ABSTRACT

Information retrieval (IR) systems are responsible for retrieval of large amounts of data in an efficient manner. The information retrieval task deals with retrieving data stored within the repositories in response to a user’s information needs. For this type of retrieval system, a performance system is necessary. This research explores the data mining techniques in order to identify the one that will offer the best performance in application to bioinformatics that will respond to genetic information. Data mining has been exploited to retrieve the valuable information in a wide spread fields, especially, in DNA microarray technology. The DNA microarray technology produces a huge amount of gene data, i.e., expression levels of thousands of genes for a very few samples.

Initially, performance issues for information retrieval in the field of microarray gene data are analyzed. The information retrieval metrics like accuracy, sensitivity, specificity are to be increased and error rate is to be reduced. As a first process in gene classification, the high dimensionality of the microarray gene data is reduced using LPP. The LPP is chosen for the dimensionality reduction because of its ability of preserving locality of neighborhood relationship. Secondly, SVM has been trained for effectual gene classification. SVM has the ability to learn with very few samples and so it is selected. SVM ensemble was utilized for the classification process in order to classify more than a single class. Thirdly, the dimensionality diminution process is carried out in order to shrink the microarray data without losing information with the aid of Principal Component Analysis
(PCA) which is one of the techniques utilized for the dimensionality reduction process.

Fourthly, a comparative study is made with the LPP and PCA-based gene retrieval techniques. The technique LPP possesses better accuracy and lesser error rate than PCA technique. Hence, the classification is developed with the blending of dimensionality reduced technique LPP and SVM results in effectual and powerful classification of gene expression data. Finally, technique to improve the retrieval performance in genomic data from commonly used clustering techniques is obtained. Fuzzy C-means has provided more accuracy, correlation and less distance and error rate rather than the k-means gene clustering technique. More accuracy and less error rate leads to effective clustering of the given microarray gene data to the actual class of the gene. Hence, the gene classification has paved the way for effective information retrieval in the microarray gene expression data.