A robust proposal of fuzzy biclustering *

Maria Brigida Ferraro, Maurizio Vichi, Paolo Giordani
Dipartimento di Scienze Statistiche
Sapienza Università di Roma
e-mail: {mariabrigida.ferraro, maurizio.vichi, paolo.giordani}@uniroma1.it

1 Abstract

Two-mode clustering consists in simultaneously clustering modes (e.g., objects, variables) of an observed two-mode data matrix. This idea arises to face with situations in which objects are homogeneous only within subsets of variables, while variables may be strongly associated only on subsets of objects. There are many practical applications presenting the above situations, for example, DNA microarrays analysis, market basket analysis and so on. By using a standard one-mode cluster analysis, the clusters of the objects are identified without considering that clusters of variables are present in the data. This problem can be overcome by simultaneously clustering the two modes. In this way all the information contained in heterogeneous datasets is completely taken into consideration.

In the literature there are different proposals of fuzzy two-mode clustering for the specific case of categorical multivariate data (Oh et al., 2001; Kummamuru et al., 2003; Tjhi and Chen, 2005, 2009). In a categorical multivariate dataset \( n \) individuals are described by a set of qualitative variables with \( p \) categories. Unfortunately, the above methods are not appropriate for heterogeneous two-mode datasets. They turn out to be variable selection procedures rather than classification ones. Starting from Double \( k \)-Means (Vichi, 2001), an extension to the fuzzy case is obtained. Unfortunately, both standard and fuzzy \( k \)-means type algorithms are affected by outliers and, in case of heterogeneity, outliers are likely to occur. They can be intermediate values between two clusters or can be far from all the remaining data. Outliers can be indicative of measurement errors or they are generated by a different data generation processes. In two-mode datasets, they can be due to different objects and variables generation mechanisms. In order to overcome these problems, we propose a robust fuzzy two-mode clustering procedure. The new method is based on the idea of "noise cluster", introduced by Davé (1991). The adequacy of the new proposal is evaluated by means of simulation and real-case studies.

References


*Presented at the second internal meeting of the FIRB (“Futuro in ricerca” 2012) project “Mixture and latent variable models for causal-inference and analysis of socio-economic data”, Rome (IT), January 23-24, 2015

