Research advances and new challenges in Cluster Analysis

Maurizio Vichi

Abstract Methodologies for *Cluster Analysis* are among the most well-known and appreciated statistical techniques of multivariate analysis. In the last twenty years they have been increasingly applied in new disciplines and frequently almost reinvented in many area of research such as computer science, engineering, bioinformatics and in specific fields including machine learning, data mining and pattern recognition. In this presentation we show recent *statistical* research advances in methodologies for clustering. The illustrated methods have in common the statistical approach of formulating a mathematical model for partitioning or hierarchical clustering multivariate observations, estimating parameters of the model and finally fitting it to data.

1. Introduction

The interpretation of the relationship within a set of objects can be helped by obtaining a hard partition of the objects into disjoint classes, with the property that objects in the same class are perceived as similar to one another, while objects in different classes are considered dissimilar. Such partitions can be achieved from the application of *Cluster Analysis* methodologies. Several methods have been proposed for clustering a set of multivariate objects. In this presentation we concentrate on the model based approach of formulating a clustering model for the data, e.g., a partition or a hierarchy specified for reconstructing data (multivariate observations or dissimilarities) and then solving the least-squares or maximum likelihood corresponding fitting problem.
The presentation in divided in three parts: model-based partitioning and hierarchical clustering of a set of units for dissimilarity data, multi-partitioning of the modes of a three and two way data matrix including multivariate observations and clustering of longitudinal multivariate observations.

2. Model-Based partitioning and hierarchical clustering

The Cluster Analysis problem of partitioning or hierarchical clustering a set of units, when dissimilarity data are observed, is here handled with the statistical model-based approach of fitting the "closest" *classification matrix* to the observed dissimilarities. A classification matrix represents a clustering model expressed in terms of dissimilarities.
Three models for partitioning a set of units from dissimilarity data, are illustrated and their estimation -via least-squares- is given together with new fast coordinate descent algorithms. Following the same statistical fitting approach a new model for hierarchical clustering objects starting from dissimilarity data is also illustrated.

3. Bi-partitioning, multi-partitioning, clustering and disjoint principal component analysis

New methodologies for three-mode (units, variables and occasions) and two-mode (units and variables) symmetrical or asymmetrical partitioning or multi-partitioning three- and two-way data are presented. In particular, by reanalyzing the *double k-means*, that identifies a unique partition for each mode of the data, a relevant extension is discussed which allows to synthesize classes of each mode symmetrically by means of mean vectors or linear combinations (components) for all modes, or asymmetrically by mixing a different strategy for each mode. Furthermore, the model allows the partition of one mode, conditionally to the partition of the other one. The performance of such *generalized double k-means* has been tested by both a simulation study and an application to gene microarray data. Clustering and disjoint principal component allows to identify a partition of the units and a partition of the variables together with a principal component for each class of the partition of variables. This technique can be seen as a special case of the generalized double k-means.

4. Clustering longitudinal multivariate observations

Longitudinal multivariate data involve repeated observations of different features of the same statistical units over a period of time. The aim is to study the developmental trends of the units across at least a part of their life span.
The dynamic evolution of the partitions of units along time is in this presentation studied in an unsupervised clustering context by using a model based clustering approach. A clustering together with a vector autoregression VAR(P) model -where P is the lag length of the VAR- are combined into a new technique that identifies an homogeneous partition in G classes for each time t and the autoregressive dynamic evolution of the clusters. The proposed clustering/VAR model can be used also to forecast a partition at time T+1. The parameters of the model are estimated both in a least-squares and maximum likelihood framework and efficient recursive algorithms are given. A simu-

lation study together with some applications of the proposed methodologies are shown to appreciate performances of the models and the quality of its estimates.
In the final part of the presentation , similarities between trajectories describing histories of units are studied. Trend, velocity and acceleration are three characteristics of trajectories considered to assess pairwise dissimilarities between trajectories. The Tucker model for three-way data, modified for clustering units together with a dimensional reduction of the observed variables, is estimated in the metric space specified by trend, velocity and acceleration. An application is given to show the performances of the methodology.

References

Martella F., Alfò M., Vichi M. (2010). Hierarchical mixture models for biclustering in microarray, *STATISTICAL MODELLING*, 11(6): 489-505.

Martella F., Vichi M. (2012) Clustering microarray data using model-based double *K*-means, *JOURNAL OF APPLIED STATISTICS*, DOI:10.1080/02664763.2012.683172.

Maruotti A., Vichi M. (2012) *Clustering Longitudinal Multivariate Observations: Model-Based Autoregressive K-means*, Submitted.

Rocci R., Gattone A., Vichi, M (2011). A New Dimension Reduction Method: Factor Discriminant *K*-means, *JOURNAL OF CLASSIFICATION*, vol 28, DOI: 10.1007/s00357-011

Vicari D., Vichi M. (2009). Structural Classification Analysis of Three-Way Dissimilarity Data. *JOURNAL OF CLASSIFICATION*, vol. 26; p. ., ISSN: 0176-4268

Vichi M., Rocci R. (2008). Two-mode Multi-partitioning. *COMPUTATIONAL STATISTICS & DATA ANALYSIS*. vol. 52, pp. 1984-2003 ISSN: 0167-9473.

Vichi M., Saporta G. (2009). Clustering and Disjoint Principal Component Analysis. *COMPUTATIONAL STATISTICS & DATA ANALYSIS* vol. 53; p. 3194-3208, ISSN: 0167-9473, doi: 10.1016/j.csda.2008.05.028,

Vichi M. (2011) Fitting Hierarchical Clustering Models to Dissimilarity Data, Submitted.

Vichi M. (2008). Fitting Semiparametric Clustering Models to Dissimilarity Data, *ADVANCES IN DATA ANALYSIS AND CLASSIFICTION*, vol, 2, 2, 121-161, DOI: 10.1007/s11634-008-0025-4