

A Comparative Study of Microarray Data Analysis for Cancer Classification

Kshipra Chitode
Research Student
Government College of Engineering
Aurangabad, India

Meghana Nagori
Asst. Professor, CSE Dept
Government College of Engineering
Aurangabad, India

ABSTRACT

Cancer is most deadly human disease. According to WHO 7.6 million deaths (around 13% of all deaths) in 2008 were caused by cancer. A Cancer diagnosis can be achieved with gene expression microarray data. Microarray allows monitoring of thousands of genes of a sample simultaneously. But all the genes in gene expression data are not informative. The relevant gene selection/extraction is the main challenge in microarray data analysis. Microarray data classification is two stage process i.e. features selection and classification. Feature selection techniques are used to extract a small subset of relevant genes without degrading the performance of classifier. The classifier uses these extracted relevant genes for cancer classification. In this review paper there is a comparative study of the feature selection and classification techniques. The evaluation criteria are applied to find out the best combination of feature selection and classification technique for accurate cancer classification

General Terms

Cancer classification

Keywords

Microarray, cancer, genes, feature selection, classification

1. INTRODUCTION

DNA microarray is an advancement in molecular genomics technologies which allows detecting expression of thousands of genes in a small sample [1]. The gene expression data from different tissues of a sample makes it possible to monitor the genes, which is helpful in diagnosis of many diseases like cancer.

On the other hand, there are two major challenges in the analysis of microarray they are small data sample available from few numbers of patients (often less than hundred) and high dimensional dataset (nearly thousands or tens of thousands of genes) [2].

In order to diagnose a disease using microarray, the gene expression data are used. All the gene expression values of a sample are not relevant to the disease, there is high redundancy. There are many genes which contain irrelevant information for the accurate classification of disease. Selection of relevant gene is most important in microarray gene expression analysis. The gene selection method reduces a large dimensional microarray dataset into a small set of genes which can classify the sample as cancer and normal samples [3].

The microarray classification is a two stage process. First stage deals with feature selection which produce a reduce set of features and the second is a classification which uses the extracted features for the classification. The feature selection techniques are broadly classified into three categories: filter,

wrapper and embedded methods [4]. Filter methods rank genes by using the intrinsic characteristics of gene expressions with the class label. Filters are of two type univariate and multivariate. The univariate filter considers each feature individually ignoring feature dependencies. Chi-square, t-statistics, Information Gain (IG), Gain Ratio (GR), Signal-to-Noise ratio are univariate filters. The multivariate filters take feature dependencies into consideration. Correlation based feature selection (CFS) is a multivariate filter. The wrapper methods interact with the classifier while ranking the genes. In embedded methods the gene selection process is embedded in learning/constructing classifier [5]. Some of the classification techniques are Support Vector Machine (SVM), K- Nearest Neighbors (KNN), Naïve Bayes (NB), Neural Network (NN) and Decision Tree (DT) [6].

In this work, the authors are comparatively studying the feature selection techniques and the classification techniques. The most efficient combination of feature selection and the classification technique leading to accurate classification of cancer is found out. The proposed system considers a different number of top ranked genes for classification. At the time of gene selection along with the gene expression value proposed system is using of biological information of the gene. There are public biological databases like KEGG, Gene Ontology (GO) are available which provides gene annotations.

The rest of the paper is organized as follows. In Section II, contents the feature selection and classification techniques. In Section III, the proposed work. The evaluation methods which will be used for the selection of best performance combination feature selection and classification techniques for cancer diagnosis are described in Section IV.

2. LITERATURE SURVEY

2.1 Microarray

Microarray is a surface on which sequences from thousands of different genes are covalently attached to a fixed location. There are two classes of the microarray. The cDNA array spots of complementary DNA apply to a glass slide (or nylon membranes) and Affymetrix array place many thousands of gene specific oligonucleotides (called probes) synthesized directly on silicon chip. After hybridization the microarray is scanned and converted to numeric data [7].

2.2 Feature selection

Feature selection techniques over DNA microarray focus on filter methods. Most of the proposed techniques are univariate i.e. each feature is considered separately. The feature relevance score is calculated, and low scoring features are removed. The top ranked genes are used to build the classifier. Following are the feature selection techniques.

2.2.1 Chi-square

Chi-square is a univariate filter method of feature selection. The statistic value of the feature (gene) is measured individually with respect to the class.

$$\chi^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i}$$

where, O_i and E_i is observed and expected expression value of i^{th} gene respectively [8].

2.2.2 Signal-to-Noise (SNR) Ratio

It is the ratio of difference of means of the two classes to the summation of standard deviations of two classes.

$$SNR = \frac{\mu_i + \mu_j}{\sigma_i + \sigma_j}$$

μ_i and μ_j denote the average expression value of i^{th} gene over all samples in normal and tumor case respectively. σ_i and σ_j denote standard deviation of i^{th} gene over all samples in normal and tumor case respectively.

2.2.3 Information Gain (IG)

Entropy value is used in Information Gain, Gain Ratio, and Symmetrical Uncertainty attribute ranking techniques [8]. Entropy of Y is

$$H(Y) = - \sum_{y \in Y} p(y) \log_2(p(y))$$

$p(y)$ is the marginal probability density function for random variable Y. the conditional entropy of Y after observing X is

$$H(Y|X) = - \sum_{x \in X} p(x) \sum_{y \in Y} p(y|x) \log_2(p(y|x))$$

$p(y|x)$ is the conditional probability of y given x. The information gained about Y after observing X is

$$IG = H(Y) - H(Y|X)$$

Information gain is symmetrical measure,

$$IG = H(Y) - H(Y|X) = H(X) - H(X|Y)$$

2.2.4 Gain Ratio (GR)

In order to predict Y we normalize information gain (IG) by dividing it by entropy of X. The Gain_ratio is

$$Gain_ratio = \frac{IG}{H(X)}$$

Gain ratio is an unsymmetrical measure. Gain_ratio falls in range [0, 1] because of normalization. Gain_ratio=1 indicates that X completely predicts Y and Gain_ratio=0 indicates that X and Y are independent.

2.2.5 Symmetrical Uncertainty (SU)

It is ratio of information gain IG and summation of entropy of X and Y

$$SU = 2 \frac{IG}{H(Y) + H(X)}$$

SU value is in range [0, 1] because of correction factor 2. Correction factor is used for normalization. SU=1 indicates

that knowledge of one completely predicts other and SU=0 indicates that X and Y are independent.

2.2.6 Fisher's Criteria

The equation used to rank the genes in fisher criteria is

$$fisher(g) = \frac{(m_1(g) - m_2(g))^2}{s_1^2(g) + s_2^2(g)}$$

m_1 and m_2 denote the mean expression value of g^{th} gene over all samples in tumor and normal case respectively. s_1 and s_2 denote standard deviation of g^{th} gene over all samples in tumor and normal case respectively[9].

2.2.7 Relief-F

Relief-F is an extension to relief algorithm. It assigns weight to each feature according to its ability of distinguishing among classes. Feature with higher weights are selected for classification purpose [8]. It finds the weight of each feature f by finding a good estimate of the following probability

$$Wf = \frac{P(\text{different value of } f \mid \text{different class})}{P(\text{different value of } f \mid \text{same class})}$$

2.2.8 Clustering based and Network analysis based Method

Shang Gao, Omar Addam and colleagues in [10] proposed these two feature reduction techniques. The proposed clustering based method uses Genetic Algorithm (GA) based clustering approach.

2.2.9 Wilcoxon based Rank Sum Method

Instead of using observed data values, it is sorted in ascending or descending order. Each data item is assigned a rank corresponding to position in the sorted list. Rather than observed data these ranks are used for further analysis.

2.2.10 Support Vector based Correlation Coefficient (SVccREF)

This technique selects support vector data points using support vector machine (SVM). These selected data points are further used for ranking the genes using correlation coefficient. The top ranked genes are used for classification.

2.3 Classification

Classification and clustering are in general considered as similar; the only difference is classification is a supervised learning process while clustering is an unsupervised learning process.

In classification the class label of each training tuple is known in advance hence called as supervised learning. The classifier built using the known instances (training set) to successfully predict the class of new instances (test set).

The accuracy of the classifier is determined as the percentage of test tuples that are correctly classified by the classifier from the test set. In other words, the classification is all about predicting the class of the new instance by learning from known instances and their class labels. SVM, NB, NN, KNN, DT are classification technique [4], [6], [8].

2.3.1 Support Vector Machine (SVM)

SVM is a popular and powerful classifier for microarray data. It is a kind of supervised learning methods for classification. SVM uses a nonlinear mapping to divide the original training data into two linearly separable classes. SVM determines an optimal hyperplane and divides the given training set of labeled sample as positive and negative labeled sample with

the maximum margin of separation. The main objective of SVM is to select optimal hyperplane as there may exist many hyperplanes which can divide the training set. The two parameters in hyperplane equation are w and b . The separating hyperplane equation is written as

$$w \cdot X + b = 0$$

where w is a weight vector and b is a scalar often referred as a bias. The samples closest to the hyperplane are considered as support vectors and play a greater role in classifying the test samples.

Given training set of instance-label pairs (x_i, y_i) , $i=1, \dots, l$, where x_i is training sample with y_i class label and l is the number of samples in the training set. The SVM classifier requires solving the following optimization problem

$$\min_{w,b,\varepsilon} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^l \varepsilon_i$$

subject to

$$y_i(w \cdot x_i + b) \geq 1 - \varepsilon_i, \varepsilon_i \geq 0, i = 1, \dots, l$$

where C is the SVM sensitivity parameter, and ε is a slack variable.

2.3.2 Naïve Bayesian

The Naïve Bayesian (NB) a probabilistic classification technique makes use of Bayes' theorem. Consider m class classification problem and let x denotes an observation with probability density $p(x)$. If class prior probability $p(m)$ and class the conditional probability $p(x|m)$ are known, according to Bayes theorem the posterior probability $p(m|x)$ is

$$p(m|x) = p(x|m) \cdot p(m) / p(x)$$

The Bayesian assigns an observation to the class that has a highest posterior probability

$$\text{Class of } x = \text{argmax}_m p(m|x)$$

2.3.3 K-Nearest Neighbor

KNN is a lazy learning algorithm which means there is no explicit training phase on the training dataset. All the training data are used in the test sample classification. It finds the k closest training points of test sample according to metric like Euclidean distance, Manhattan distance. In order to classify the test sample, class labels of these k closet points are used.

2.3.4 Neural Network

Neural networks are comprised of a set of connected input/output units. In NN each connection has a weight associated with it. In order to predict the correct class labels of input tuple, during the learning phase the network learns by adjusting the weights. Back-propagation neural network is multilayer feed forward neural network having input layer, hidden layers and output layers.

2.3.5 Decision Tree

The Decision tree is a supervised learning process which generates a flow chart like tree structure using the training samples through an iterative process.

ID3, C4.5, Classification and Regression Trees (CART) are decision tree algorithms

3. PROPOSED SYSTEM

In this paper the authors are comparatively studying both feature selection and classification techniques for microarray

data analysis for accurate classification of cancer. The features are extracted from the training dataset and the classifier is built. Actual classification is performed on test dataset.

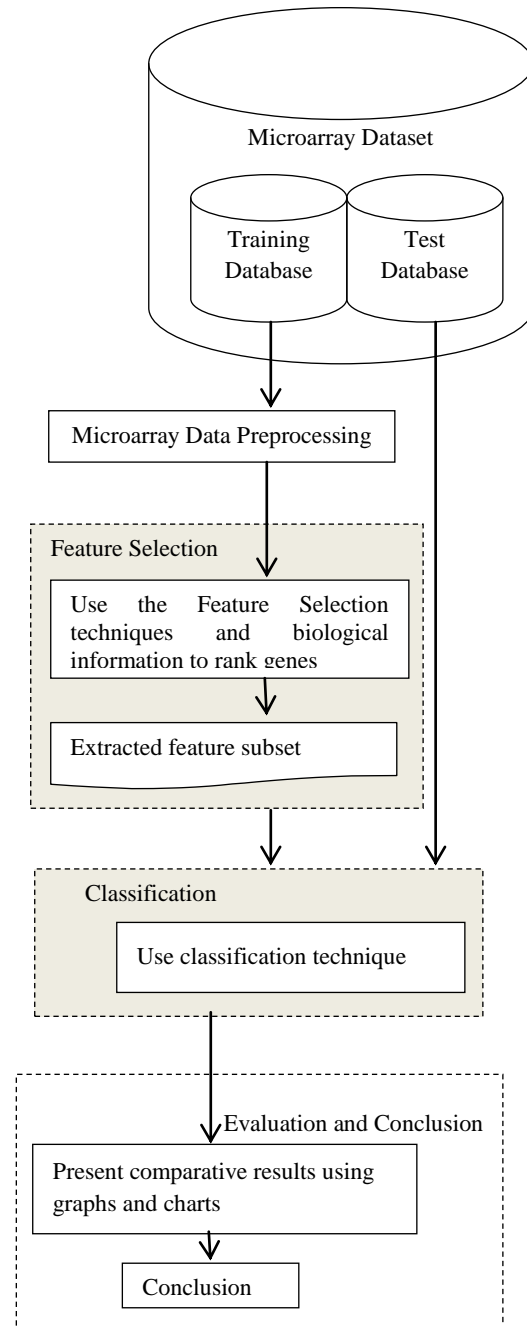


Fig 1: Proposed microarray data analysis system

The workflow of the proposed microarray data analysis system and comparative analysis of different techniques leading to final conclusion is shown in Figure 1. Different number of top ranked genes will be selected in order to improve the performance of the classifier.

There are many online repositories which make the biomedical data set available which include gene expression value, protein informative data and genomic sequence data.

Colon cancer dataset: The dataset contains expression of 2000 genes across 62 samples collected from Colon Tumor

patients. Among these samples, 40 are tumor tissues (labeled as “negative”) and 22 are normal (labeled as positive”). The genes are placed in descending order of minimal intensity.

Acute Myeloid Leukemia (AML) Dataset: The dataset contains 78 samples among which 39 are training set and 39 are testing dataset which is publicly available on Gene Expression Omnibus (GEO) Data Repository.

Table 1. No of samples and genes per sample in dataset

Dataset	No. of genes per sample	No. of training samples	No. of testing samples
Colon Cancer	2000	40	22
Acute Myeloid Leukemia (AML)	22000	39	39

The Affymetrix scanner generates a DAT file the image of scanned array. After image analysis a CEL file (Cell Intensity File) is generated, which contains raw probe level expression data. Signal intensity values are computed using normalization method like Microarray Analysis Suite (MAS 5.0), GCOS or Roust Multi-chip Analysis (RMA). These signal intensity values are used in the proposed system for cancer classification.

Gene Ontology (GO), KEGG are reliable public source of information on genes [11]. These gene annotations can be used to improve performance of feature selection techniques.

The proposed system considers a different number of top ranked genes like 5 genes, 10 genes, 20 genes, 50 genes, 100 genes. According to performance evaluation, combination of efficient feature selection and accurate classification technique is found.

4. EVALUATION METHODS

4.1. AUC (Area under Receiver Operating Characteristic Curve)

The AUC evaluation criteria can be used for the evaluation of the accuracy of various developed classifiers. ROC graph is a technique used for visualizing, organizing and selecting the best classifier based on their performance. For a given classifier and an instance there are four possible outcomes for that instance. (1) True positive (TP) if the instance is positive and it is classified as positive (2) False negative (FN) if positive instance is classified as negative (3) True negative (TN) if the instance is negative and it is classified as negative (4) False positive (FP) if the negative instance is classified as positive.

A confusion matrix can be built from this methodology as

$$\text{Confusion matrix} = \begin{bmatrix} TP & FP \\ FN & TN \end{bmatrix}$$

The upper left or lower right elements of confusion matrix represent the correct decisions made and the upper right and lower left element represent the errors [12].

True positive (TP rate) of a classifier can be estimated as

$$TP \text{ rate} = \frac{\text{Positives correctly classified}}{\text{Total positive}}$$

False positive (FP rate) of a classifier is

$$FP \text{ rate} = \frac{\text{Negatives incorrectly classified}}{\text{Total negative}}$$

The ROC graphs are two dimensional graphs in which FP rate value of classifier is plotted on the X-axis and TP rate value of the classifier is plotted on Y-axis.

Accuracy, sensitivity and specificity are the three measure performance of different classifier [9]. The accuracy of a classifier is the fraction of the correctly classified samples to all samples.

$$\text{Accuracy} = \frac{TN + TP}{TN + FN + FP + TP}$$

Sensitivity of a classifier is a fraction of the real positive samples that are correctly predicted as positive.

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

Specificity of a classifier is a fraction of the real negative samples that are correctly predicted as negative

$$\text{Specificity} = \frac{TN}{TN + FP}$$

4.2. LOOCV and 10 fold Cross Validation

Leave One Out Cross Validation (LOOCV) and 10-fold cross validation is used when genuine dataset not available. The random division of the data set into a training set and test set is performed and different predictors are compared. In 10 fold cross validation the data are randomly divided into 10 mutually exclusive sets of nearly equal size. Each one is treated as a test set in turn and remaining as training set. This process is repeated 10 times to make each set as test set. When a small number of samples are available 4 fold cross validation is also performed [9].

5. CONCLUSION

In this paper authors have comparatively studied various feature selection and classification techniques. The extraction of the most relevant features is important for the accurate cancer classification. The proposed method can be used to find out the most efficient and accurate combination of feature selection and classification technique. In the future, the efficiency of the feature selection process can be improved by using a combination of two or more feature selection techniques to extract most relevant and informative feature from the dataset.

6. ACKNOWLEDGEMENT

The authors feel a deep sense of gratitude to Prof V. P. Kshirsagar HOD of Computer Science and Engineering Department for his motivation and support during this work. The authors are also thankful to the Principal, Government College of Engineering, Aurangabad for being a constant source of inspiration.

7. REFERENCES

- [1] Jun S. Liu Department of Statistics Harvard University "Bioinformatics: Microarrays Analyses and Beyond".
- [2] Alireza Osareh and Bitia Shadgar "Microarray Data Analysis for Cancer classification" IEEE Antalya, Turkey 2009, pp.125-132.
- [3] C. Shang and Q. Shen, "Aiding classification of gene expression data with feature selection: a comparative study." vol. 1, 2005, pp. 68-76.
- [4] Piyushkumar A. Mundra and Jagath C. Rajapakse "Support Vectors Based Correlation Coefficient for Gene and Sample Selection in Cancer Classification" IEEE 2010.
- [5] Yvan Saeys, Inaki Inza and Pedro Larranaga "A review of feature selection techniques in bioinformatics" 2005 pp 1–10.
- [6] "Data Mining: Concepts and Techniques", second edition by Jaiwei Han and Micheline Kamber Chapter 6.
- [7] "Microarray Data Management An Enterprise Information Approach: Implementations and Challenges" by Willy A. Valdivia-Granda and Christopher Dwan, Chapter 6.
- [8] Jasmina Novaković, Perica Strbac, Dusan Bulatovic, "Toward optimal feature selection using ranking methods and classification algorithms" Yugoslav Journal of Operations Research 21 (2011), Number 1, 119-135.
- [9] Azadeh Mohammadi, Mohammad H Saraee, Mansoor Salehi, "Identification of disease-causing genes using microarray data mining and Gene Ontology" Mohammadi et al. BMC Medical Genomics 2011.
- [10] Shang Gao, Omar Addam and colleagues, "Robust Integrated Framework for Effective Feature Selection and Sample Classification and Its Application to Gene Expression Data Analysis" IEEE 2012 pp. 112-119.
- [11] One Huey Fang, Norwati Mustapha, Md. Nasir Sulaiman "Integrating Biological Information for Feature Selection in Microarray Data Classification" 2010 Second International Conference on Computer Engineering and Applications IEEE, pp. 330-334.
- [12] Sri Harsha Vege "Ensemble of Feature Selection Techniques for High Dimensional Data" Western Kentucky University 2012.