



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

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

Rabindra Kumar Singh¹, Dr. M. Sivabalakrishnan²

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

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.

© 2015 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

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.

DNA microarrays are created by robotic machines that arrange large amounts of hundreds or thousands of gene sequences on a single microscope slide. Researchers have a database of over 40,000 gene sequences that they can use for this purpose. When a gene is activated, cellular machinery begins to copy certain segments of that gene. The resulting product is known as messenger RNA (mRNA), which is the body's template for creating proteins. The mRNA produced by the cell is complementary, and therefore will bind to the original portion of the DNA strand from which it was copied.

To determine which genes are turned on and which are turned off in a given cell, a researcher must first collect the messenger RNA molecules present in that cell. The researcher then labels each mRNA molecule by using a reverse transcriptase enzyme (RT) that generates a complementary cDNA to the mRNA. During that process fluorescent nucleotides are attached to the cDNA. The tumor and the normal samples are labeled with different fluorescent dyes (Fig. - 1). Next, the researcher places the labeled cDNAs onto a DNA microarray slide. The labeled cDNAs that represent mRNAs in the cell will then hybridize – or bind – to their synthetic complementary DNAs attached on the microarray slide, leaving its fluorescent tag. A researcher must then use a special scanner to measure the fluorescent intensity for each spot/areas on the microarray slide.

If a particular gene is very active, it produces many molecules of messenger RNA, thus, more labeled cDNAs, which hybridize to the DNA on the microarray slide and generate a very bright fluorescent area. Genes that are somewhat less active produce fewer mRNAs, thus, less labeled cDNAs, which results in dimmer fluorescent spots. If there is no fluorescence, none of the messenger molecules have hybridized to the DNA, indicating that the gene is inactive. Researchers frequently use this technique to examine the activity of various genes at different times. When co-hybridizing Tumor samples (Red Dye) and Normal sample (Green dye) together, they will compete for the synthetic complementary DNAs on the microarray slide. As a result, if the spot is red, this means that that specific gene is more expressed in tumor than in normal (up-regulated in cancer). If a spot is Green, that means that that gene is more expressed in the Normal tissue (Down regulated in cancer). If a spot is yellow that means that that specific gene is equally expressed in normal and tumor.

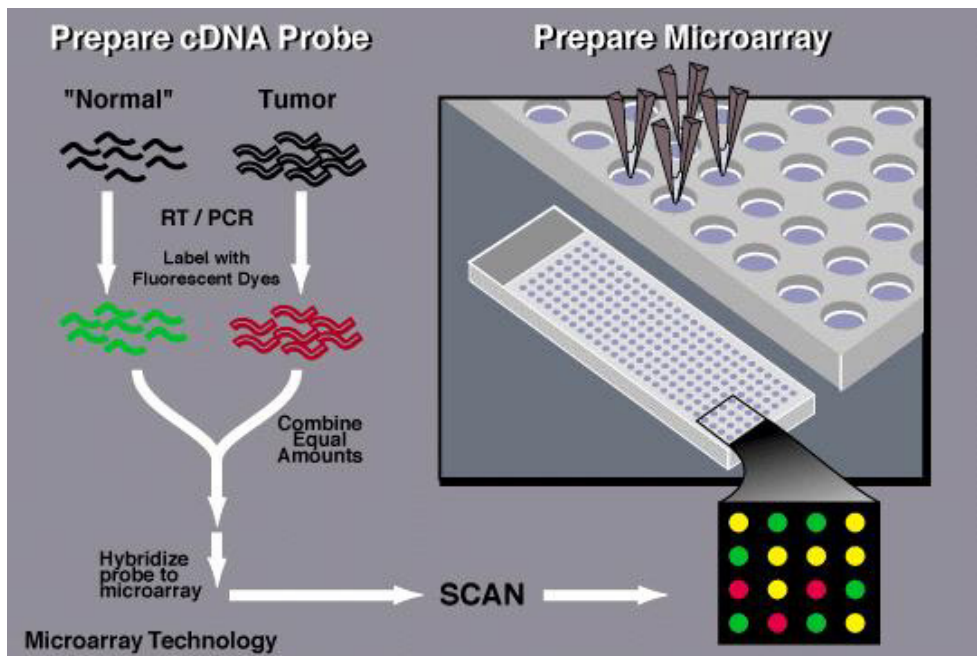


Fig. – 1 Microarray Technology

2. Related Work

Cancer is one of most deadly diseases and lot of people die world- wide because of cancer. As per the WHO statistics in 2014 more than 14 million new cases were identified and around 8.2 million cancer related death occur. The number of new cases is expected to rise by about 70% over the next 2 decades[1]. It has been identified long ago that cancer occurs because of gene disorder. Gene expression is nothing but level of production of protein molecules defined by a gene. Monitoring of gene expression is one of most fundamental approach in genetics. The technique for measuring gene expression is to measure the mRNA instead of protein, because mRNA sequences hybridize with their complementary RNA or DNA sequence while this property lacks in protein. The DNA arrays, are novel technologies that are designed to measure gene expression of tens of thousands of genes in a single experiment. Gene expression data usually contain a large number of genes (in thousands) and a small number of experiments (in dozens). In machine learning terminology, these data sets are usually of very high dimensions with undersized samples. The purpose of Gene selection is to find a set of genes that best discriminate biological samples of different types. The selected genes are “biomarkers,” and they form a “marker panel” for analysis. For analyzing the marker panel rank based scheme such information gain was used. It was observed that the information gain with large group was not accurate, therefore in paper [2] they proposed model-based approach to estimate the entropy on the model, instead of on the data themselves. Here, they used multivariate Gaussian generative models, which model the data with multivariate normal distributions.

Feature selection method

There are two types of feature selection methods have been studied: filter methods [3] and wrapper methods [4].

Filter methods are essentially data preprocessing or data filtering methods. Features are selected based on the intrinsic characteristics that determine their relevance or discriminative powers with regard to the target classes.

In wrapper methods, feature selection is “wrapped” around a learning method: the usefulness of a feature is directly judged by the estimated accuracy of the learning method. Wrapper methods typically require extensive computation to search for the best features.

Basic feature selection algorithm

Input:

S - Data sample f with features X , $|X| = n$

J - Evaluation measure to be maximized

GS – successor generation operator

Output:

Solution – (weighted) feature subset

L: = Start Point(X);

Solution: = {best of L according to J };

Repeat

L: = Search Strategy (L, GS (J), X);

X':= {best of L according to J};

If $J(X') = J(\text{Solution})$ or $(J(X') = J(\text{Solution}) \text{ and } |X'| < |\text{Solution}|)$ then

Solution: =X';

Until Stop (J, L).

The discriminating criteria is being used by filter method for feature selection. The correlation coefficient or statistical test like t-test or f-test is used to filter the features in the filter feature selection method.

Many interesting results were obtained by researchers aiming to distinguish between two or more types of cells (e.g., diseased versus normal, or cells with different types of cancers), based on gene expression data in the case of DNA microarrays. Since microarray data have large amount of data and attributes, which makes complex for researcher to do analysis. A small subset of genes is easier to analyze as opposed to the set of genes available in DNA microarray chips. Therefore it is important to focus on very few genes to give insight into the class association for a microarray sample. It also makes it relatively easier to deduce biological relationships among them as well as to study their interactions.

In paper [5] they obtained feature selection algorithms for classification with tight realizable guarantees on their generalization error. The proposed approaches are a step toward which are more general learning strategies that combine feature selection with the classification algorithm

and have tight realizable guarantees. They classified microarray data, where the attributes of the data sample correspond to the expression level measurements of various genes was considered. They chosen decision stumps as learning bias, which is in part been motivated by this application.

[6] In this paper they introduced an evolutionary rough feature selection algorithm for classifying microarray gene expression pattern. Microarray data typically consist of large number of redundant features; therefore an initial redundancy reduction of attributes was done to enable faster convergence. The main aim was to retain only those genes that play a vital role in discerning between objects. Rough set theory was employed to generate reducts, which represent the minimal sets of non redundant features capable of discerning between all objects, in a multiobjective framework.

3. Conclusion:

This review of feature selection algorithms shows that the feature selection algorithm consistently improves the accuracy of the classifier. Each feature selection methodology has its own advantages and disadvantages. Each algorithm has different behavior which shows that using single algorithm for different dataset is infeasible. The feature selection algorithms are one which decides the accuracy of the classification of different datasets. The feature selection algorithm must select the relevant features and also remove the irrelevant and inconsistent features which cause the degradation of accuracy of the classification algorithms. Feature selection algorithm is playing a major role in accurate classification of large data set like gene expression. Therefore proper cancer classification can be achieved using feature selection algorithms, and on time and accurate treatment may be provided to the patients.

4. References:

- [1] <http://www.who.int/mediacentre/factsheets/fs297/en/>
- [2] Shenghuo Zhu, Dingding Wang, Kai Yu, Tao Li, and Yihong Gong “Feature Selection for Gene Expression Using Model-Based Entropy” *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2010 , pp 25-36
- [3] P. Langley, “Selection of Relevant Features in Machine Learning,” *Proc. AAAI Fall Symp. Relevance*, pp. 140-144, 1994.
- [4] R. Kohavi and G.H. John, “Wrappers for Feature Subset Selection,” *Artificial Intelligence*, vol. 97, no. 1-2, pp. 273-324, 1997.
- [5] Mohak Shah, Member, IEEE, Mario Marchand, and Jacques Corbeil “Feature Selection with Conjunctions of Decision Stumps and Learning from Microarray Data” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 34, No. 1, January 2012
- [6] Mohua Banerjee, Sushmita Mitra “Evolutionary Rough Feature Selection in Gene Expression Data” *IEEE Transactions on Systems, Man, And Cybernetics—Part C: Applications And Reviews*, Vol. 37, No. 4, July 2007