Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 50 (2015) 52 – 57

Procedia Computer Science

2nd International Symposium on Big Data and Cloud Computing (ISBCC'15)

Feature Selection of Gene Expression Data for Cancer Classification: A Review

Rabindra Kumar Singh [1] , Dr. M. Sivabalakrishnan [2]
[1] Assistant Professor rabindrakumar.singh@vit.ac.in,
[2] Associate Professor sivabalakrishnan.m@vit.ac.in
School of Computer Science and Engineering
VIT Univeristy , Chennai Campus

Abstract

The DNA microarray technology has capability to determine the levels of thousands of gene simultaneously in a single experiment. Analysis of gene expression is important in many fields of biological research in order to retrieve the required information. As time progresses, the illness in general and cancer in particular have become more and more complex and complicated, in detecting, analyzing and curing. We know cancer is deadly disease. Cancer research is one of the major area of research in medical field. Predicting precisely of different tumor types is a great challenge and providing accurate prediction will have great value in providing better treatment to the patients. To achieve this, data mining algorithms are important tools and the most extensively used approach to achieve important feature of gene expression data and plays an important role for gene classification. One of major challenges is to discover how to extract useful information from huge datasets. This paper presents recent advances in the machine learning based gene expression data analysis with different feature selection algorithms.

Gene expression profiles, which represent the state of a cell at a molecular level, have great potential as a medical diagnosis tool. But compared to the number of genes involved, available training data sets generally have a fairly small sample size for classification. These training data limitations constitute a challenge to certain classification methodologies. Feature selection techniques can be used to extract the marker genes which influence the classification accuracy effectively by eliminating the un wanted noisy and redundant genes This paper presents a review

Peer-review under responsibility of scientific committee of 2nd International Symposium on Big Data and Cloud Computing (ISBCC'15)
doi:10.1016/j.procs.2015.04.060

of feature selection techniques that have been employed in micro array data based cancer classification and also the predominant role of SVM for cancer classification.

Peer-review under responsibility of scientific committee of 2nd International Symposium on Big Data and Cloud Computing (ISBCC'15)

Keyword – Gene Expression , Cancer Classification, Feature selection

1. Introduction

Feature selection is an active research area in pattern recognition, statistics, and data mining communities. The main idea of feature selection is to choose a subset of input variables by eliminating features with little or no predictive information. Feature selection can significantly improve the comprehensibility of the resulting classifier models and often build a model that generalizes better to unseen points. Further, it is often the case that finding the correct subset of predictive features is an important problem in its own right. For example, physician may make a decision based on the selected features whether an expensive surgery is necessary for treatment or not.

DNA Microarray
Microarray technology is a developing technology used to study the expression of many genes at once. It involves placing thousands of gene sequences in known locations on a glass slide called a gene chip. A sample containing DNA or RNA is placed in contact with the gene chip. Complementary base pairing between the sample and the gene sequences on the chip produces light that is measured. Areas on the chip producing light identify genes that are expressed in the sample.

Microarray technology provided an opportunity for the researchers to analyze thousands of gene expression profiles simultaneously that are relevant to different fields including medicine especially cancer. The categorization of patient gene expression profile has become a common study in biomedical research. The real problem is managing microarray data with its dimension. Since the dimension of microarray is large, classifying and handling the algorithms becomes too complex to study the gene expression characteristics. Due to the presence of more improper attributes in the dataset, the accuracy of the classification algorithm also gets affected significantly. The aim of feature selection algorithm is to isolate the most important features from the microarray data to minimize the feature space in order to improve the accuracy of the classification.

A microarray gene expression data set can be represented in a tabular form, in which each row represents to one particular gene, each column to a sample or time point, and each entry of the matrix is the measured expression level of a particular gene in a sample or time point, respectively.