Matilde Bini, Pietro Amenta, Antonello D'Ambra, Ida Camminatiello
Editors

SIS Società Italiana di Statistica

Statistical Methods for Service Quality Evaluation

Book of short papers of IES 2019, Rome, Italy, July 4-5

SIS Società Italiana di Statistica

Cuzzolin

Matilde Bini, Pietro Amenta, Antonello D'Ambra, Ida Camminatiello
Editors

SIS Società Italiana di Statistica

Statistical Methods for Service Quality Evaluation

Book of short papers

9th International Conference **IES 2019** - Innovation & Society -

Statistical evaluation systems at 360°: techniques, technologies and new frontiers

organized by Statistics for the Evaluation and Quality in Services Group of the Italian Statistical Society and European University of Rome

SIS Società Italiana di Statistica

Cuzzolin

Matilde Bini - European University of Rome, Italy
Pietro Amenta - University of Sannio, Italy
Antonello D'Ambra - University of Campania "L. Vanvitelli", Italy
Ida Camminatiello - University of Campania "L. Vanvitelli", Italy
Editors

Prima Edizione: Luglio 2019

Traversa Michele Pietravalle, 8 - 80131 Napoli
Tel. 081 5451143 - Fax 081 7707340
cuzzolineditor@cuzzolin.it
www.cuzzolineditore.com

ISBN: 978-88-86638-65-4

CUZZOLIN EDITORE

Robust multivariate analysis for mixed-type data

Analisi multivariata robusta per dati misti

Aurea Grané and Silvia Salini

Abstract In this work we deal with classifying mixed type data using a hierarchical approach based on Forward Search. The identification of groups is based on the identification of similar trajectories and then linked to very intuitive two-dimensional maps. The proposed algorithm can implement different measures for mixed type data, such as Gower or Related Mertric Scaling. The algorithm will be applied to data related to a large set of countries. In economics it is a shared idea that a country's economic performance, but also more or less sustainable development policies depend on their deep institutional characteristics. These characteristics are described by indicators of mixed nature

Abstract *Nel presente lavoro ci occupiamo di classificare dati misti utilizzando un approccio gerarchico basato sulla Forward Search. L'individuazione dei gruppi si basa sull'identificazione di traiettorie simili e poi collegate a mappe bidimensionali molto ituitive. L'algoritmo proposto pu implementare diverse misure per dati misti, per esempio Gower o Related Mertric Scaling. L'algoritmo sar applicato a dati relativi a un ampio insieme di paesi. In economia una idea condivisa che le performance economiche, ma anche le politiche di sviluppo pi o meno sostenibile, di un paese dipendono dalle loro profonde caratteristiche istituzionali. Tali caratterisitche sono descritte da indicatori di natura mista*

Key words: Forward Search, Mixed Type Data, Outliers, Robustness

Aurea Grané
Universidad Carlos III de Madrid , e-mail: agrane@est-econ.uc3m.es

Silvia Salini
University of Milan e-mail: silvia.salini@unimi.it

Bini M., Amenta P., D'Ambra A., Camminatiello I. (Editors). Statistical Methods for Service quality evaluation. Cuzzolin Publisher.

1 Introduction

Mixed type data comprises both numeric and categorical features, and mixed datasets occur frequently in many domains, such as economics, health, finance, marketing, as well as, data coming from socio-demographic surveys. Clustering is often applied to mixed datasets to find structures and to group similar objects for further analysis. However, clustering mixed type data is challenging because it is difficult to directly apply mathematical operations, such as summation or averaging, to the feature values of these datasets. In literature some proposals exist in order to cluster mixed-type data [1], but robustness is not explored in this context. [6] proposed a statistic to identify multivariate outliers in the framework of mixed-type data. The statistic is related to a distance-based proximity function [5] and its effectiveness was studied through the analysis of several contaminated data sets. The aim of this work is to combine the measure proposed by [6] with Forward Search algorithm. The Forward Search is a powerful general method, incorporating flexible data-driven trimming, for the detection of outliers and unsuspected structure in data and so for building robust models. Starting from small subsets of data, observations that are close to the fitted model are added to the observations used in parameter estimation. As this subset grows we monitor parameter estimates, test statistics and measures of fit [2].

Section 2 is devoted to the proposed algorithm. Section 3 presents an application in the economical context. Section 4 describe future directions of the research.

2 Method

We apply the Forward Search method to mixed type data following this algorithm:

1. Data matrix of mixed-type data $n \times p$.
2. Select a distance measure. In this first example we use Gower's similarity coefficient. Given two p-dimensional vectors $\mathbf{z}_i$ and $\mathbf{z}_j$, Gower's similarity coefficient is defined as

$$s_{ij} = \frac{\sum_{h=1}^{p_1} \left(1 - |z_{ih} - z_{jh}|/R_h\right) + a + \alpha}{p_1 + (p_2 - d) + p_3}, \quad 0 \leq s_{ij} \leq 1,$$

   where $p = p_1 + p_2 + p_3$, p_1 is the number of continuous variables, a and d are the number of positive and negative matches, respectively, for the p_2 binary variables, α is the number of matches for the p_3 multi-state categorical variables, and R_h is the range of the h-th continuous variable. Gower's distance is defined as $\delta^2(\mathbf{z}_i, \mathbf{z}_j) = 1 - s_{ij}$, which are the entries of the matrix of squared distances Δ.
3. Select a subset size ($m < n$).
4. Select the units inside the starting subset which lowest distance measure.
5. Calculate the geometric variability of the subset V_Δ. Let $\{\mathbf{z}_i, 1 \leq i \leq n\}$ be n p-dimensional vectors containing the information of n individuals and consider

a matrix Δ of squared distances, with entries $\delta^2(\mathbf{z}_i, \mathbf{z}_j)$, for $1 \leq i, j \leq n$. The geometric variability of Δ is

$$V_\Delta = \frac{1}{2n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \delta^2(\mathbf{z}_i, \mathbf{z}_j).$$

6. Calculate for each units outside the subset the *distance-based proximity* $\phi(i)$ to the subset. Given a new individual $\mathbf{z}_0 \in \mathbb{R}^p$, the distance-based proximity of $\mathbf{z}_0$ to the set $\{\mathbf{z}_i, 1 \leq i \leq n\}$ is

$$\phi(\mathbf{z}_0) = \frac{1}{n} \sum_{i=1}^{n} \delta^2(\mathbf{z}_0, \mathbf{z}_i) - V_\Delta.$$

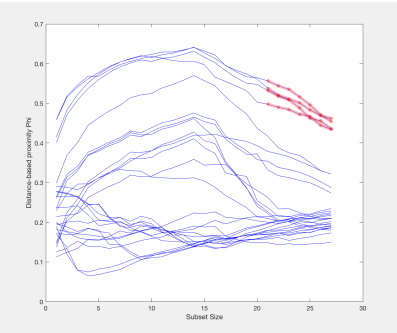
7. Include in the subset the unit with the minimum value of $\phi(i)$.
8. Iterate the procedure until n.
9. Monitoring $\phi(i)$ for each unit on the subset size.
10. Plot the trajectory in MDS maps and identify groups and outliers.

3 Application

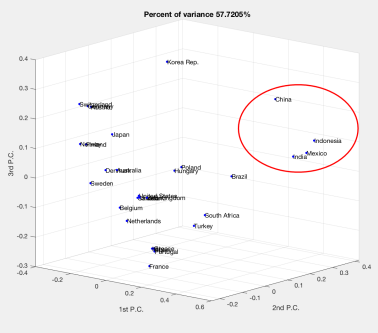
Regarding the data selected, for a set of 35 countries, we consider three types of variables, described in [7]. First, we focus on the legal origins of a country, which can be either British, French, Socialist, German and Scandinavian, these variables are binary. Second we consider the quality of the country's bureaucratic environment, proxies by the level of tax compliance, an index for bureaucratic delays, and an index measuring comparison, these variables are numerical. Third, we focus on indexes describing rights: political rights, property rights, an index of business regulation and a score for democracy, these variables are categorical. All these variables are meant to help identify economies with similar underlying structure. The benefit of using these data to build a similarity indicator is its potential use for the generation of instrumental variables in economic analyses.

The following figure shows the forward plot and the multidimensional plot using Gower's distance. In the first plot it is possible to identify similar trajectories that enter in the search in the final steps. These trajectories represent similar countries that differ from the others. The same countries can be identified, brushing the forward plot, in the multidimensional scaling plot, they are India, Mexico, China and Indonesia.

The application is relevant for example in the economic growth context. The potential applicability of this approach ranges from the evaluation of environmental and energy policy indexes to the role of monetary policy. In all these instances, endogeneity plagues economic analysis, and deep country characteristics are perfect candidates to build for instrumental variables.



Forward Search plot



MDS Plot

4 Conclusions

Future intent is to implement, in the step 2 of the algorithm, a robust Gower measure (Mahalanobis instead of Manhattan) as well as to use Related Metric Scaling (RelMS) instead of Gower. RelMS is a multivariate technique that allows to obtain a unique representation of a set of individuals from several distance matrices computed on the same set of individuals. The method is based on the construction of a joint metric that satisfies several axioms related to the property of identifying and discarding redundant information [3, 4]. The methods, in the final paper, will be tested in a larger datasets, when anomalous data and contaminations are present. We plan to develop the proposed method inside the common and flexible computational framework provided by the FSDA Toolbox of Matlab [8].

References

1. Ahmad A, Khan S (2019) Survey of State-of-the-Art Mixed Data Clustering Algorithms. IEEE Access 7: 31883-31902
2. Atkinson AC, Riani M, Cerioli A (2010) The forward search: Theory and data analysis. Journal of the korean statistical society 39(2):117–134
3. Cuadras CM (1998) Multidimensional dependencies in classification and ordination. Analyses Multidimensionelles des Données pp 15–25
4. Cuadras CM, Fortiana J (1998) Visualizing categorical data with related metric scaling. In: Visualization of Categorical Data, Elsevier, pp 365–376
5. Cuadras CM, Fortiana J, Oliva F (1997) The proximity of an individual to a population with applications in discriminant analysis. Journal of Classification 14(1):117–136
6. Grané A, Romera R (2018) On visualizing mixed-type data: A joint metric approach to profile construction and outlier detection. Sociological Methods & Research 47(2):207–239
7. La Porta R, Lopez-de Silanes F, Shleifer A, Vishny R (1999) The quality of government. The Journal of Law, Economics, and Organization 15(1):222–279
8. Riani M, Perrotta D, Torti F (2012) FSDA: a matlab toolbox for robust analysis and in- teractive data exploration. Chemometrics and Intelligent Laboratory Systems 116:17 32