ABSTRACT

Microarray data takes an essential part in diagnosing and detecting cancer because microarray analysis can be used to look at levels of gene expression in specific cell samples that serve to analyze thousands of genes simultaneously. However, microarrays data have very little sample data and have high data dimensions. So to classify the microarray data, it requires dimensional reduction process. The dimensional reduction can eliminate redundancy of data thus features used in classification are features that have a high correlation to its class. There are two types of dimensional reduction: feature selection and feature extraction. In this study, it will use feature selection, using the k-means clustering algorithm to classify features that have similarity level at 1 cluster, and then each cluster is ranked by applying Relief method. After that best-scored in each element, a cluster is selected and to be a subset feature for the classification process. Its purpose is to remove redundancy from the data that can decrease the accuracy of the classification. Next, on the classification process, it will use Random Forest algorithm. From the results of this study obtained the results of accuracy for each dataset, namely Colon 85.87%, Lung Cancer 98.9%, and Prostate Tumor 89%. Accuracy results obtained are higher than previous studies, which only use the Random Forest algorithm as gene selection and classification, so it can be inferred from this research prove that the clustering approach can be used to remove redundancy dimension in microarray data.

Keywords: Microarray, high dimensional, dimensional reduction, feature selection, feature extraction, clustering, k-means, information entropy, classification, random forest.