Abstract

Clustering is the process of grouping a set of objects into classes of similar objects. A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters. Clustering is a method of unsupervised learning, as it partitions the set of objects into groups based on data similarity and then assigns labels to the groups.

K-means algorithm is a scalable and efficient clustering algorithm which partitions a set of objects into k clusters. K-means algorithm converges to one of the local minima. The algorithm requires random selection of initial points for clusters due to which it is sensitive to the initial starting conditions. We present a method for selecting a starting condition consisting of k-furthest points. We demonstrate that application of this method to k-means algorithm convergence at a better local minimum and improve its performance.

Mixed type data is common in real life datasets. Most of the algorithm work effectively either on pure numeric data or on pure categorical data. We propose another algorithm which effectively clusters mixed type datasets as well as both pure numeric and pure categorical datasets.

Keywords: Cluster, clustering, datasets, algorithm, mixed type data