Abstract

Data is an important resource that can be processed into an useful information. To generate the data to be an information, the data must be processed properly. Data mining is a process to find an interesting information from very large data. In data mining there are many techniques to process the data, one of which is clustering. Clustering is the process of grouping the data based on characteristics of data into cluster(group) that has a same characteristics (similarity) in the same cluster and has a different characteristics (dissimilarity) with object in the other cluster. To measure the similarity between the data, the distance measurement can be used for numeric data type, such as Euclidean distance. Euclidean distance is very easy to be use and working good when the type of data is numeric, but when the type of data is categorical or mixed, Euclidean distance is cannot be used. In real world, a lot of data is categorical or mixed. In this Final project will be implemented clustering mixed type data with Incremental Genetic K-means Algorithm.

In this Final Project the distance will be calculated separately for numerical and categorical data. Incremental Genetic K-means Algorithm starts with the randomized initialization of population, and calculate the fitness function and selection, then the individual who are selected will be transferred with mutation operator. After the mutation phase all individuals will be processed with k-means operator. This process will continue until the end of iteration or until the center of each cluster has not changed. From result and the analysis found that the number of clusters can affect the silhouette coefficient. The larger number of clusters will make the silhouette coefficient tends to increasingly rise.

Keywords: *data mining, clustering, IGKA, mixed data type*