Abstract

Dimensionaliy is the one of data mining challenge, this challenge include large number of attribute this also called the curse of dimensionality. the greater number of attribute is more time consuming and need more excessive computational effort to handle. the things that needed to handle this challenge is to reduce the dimension of data.

The reduction technique that mentioned in this final task is to use K-Means algorithm to grouping the data into each cluster. The use of this algorithm is to reduce the record and then GA as feature selection to select the optimal attribute with higher fitness value. The search for fitness value can be done by classification method SVM.

The result of the system examination is data reduced by K-Means have lowest accuracy in certain dataset compared to without using K-Means. The result of optimal atribut which GA produce is varies based on the use of different parameter. The data that used is the high dimensional disease data in the form of gene expression, i.e. colon tumor and leukemia. The best average accuracy for colon tumor is 92.86% with selected attribute 983 attributes while leukemia always produce best attribute with average accuracy 100%.

Keywords : *dimensionality, data mining, K-Means, GA , SVM*