Chapter 51
New Hybrid Gene Selection-Sample Classification Method in Microarray Data

Chandra Das
Netaji Subhash Engineering College, India

Shilpi Bose
Netaji Subhash Engineering College, India

Sourav Dutta
Netaji Subhash Engineering College, India

Kuntal Ghosh
Indian Statistical Institute, India

Samiran Chattopadhyay
Jadavpur University, India

ABSTRACT

The gene expression dataset generated by DNA microarray technology contains expression profiles of huge quantities of genes for very small samples. Among these genes, a very small number of genes are informative for cancer sample identification and classification. Informative genes finding is an essential task of microarray gene expression data analysis. Here, a new hybrid gene selection-sample classification model (NHGSSC) is proposed for selection of relevant genes and classification of cancer samples. The NHGSSC performs two tasks-gene selection and sample classification. For gene selection, a new hybrid single filter and α-depth limited best first search based single wrapper method (SFα-BFSSW) is proposed. From these subsets, highly informative genes are selected by counting frequency of occurrence (FO) of every gene. Then SFα-BFSSW method-based ensemble classifier (SFα-BFSSWEC) is built by combining the classifiers created for the selected gene subsets. Experimental results demonstrate the superiority of the NHGSSC to other existing models.

DOI: 10.4018/979-8-3693-3026-5.ch051

INTRODUCTION

Research in the medical science related field is very crucial as it will be beneficial for revelation of the secrets of life or to find new information for preventing certain currently non-curable diseases. The increasing and advancement of new high throughput technologies in the field of engineering, and application of some of those techniques in the field of medical science is the key reason of rapid progress in the diagnosis and taking preventive measures for occurrence of deadly diseases. Proper analysis and utilization of information obtained from the human body can improve medicine quality and as a consequence human health and life quality can be improved.

From a human body huge amount of biological information can be collected and due to advancement of new high-throughput technologies enormous amount of biological information are collected from different human bodies and gathered in different databases by different scientific organizations. To analyze this huge amount of biological information, laboratory experimentations are the most effective methods but these methods are costly, labor dependent and very time consuming to handle these enormous volume of biological data. So, computational methods and techniques are applied to analyze this large volume of biological data. A new field named Bioinformatics has evolved to apply advanced computational techniques to analyze molecular biology related data (in the form of DNA sequence, protein sequence, gene expression data etc.) to satisfy the above-mentioned issues.

Since last few decades DNA microarray technology (Schena et al., 1995) has been becoming one of the most important high throughput bio-technology to researchers involved in biomedical field. Rapid advances in DNA microarray technology (Schena et al., 1995), has improved its prospect significantly toward cancer diagnosis and prognosis (Schena et al., 1995, Golub et al.,1999), which is one of the vital research area in biomedical field. Cancer (Schena et al., 1995, Golub et al.,1999) is a deadly disease and involves different genes, proteins, and many different pathways. Although a tremendous effort of research is carried out in this field, still now information extracted by this research is not sufficient enough to understand the whole mechanism of cancer development process. Experts in medical field are successful to cure cancer patients in case of early detection. At early stage, cancer treatment is much more beneficial for the cancer patients but detection of cancer at early stage is a great challenge yet. However, later stage cancer treatment is not so advanced and as a consequence is not effective so much. So, research in this field has been carried out rigorously. For this purpose analysis of data generated by DNA microarray technology is very crucial still now.

The main feature of microarray technology (Schena et al., 1995) is that it can help to asses expression level of several thousands of genes for a number of different samples or for different time point values of a particular sample. The outcome of DNA microarray technology is a gene expression data matrix, in which every row denotes a sample, every column denotes a gene, and a matrix cell entry represents expression level for the sample corresponds to a gene correspond to a sample.

Among several scientific analysis tasks of microarray data, sample classification that means classification of cancer sample and normal samples or classification of several types of cancer samples is a very important task (Golub et al.,1999, Huawen et al., 2010, Maji et al., 2012, Maji, 2012 & Ang, 2015). This type of analysis has great importance in diagnosis and predicting proper therapy for cancer. A number of classifiers (Ying Lu, 2003) in the past like Fisher's linear discriminant analysis, Weighted Voting of Informative Genes - GS Method, decision tree, nearest neighbor, naïve bayes, support vector machine, neural network based methods etc. have been applied for this task and still now researchers have been continuing their research work in this field for developing new classifiers to improve prediction accuracy.