Clustering models for high dimensional, temporal, and dissimilarity data

Maurizio VICHI

Abstract:
In this talk the cluster analysis problem of partitioning a set of objects from dissimilarity data is handled with the statistical model-based approach of fitting the "closest" classification matrix to the observed dissimilarities. A classification matrix represents a clustering structure expressed in terms of dissimilarities. In cluster analysis there is a lack of methodologies widely used to directly partition a set of objects from dissimilarity data. In real applications, a hierarchical clustering algorithm is applied on dissimilarities and subsequently a partition is chosen by visual inspection of the dendrogram. Alternatively, a "tandem analysis" is used by first applying a Multidimensional Scaling (MDS) algorithm and then by using a partitioning algorithm such as k-means applied on the dimensions specified by the MDS. However, neither the hierarchical clustering algorithms nor the tandem analysis is specifically defined to solve the statistical problem of fitting the closest partition to the observed dissimilarities. This lack of appropriate methodologies motivates the study of three new object partitioning models for dissimilarity data, their estimation via least-squares and the introduction of three new fast algorithms.

Keywords:
Cluster Analysis, dissimilarity data

References
Fitting Semiparametric clustering models to dissimilarity data, Advances in Data Analysis and Classification, Volume 2, Number 2, 2008.