Graphical Determination of Groups and Outliers in Distance-Based Cluster Analysis

Luis F. Rivera-Galicia*
University of Alcalá, Alcalá de Henares, Spain luisf.rivera@uah.es

Cluster analysis is a popular unsupervised learning method. Its goal is to find a partition of a dataset of N objects into k well separated groups: elements within a group must be similar (in some sense) to one another, and different to elements in other groups. The fundamental problem of cluster analysis is to determine the real number of groups (k) in the dataset.

In this paper, a new method of clustering is presented, to simultaneously determine the number of groups and the clustering in a dataset. This method is based on graph theory. Dissimilarity data between objects is used to form a dissimilarity graph, in which the vertices are the objects in dataset, and the edges are weighted according to the dissimilarity between the objects. Two vertices are then connected by an edge, when the dissimilarity among them is under some certain threshold. A statistical procedure is proposed to determine the appropriate threshold to split the graph into its connected components. As an additional result, cases which are isolated can be considered as outliers and may need to be further analyzed. This method has been tested on some different datasets, and results obtained are analyzed taking into account the resulting clustering.

Key words: clustering, number of groups, outliers.