Computation of Initial Modes for K-modes Clustering Algorithm using Evidence Accumulation

Prathmesh Gaonkar (B20ME032)
gaonkar.1@iitj.c.in

Department of Mechanical Engineering
Indian Institute of Technology Jodhpur

Abstract

In machine learning, there are two basic approaches to learning a dataset and training our model. These are supervised and unsupervised learning. Supervised learning uses labeled data, whereas we do not have labeled data in unsupervised learning. The algorithms of the unsupervised datasets discover some meaningful structures and hidden patterns by analyzing the data, thus forming clusters of these unlabeled data.

Clustering is one of the most useful and popular unsupervised learning techniques. It is a data mining technique used to group unlabelled data based on their similarities or differences. It aims to group the set of data points in clusters such that objects within the same cluster are more similar than those in other clusters.

K means clustering algorithm is one of the most useful and popular clustering algorithms because of its efficiency in clustering large datasets. It is a centroid-based clustering technique. The method aims to separate the n observations in the dataset into K clusters in which each observation belongs to that cluster whose mean is nearest to it. With each iteration(forming groups and assigning new means), the algorithm aims to reduce the following objective function

$$\min_{\mathbf{m}} \min_{\mathbf{C}} \sum_{i=1}^{k} \sum_{x \in C_i} |x - m_i| \tag{1}$$

However, this algorithm fails to work with categorical data as it minimizes the cost function, which is numerically measured. So to remove the numeric-only limitation of the K means-algorithm proposed the K modes algorithm, which extends the K means algorithm by using a simple matching dissimilarity measure for categorical data. This algorithm uses modes instead of means and a frequency-based approach to update modes in the clustering process to minimize the clustering cost function.

The accuracy of this partitional clustering algorithm depends on the choice of initial data points to instigate the clustering process. The choice of the initial modes is random. The clustering obtained in such a case is not reproducible, and hence the results cannot be relied upon with confidence. This paper offers an approach to compute these initial modes. Here, the idea used is Evidence Accumulation for combining the results of multiple clusterings. Initially, we apply the K modes algorithm, which performs the decomposition of n-F dimensional data. Several clusters are obtained by N random initialization of the K modes algorithm. The modes that we get in each iteration are stored in a Mode-pool, Pn. The aim is to investigate the contribution of those data objects that are less vulnerable to a random selection of modes and choose the most diverse set of modes from the available Mode-pool that can provide us with consistent results. Later, we can test that the modes obtained by this process are very similar to the desired modes and gives better clustering results with minor variance compared to the traditional method.