 4$ and a trimming proportion alpha $= 0.04$ (*bottom*). The unclassified data are plotted with *circles*

this number of clusters seems to be excessive. In fact the four components covering the dense area seem to be sub-samples of the same group.

Let us now consider the Gallego's algorithm (Garcia-Escudero et al., 2010), implemented in the TCLUST function in R-software, which is more sophisticated than MCLUST due to the possibility of data trimming. Based on the Trimmed Likelihood Curves shown in the top panels of Fig. 4, one should observe the correct number of clusters (say k) and the corresponding trimming level (alpha). However, the interpretation of these panels is not clear, in the sense that it is not obvious how to choose the smallest value k and alpha such that no big change in the classification trimmed likelihood curves are found when increasing from k to $k + 1$. Since here we wanted to stay on a few number of clusters and a relative small trimming, we decided in a rather arbitrary way that $k = 4$ and alpha $= 0.04$ were somehow reasonable. The ellipses associated with the four clusters produced by TCLUST method are drawn in the bottom panel of Fig. 4 together with the 4 % unclassified units shown with circles.

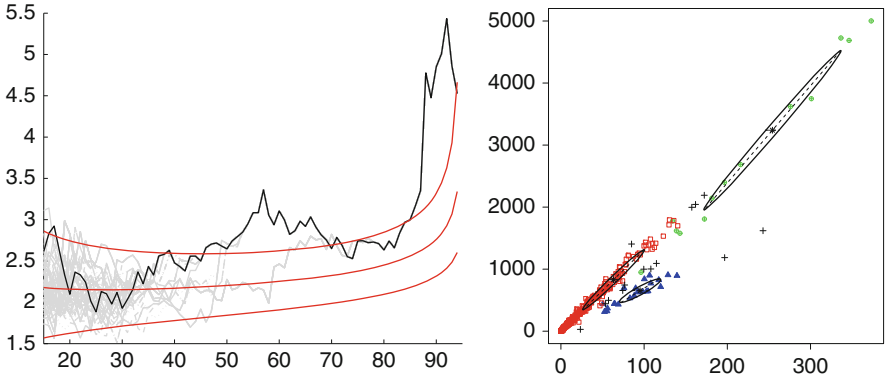

Fig. 5 Minimum deletion residual trajectories from 200 FS random starts on the gridded data (*left panel*). The final FS classification of the Fishery dataset based on centroids found on the gridded data with the FS (*right panel*)

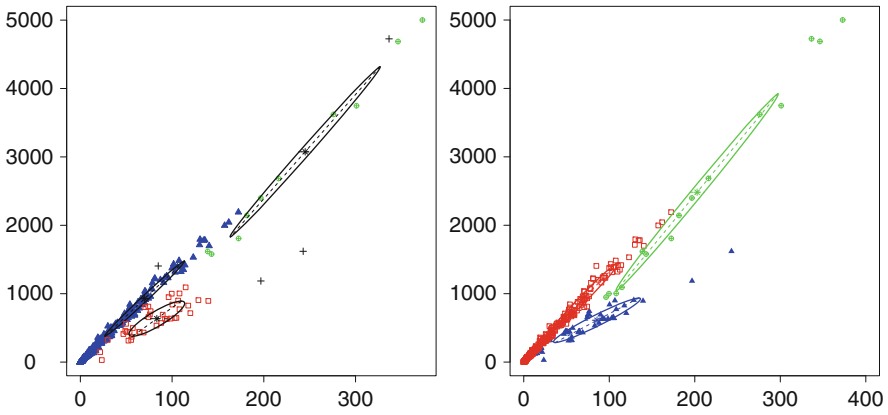

Fig. 6 Final classification of the Fishery dataset based on centroids found on the gridded data with TCLUST (*left panel*) and MCLUST (*right panel*)

When we apply the three methods to the gridded data, we obtained a more meaningful number of components as with the original data. First of all in all cases the estimated number of clusters was always three. We then classify the rest of the data based on the smallest Mahalanobis distance computed using the centroids and the variance-covariance matrices of the groups found on the gridded data. The final data classification for the three clustering methods is shown in Figs. 5 and 6. The units which remain unassigned in the FS and TCLUST are represented with the plus symbol. Compared to the previous clustering without gridding, all these new clusters are much more meaningful thanks to the data balancing of the gridding technique.

4 Conclusions

In this paper we have shown how different model-based-clustering methods are bad-performing in presence of densely populated data points with high degree of overlapping. We have therefore proposed to precede each clustering method with a technique that selects a point in each cell of a grid defined along the Principal Component axes of the data, in order to identify a sub-sample that preserves the general structure of the data, on which to apply a clustering technique.

References

Atkinson, A. C., & Riani, M. (2007). Exploratory tools for clustering multivariate data. *Computational Statistics and Data Analysis, 52*, 272–285.

Atkinson, A. C., Riani, M., & Cerioli, A. (2004). *Exploring multivariate data with the forward search*. New York: Springer.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. New York: Springer.

Cochran, W. G. (1977). *Robust sampling techniques* (3rd ed.). New York: Wiley.

Fraley, C. (1998). Algorithms for model-based Gaussian hierarchical clustering. *SIAM Journal on Scientific Computing, 20*, 270–281.

Garcia-Escudero, L. A., Gordaliza, A., Matran, C., & Mayo-Iscar, A. (2008). A general trimming approach to robust cluster analysis. *Annals of Statistics, 36*, 1324–1345.

Garcia-Escudero, L. A., Gordaliza, A., Matran, C., & Mayo-Iscar, A. (2010). A review of robust clustering methods. *Advances in Data Analysis and Classification, 4*, 89–109.

Garcia-Escudero, L. A., Gordaliza, A., Matran, C., & Mayo-Iscar, A. (2011). Exploring the number of groups in robust model based clustering. *Statistics and Computing, 21*(4), 585–599.

Perrotta, D., & Torti, F. (2009). Detecting price outliers in European trade data with the forward search. In N. C. Lauro, F. Palumbo, & M. Greenacre (Eds.), *Data analysis and classification: From exploration to confirmation* (Springer studies in classification, data analysis, and knowledge organization, pp. 415–423). Berlin: Springer.

Riani, M., Cerioli, A., Atkinson, A. C., Perrotta, D., & Torti, F. (2008). Fitting robust mixtures of regression lines to European trade data. In: F. Fogelman-Soulie, et al. (Eds.), *Mining massive datasets for security applications*. Amsterdam: IOS Press.

Riani, M., Atkinson, A. C., & Cerioli, A. (2009). Finding an unknown number of multivariate outliers. *Journal of the Royal Statistical Society, Series B – Statistical Methodology, 71*, 447–466.

Riani, M., Perrotta, D. and Torti, F. (2012). FSDA: A MATLAB toolbox for robust analysis and interactive data exploration. In: *Chemometrics and Intelligent Laboratory Systems, 116,* 17–32.

Rissanen, J. (1986). Stochastic complexity and modeling. *Annals of Statistics, 14*, 1080–1100.

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*(2), 461–464.