

Clustering Performance Analysis using New Correlation-Based Cluster Validity Indices

Authors : Nathakhun Wiroonsri

Abstract : There are various cluster validity measures used for evaluating clustering results. One of the main objectives of using these measures is to seek the optimal unknown number of clusters. Some measures work well for clusters with different densities, sizes and shapes. Yet, one of the weaknesses that those validity measures share is that they sometimes provide only one clear optimal number of clusters. That number is actually unknown and there might be more than one potential sub-optimal option that a user may wish to choose based on different applications. We develop two new cluster validity indices based on a correlation between an actual distance between a pair of data points and a centroid distance of clusters that the two points are located in. Our proposed indices constantly yield several peaks at different numbers of clusters which overcome the weakness previously stated. Furthermore, the introduced correlation can also be used for evaluating the quality of a selected clustering result. Several experiments in different scenarios, including the well-known iris data set and a real-world marketing application, have been conducted to compare the proposed validity indices with several well-known ones.

Keywords : clustering algorithm, cluster validity measure, correlation, data partitions, iris data set, marketing, pattern recognition

Conference Title : ICDMCA 2022 : International Conference on Data Mining Cluster Analysis

Conference Location : Bangkok, Thailand

Conference Dates : December 20-21, 2022