Abstract

The K-means clustering algorithm is an old algorithm that has been intensely researched owing to its simplicity of implementation. However, there have also been criticisms on its performance, in particular, for demanding the value of K a priori. It is evident from previous researches that providing the number of clusters a priori does not in any way assist in the production of good quality clusters. The objective of this paper is to investigate the usefulness of the K-means clustering in the clustering of high and multi-dimensional data by applying it to biological sequence data which is known for high and multi-dimension. The squared-Euclidean distance and the cosine measure are used as the similarity measures. The silhouette validity index is used first to show K-means algorithm’s inefficiency in the clustering of high and multi-dimensional data irrespective of the distance or similarity measure employed. A further study was to introduce a preprocessor scheme to the K-means algorithm to automatically initialize a suitable value of K prior to the execution of the K-mean algorithm. The dimensionality problem investigated suggests that the use of the preprocessor improves the quality of clusters significantly for the biological data sets considered. Furthermore, it is then shown that the K-means algorithm with preprocessor produces good quality, compact and well-separated clusters of the biological data obtained from a high-dimension-to-low- dimension mapping scheme introduced in the paper.

References
More work on K-Means Clustering Algorithm: The Dimensionality Problem

11.

**Index Terms**

Computer Science Pattern Recognition

**Keywords**

Clustering Dimensionality Categorical Data Silhouette Validity Index