Abstract

The inter-networked society has been experiencing an explosion of biological data. However, the explosion is paradoxically acting as an impediment to acquiring knowledge. The meaningful interpretation of these large volumes of biological data is increasingly becoming difficult. This analysis is very much crucial to elucidate several secrets of life and several aspects of medical sciences. The most efficient method to investigate these data is via laboratory experiments, but it involves lots of time, money, and manpower. So, effective and efficient computational tools are needed to store, analyze, and interpret these diverse types of biological data. Data mining bridges this gap. Data mining techniques are of two types: data management techniques and data analysis techniques. Among the different data mining techniques, machine learning based data mining techniques are widely used to mine biological data.

Among the different types of bio-molecular data, gene expression data is highly impactful. Microarray techniques such as DNA and high density oligonucleotide chips are powerful biotechnologies as they are able to record the expression levels of thousands of genes simultaneously. Microarray data analysis has great impact in a number of studies over a broad range of biological and medical disciplines, including identification of functions of novel genes, identification of pathway in gene regulatory network, cancer classification by class discovery and prediction, identification of unknown effects of a specific therapy, identification of genes relevant to a certain diagnosis or therapy, and cancer prognosis etc. So, microarray gene expression data analysis plays an important role in real life applications.

Due to several shortcomings of microarray experiments, considerable missing values (MVs) are introduced in the resultant matrix. Such incomplete matrices pose a problem in analysis algorithms as they need complete matrices. It is not feasible to repeat microarray experiments as they are overwhelmingly costly. So, designing algorithms for predicting these missing values accurately have become a mandatory preprocessing step before analysis.

The challenge in this thesis, overall, is to devise powerful machine learning methodologies by symbiotically combining different tools to mine gene expression data in more efficient ways. In this regard, this thesis presents some new supervised and unsupervised learning methodologies, which are efficient in terms of prediction accuracy. The proposed methodologies are used to solve certain problems related to DNA microarray gene expression data.

In the first and second works of this thesis, clustering techniques are used to develop new imputation methods for prediction of missing values more accurately. In the first work, we have introduced a new robust framework which is embedded in the *K*-nearest neighbor imputation method (*K*NNimpute), as well as its different versions to achieve better neighborhood formation, in order to improve the prediction accuracies. Apart from this a new version of *K*NNimpute is proposed here. From the experimental results it has been found that in each and every case, the proposed modified method with new framework significantly outperform their corresponding

traditional versions and are also comparable with the existing robust numerical methods. In the second work we have given focus on developing a new imputation method via combining clustering and numerical approach to improve prediction accuracy. Existing numerical methods are very complex and hard to implement. The proposed work is simple compared to existing numerical methods and shows its superiority.

In third and fourth work new supervised and unsupervised learning techniques are developed to analyze gene expression data in more efficient manner. Here, in the third work, we have designed a novel framework using different partition-based clustering algorithms (mainly different versions of *K*-Means) to provide an intuitive model for eliminating noise and also generating functional gene clusters. The model is also capable of clustering genes without using any predefined *K* as *K* is automatically detected here. The effectiveness of the algorithm, along with a comparison with other algorithms, is demonstrated on different microarray datasets.

In the fourth work, a new ensemble machine learning classification model named Multiple Filtering and Supervised Attribute Clustering algorithm based Ensemble Classification model (MFSAC-EC) is proposed to classify cancer samples more accurately which can handle class imbalance problem and high dimensionality of microarray datasets. The superiority of the algorithm, along with a comparison with other algorithms, is demonstrated on different microarray data sets.