UGC Minor Research Project Report on

Analysis of Microarray Data using Artificial Intelligence Based Techniques

F.No. 42-1019/2013(SR)

Submitted to

University Grants Commission
Bahadurshah Zafar Marg
New Delhi-110002

By

Dr. Khalid Raza
Assistant Professor
Department of Computer Science
Jamia Millia Islamia, New Delhi
kraza@jmi.ac.in | www.kraza.in

April 2015

Abstract

Due to rapid advancement in microarray technology it is possible to measure expression of tens of thousands of genes simultaneously and as a result we have flood of data that need to be analyzed for the discovery of fruitful knowledge. Due to advancement in artificial intelligence based approaches, such as computational intelligence, it is possible to analyse microarray data and infer fruitful information for better understanding of functioning of genes and proteins that lead to better diagnosis and therapy of various diseases.

Prostate cancer is among the most common cancer in males and its heterogeneity is well known. Its early detection helps making therapeutic decision. There is no standard technique or procedure yet which is full-proof in predicting cancer class. The genomic level changes can be detected in gene expression data and those changes may serve as standard model for any random cancer data for class prediction. Various techniques were implied on prostate cancer data set in order to accurately predict cancer class, including machine learning techniques. Huge number of attributes and few numbers of samples in microarray data leads to poor machine learning, therefore the most challenging part is attribute reduction or non-significant gene reduction. This project report is organized in six chapters. In Chapter 2, we have compared several machine learning techniques for their accuracy in predicting the cancer class i.e., Tumor or Normal. Attribute reduction is absolutely required in order to make the data more meaningful as most of the genes do not participate in tumor development and are irrelevant for cancer prediction. Combination of statistical techniques such as inter-quartile range and t-test were used and were effective in filtering significant genes and minimizing noise from data. Further, a comprehensive evaluation of ten state-of-the-art machine learning techniques was done for measuring their accuracy in class prediction of prostate cancer consists of 12600 genes and 102 samples for training, test data set has the same number of genes but 34 different samples. After applying inter-quartile range followed by a t-test statistics for attribute reduction we got 856 most significant genes. Different machine learning techniques were trained with these significant genes. Out of those techniques Bayes Network out performed with an accuracy of 94.11% followed by Navie Bayes. To cross validate our results, we modified our training dataset in six different ways and found that average sensitivity, specificity,

precision and accuracy for Bayes Network is highest among all other machine learning techniques used. The results were compared with the others works on the same kind of dataset and it was found that our results are better than others.

In Chapter 3, we applied clustering techniques to cluster genes. Clustering is a well-known unsupervised learning approach that clubs a set of similar objects in groups that forms clusters. Here, we applied four different types of clustering such as k-means, hierarchical, density-based and expectation maximization approaches, on five different kinds of cancerous gene expression data (lung, breast, colon, prostate, breast and ovarian cancer) for their analysis.

Reconstruction of gene interaction network from gene expression profiles is an important task in systems biology research. Reconstruction of GRNs or 'reverse-engineering' is a process of identifying gene interaction networks from experimental microarray gene expression profile through computation techniques. In Chapter 4, we tried to reconstruct cancer-specific GRNs using information theoretic approach, i.e. mutual information. The considered microarray data consists of large number of genes with 20 samples, where 12 samples were from colon cancer patient and 8 where from normal cell. The data has been preprocessed and normalized. A t-test statistics has been applied to filter differentially expressed genes. The interaction between filtered genes has been computed using mutual information and ten different networks has been constructed with varying number of interactions ranging from 30 to 500. We performed the topological analysis of the reconstructed network, revealing a large number of interactions in colon cancer. Finally, validation of the inferred results has been done with available biological databases and literature.

Inferring key interaction in a given gene interaction network is an important task. The key interactions play an important role in identifying biomarkers for disease that further helps in drug design. Ant colony optimization (ACO) is a nature-inspired swarm-based optimization algorithm that has been used in many optimization problems. In Chapter 5, we applied ACO algorithm for inferring the key interactions in a GRN from gene expression profiles. The algorithm has been tested on two different benchmark datasets and observed that it successfully identify some important key gene interactions. Finally, in Chapter 6 we concluded the project report and presented direction for future works.