Abstract

Clustering or grouping is one of the method used to solve problems in data mining. There are several algorithms available to solve such problems in clustering, one of which is k-means clustering. However, there are some weaknesses in the k-means algorithm so that it reduce the accuracy and efficiency of the result. One of these weaknesses is the determination of the cluster center or centroid initialization randomly can cause inaccurate results because the cluster center does not spread within the entire dataset. In this study some methods will be discussed to minimize these weaknesses hence improving accuracy and efficiency of the process. Determining cluster center or initial cluster centroid algorithm using K-Means++ led to distribution of clusters in the data point is more prevalent when compared to using K-Means which choosing random cluster centers are intended. Canopy method or canopy clustering is a method commonly used as an initial step of clustering methods such as K-Means and Expectation Maximization (EM). Canopy clustering has a simple, fast, and cheap process to classify the data into specific groups. By combining Canopy with other clustering methods such as K-Means will reduce the distance calculation between the data points that usually taken by forming a canopy to limit the number of data points to be calculated. By using canopy and k-means ++ generating the accuracy of 98.3% for dim1024 dataset and 87.7% for Iris dataset. In terms of running time, canopy and k-means ++ get the results within 7 seconds on dim1024 dataset and 0.77 seconds in the iris dataset.

Keywords: Data mining, Clustering, K-means Clustering, Canopy Clustering, Centroid Initialization, K-Means++