Abstract ID: P38

DETERMINING THE OPTIMAL NUMBER OF CLUSTERS USING DISTANCE BASED *K*-MEANS ALGORITHM

M.C. Alibuhtto*

Department of Mathematical Sciences, Faculty of Applied Sciences,
South Eastern University of Sri Lanka, Sammanthurai, Sri Lanka.

**mcabuhtto@seu.ac.lk*

Abstract

In the current digital era, data is generated enormously with fast growth from different sources, and managing such huge data is a big challenge. Clustering algorithm is able to find hidden patterns and extract useful information from huge datasets. Among the clustering techniques, *k*-means clustering algorithm is the most commonly used unsupervised classification technique to determine the optimal number of clusters (*k*). However, the choice of the optimal number of clusters (*k*) is a prominent problem in the process of the *k*-means clustering algorithm. In most cases, clustering huge data, *k* is pre-determined by researcher, and incorrectly chosen *k* leads to increase computational cost. In order to obtain the optimal number of clusters, a distance-based *k*-mean algorithm was proposed with a simulated dataset. In the *k*-means algorithm, two distance measures were considered namely Euclidean and Manhattan distances. The results based on simulated data reveal that the *k*-means algorithm with Euclidean distance yields the optimal number of clusters compared to Manhattan distance. Testing on real datasets shows consistent results as the simulated ones.

Keywords: *huge data, digital era, distance measure, K-means algorithm*