Abstract

The advent of DNA microarray technology has enabled biologists to monitor the expression levels of thousands of genes with a single experiment. This helps in classifying diseases according to varying expression levels in normal and tumour cells, uncovering gene-gene relationships, and identifying genes responsible for the development of diseases. However, such experiments often result in the evaluation of a large number of features on a small number of samples. The selection of the most relevant features and making use of the limited data samples are the key issues in microarray classification. Data mining techniques have proven to be useful in understanding gene function, gene regulation, cellular processes and subtypes of cells. Most data mining algorithms developed for gene expression deal with the problem of classification and clustering. The purpose of this thesis is to study various supervised and unsupervised ensemble methods for analyzing microarray data. Ensemble methods are characterized by their ability to deal with small sample size and high dimensionality; hence they have been widely accepted as an effective method for microarray data analysis.

The first contribution of this work is an exhaustive and comprehensive survey of the supervised, unsupervised and semi-supervised methods and their applicability to analyzing microarray and high dimensional gene expression data. A cost effective supervised ensemble method has been developed which is not influenced by the biasness of the base classifiers. A meta-ensemble has also been developed using the supervised ensemble and the experimental results establish the effectiveness of the proposed model, validated over nine cancer datasets.

To improve upon the analysis of the results that were obtained from the supervised ensemble method proposed by us, a method involving unsupervised ensemble methods has been developed to integrate existing biological knowledge, such as the GO database, to assist in the cluster validation process of cancer datasets. The clustering results arrived at are validated using internal validity measures and external validity measures such as semantic and sequence similarity measures. The approach was tested on several benchmark cancer datasets and the experimental results have been found to be excellent.

The thesis also incorporates an evaluation and review of eight state-of-the-art techniques used for prediction of protein complexes. To take care of the limitations of traditional protein complex prediction approaches as uncovered during the experimental study, an ensemble framework has been developed that integrates Gene Ontology information with PPI network data. The approach is able to identify larger, denser clusters with improved biological significance, by combining the output of several algorithms and thus improve the overall prediction accuracy.

Keywords: Gene expression, cancer data, protein interactions, classification, clustering, ensemble, semantic similarity, sequence similarity, protein complexes.