Recent Advances in Partitioning Clustering Algorithms for Interval-Valued Data

Francisco de Assis Tenório de Carvalho

Universidade Federal de Pernambuco
Recife, Brésil.
fatc@cin.ufpe.br
http://www.cin.ufpe.br/~ fatc/

Cluster analysis have been widely used in numerous fields including pattern recognition, data mining and image processing. Their aim is to organize a set of objects into clusters such that items within a given cluster have a high degree of similarity, whereas items belonging to different clusters have a high degree of dissimilarity. In particular, partitioning clustering models aims to organize a set of items into a pre-defined number of clusters. Our reference clustering algorithms are the so-called partitioning dynamic cluster algorithms. They are iterative two steps relocation clustering algorithms involving at each iteration the construction of the clusters and the identification of a suitable representative or prototype (means, factorial axes, probability laws, etc.) of each cluster by locally optimizing an adequacy criterion between the clusters and their corresponding prototypes. Often, objects to be clustered are represented as a vector of quantitative features. However, the recording of interval-valued data has become a common practice in real world applications and nowadays this kind of data is often used to describe objects. Symbolic Data Analysis (SDA) is an area related to multivariate analysis, data mining and pattern recognition, which has provided suitable data analysis methods for managing objects described as a vector of intervals. In this talk, we review partitioning clustering algorithms for interval-valued data having as reference the dynamic clustering algorithm. For each clustering algorithm, it is given the clustering criterion, the best prototype of each cluster, the best distance associated to each cluster (if any) as well as the best partition in a fixed number of clusters. Moreover, various tools for the partition and cluster interpretation of interval-valued data furnished by these algorithms are also presented. Finally, in order to show the usefulness of these algorithms and the merit of the partition and cluster interpretation tools, experiments with real interval-valued data sets are given.