

## Abstract

Clustering is a process of grouping data into a class or cluster, so that the objects in a cluster has a very large similarity with other objects in the same cluster, but not similar to objects in other clusters.

One commonly used algorithm for data clustering process is the k-means algorithm. K-means is very popular in clustering data process because its efficiency for clustering data. However, this algorithm is limited to numerical data grouping, whereas in fact, in the real world there are many valuable attributes of categorical data.

To handle the problem of categorical data, in this Final Project will be discussed an algorithm called the k-modes which is a variant of k-means algorithm. Just as k-means algorithm, k-modes algorithm produces local optimum solution. This is related to the initialization process in determining the initial cluster centroid. This Final Project explains about the methods for determining first initialization of k-modes algorithm by randomly, and using frequency-based method.

It is shown in this Final Project that the selection method of first k initialization using frequency-based method which has better accuracy in grouping data compared with random initialization.

**Keywords:** Clustering, k-means, k-modes, frequency-based