Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Software**

Colon cancer prediction with genetic profiles using intelligent techniques

Subha Mahadevi Alladi[1], Shinde Santosh P.[1], Vadlamani Ravi[2,] * and Upadhyayula Suryanarayana Murthy[1]

[1]Bioinformatics Group, Biology Division, Indian Institute of Chemical Technology, Tarnaka, Hyderabad 500007, Andhra Pradesh, India;
[2]Institute for Development and Research in Banking Technology, Castle Hills Road, Masab Tank, Hyderabad 500057, India;
Vadlamani Ravi* - Email: rav_padma@yahoo.com; * Corresponding author

received June 10, 2008; revised August 28, 2008; accepted September 13, 2008; published November 04, 2008

Abstract:
Micro array data provides information of expression levels of thousands of genes in a cell in a single experiment. Numerous efforts have been made to use gene expression profiles to improve precision of tumor classification. In our present study we have used the benchmark colon cancer data set for analysis. Feature selection is done using t –statistic. Comparative study of class prediction accuracy of 3 different classifiers viz., support vector machine (SVM), neural nets and logistic regression was performed using the top 10 genes ranked by the t–statistic. SVM turned out to be the best classifier for this dataset based on area under the receiver operating characteristic curve (AUC) and total accuracy. Logistic Regression ranks as the next best classifier followed by Multi Layer Perceptron (MLP). The top 10 genes selected by us for classification are all well documented for their variable expression in colon cancer. We conclude that SVM together with t-statistic based feature selection is an efficient and viable alternative to popular techniques.

Keywords: gene expression; tumor classification; t-statistic; feature selection; SVM neural network; logistic regression

Background:
DNA micro arrays have enabled researchers to monitor thousands of genes simultaneously. The role of micro array expression data in cancer diagnosis is very significant. Mining for useful information from such micro array data consisting of thousands of genes and a small number of samples is often a tough task. Colon cancer is the second most common cause of cancer mortality in Western countries **[1]**. According to the WHO 2006 report colorectal cancer causes 655,000 deaths worldwide per year. Precise predictions of tumors are very important for treatment and diagnosis.

All the genes used in the expression profile are not informative; also many of them are redundant. Reducing the number of genes by feature selection and still retaining best class prediction accuracy for the classifier is vital in case of tumor classification. The emphasis in cancer classification is both on methods of gene selection and on choice of classifier. Towards the objective of selecting the features important for colon cancer classification several methods like t-statistic **[2]**, Fisher's F statistic, Principal component analysis, SVM-RFE are in use. Several machine learning techniques have also been successfully applied, for example decision trees, naïve Bayesian methods and support vector machines.

Furey and colleagues have used signal to noise ratio for feature selection and SVM as classifier resulting in 90.3% accuracy in prediction **[3]**. Li and colleagues have made use of an approach combining GA and KNN to identify genes that can jointly discriminate between the tumor and normal classes **[4]**. It is a stochastic supervised pattern recognition method and they have achieved 94.1% accuracy with it. Sun and colleagues have used wavelet transformation to reduce feature space to a lower dimension and classified colon cancer data with PNN to obtain 92% accuracy **[5]**. Chen and colleagues used multiple kernel support vector machine (MK-SVM) scheme, consisting of feature selection, rule extraction and prediction modeling to improve the explanation capacity of SVM. They used two gene expression datasets viz., leukemia dataset and colon tumor dataset to demonstrate the performance of this approach. Using the small number of selected genes, MK-SVM achieves encouraging classification accuracy of more than 90% for both two datasets **[6]**. Kim and colleagues have used information gain and tested evolutionary neural networks for their prediction model **[7]**. Mahata and colleagues have ranked genes depending on the minimum probability of classification errors (MPE) for each gene and classification accuracy was obtained using SVM and a modified naïve Bayes classifier **[8]**.

In our study, we have worked on the prediction of colon cancer, based on gene expression data. We have used the benchmark colon cancer dataset having expression pattern for 40 tumor and 22 normal colon tissue samples analyzed with Affymetrix Oligonucleotide array **[9]** to compare 3 different classifiers. Feature extraction is done using t–statistic. SVM, Neural nets and Logistic Regression are employed and their performance in terms of classification accuracy on the micro array data was compared.

Methodology:
Dataset
The colon cancer data set was taken from Kent Ridge Biomedical Data Repository **[10]**. It has gene expression samples that were analyzed with an Affymetrix

ISSN 0973-2063 (online) 0973-2063 (print)
Bioinformation 3(2): 130-133 (2008)

Oligonucleotide array complementary to more than 6500 human genes. The data set contains 62 samples collected from colon-cancer patients. Among them, 40 tumor biopsies are from tumors (labeled as "negative") and 22 normal (labeled as "positive") biopsies are from healthy parts of the colons of the same patients. Two thousand out of around 6500 genes were selected based on the confidence in the measured expression levels. The source of the dataset did not mention the number for the confidence of the expression levels. The data of all samples in a micro array are presented in a table constructing the gene expression matrix. The rows of the matrix correspond to the single genes and the columns to the single samples. This gene expression matrix is the input to a classification system.

Data preparation

The first step in the analysis of micro array experiments is the normalization of the data. The purpose of normalization is to adjust for any bias arising from the variation in micro array technology rather than from biological differences between the RNA samples or the printed probes. The micro array expression data is also highly heterogeneous. We have standardized the data in order to reduce the range over which calculations need to be made. We have taken the normal samples as class 0 and tumor samples as class 1.The equation for the standardized value is: $Z=[X-\mu]/\sigma$, where X is the attribute to be standardized, μ is the arithmetic mean of the attribute and σ is its standard deviation.

Feature selection

Feature selection is the technique commonly used in machine learning, to select a subset of relevant features for building robust learning models. When applied in biology domain, the technique is also called discriminative gene selection which detects influential genes based on DNA micro array experiments. By removing most irrelevant and redundant features from the data, feature selection helps improve the performance of learning models. Feature selection identifies the subset of differentially expressed genes that are potentially relevant for distinguishing the classes of samples. The aim is to reduce the initial gene pool from 7,000–10,000 to 100–200.

(a)Table comparing classifier performance with different number of Genes

No of genes		10	50
		P %	P %
LR		78.18	78.18
NN	MLP	80.91	76.36
	RBF	62.73	67.27
SVM	Linear	81.81	74.44
	polynomial	75.45	80
	RBF	85.45	82.72
	Sigmoid	85.45	63.63

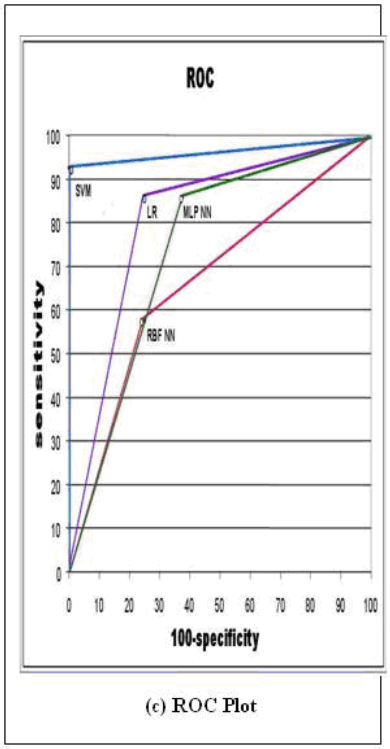
P%= Prediction accuracy

(b) 10 fold Cross validation for SVM.

	Linear	polynomial	RBF	Sigmoid
Set1	83.87%	77.41%	83.87%	83.87%
Set2	83.87%	79.03%	83.87%	87.09%
Set3	83.87%	77.41%	83.87%	87.09%
Set4	85.48%	75.80%	83.87%	83.87%
Set5	83.87%	77.41%	83.87%	87.09%
Average	84.19%	77.41%	83.87%	85.80%

% represents cross validation accuracy.

SVM- Support Vector Machine, **LR**-Logistic Regression, **MLP**-Multi Layer Perceptron, **RBF NN**-Radial Basis Function Neural Network.



(c) ROC Plot

Figure 1: (a) Table comparing classifier performance with different number of genes; (b) 10 fold cross validation for SVM; (c) ROC plot.

We have used t -statistic for selecting the top genes to be used by the classifiers. Each sample belongs to one class 0 (normal) or 1 (tumor). For each gene we have calculated the mean and standard deviation for both the classes. Assuming unequal variances a score **t** is given by the formula in equation 1 (see supplementary material). The top 50 genes are identified on the basis of t statistic of each of the 2000 genes originally considered and used for classification.

ISSN 0973-2063 (online) 0973-2063 (print)
Bioinformation 3(2): 130-133 (2008)

KNIME and Lib-SVM software

For analyzing the classification efficiency of Logistic regression and neural nets we have used the LR, MLP and RBF options available in the software tool KNIME (a freely available software package) **[11]**. For analyzing colon cancer dataset with SVM we used the Lib-SVM 2.85 tool (a freely available software package). **[12]**.

Class prediction procedure

The dataset was split into training and test set in the ratio 80:20, 70:30, 65:35 and 60:40. We obtained best prediction accuracy with the 65:35 split. Consequently, we continued training and prediction using the classifiers with the 65:35 data sets. Five different 65:35 data sets were generated set1, set2, set3, set4, set5) and used to predict the average classification accuracy of the models. We applied Logistic Regression, neural net (MLP, RBF) and SVM to the colon cancer data. 10 fold cross validation was performed with SVM.

Specificity, sensitivity and ROC curve

The different classifiers are compared against one another on the basis of specificity, sensitivity, total accuracy and the value of area under the ROC curve (AUC). Sensitivity of each classifier was calculated using the formula (TP/TP+FN) where TP is number of true positives and FN is number of false negative cases. Specificity is measured by the formula (TN/TN+FP) where TN is number of true negatives tested and FP is the number of false positive cases.

Discussion:

The top 50 genes are identified on the basis of t –statistic of each of the 2000 genes originally considered. We compared the average prediction accuracy using 10 and 50 genes and found less error in prediction using 10 genes. The top 10 genes are ranked in ascending order and their GenBank Accession numbers are H08393, X63629, M22382, J05032, H40095, M63391, M26697, T56604, X12671, T47377. Figure 1a represents the Table comparing SVM classifier performance with different classifiers. Figure 1 (b) presents the results of 10 fold cross validation for SVM and finally figure 1c depicts the ROC Plot.

Figure 1a presents the comparison of classifier performance with different number of features selected. 10 fold cross validation was performed on the full data set with out splitting it and used in Lib-SVM. This was done with each of the five sets and an average cross validation percentage was generated, as shown in Figure 1b. Specificity and sensitivity of the different classifiers are evaluated for the colon cancer data set with 10 genes. The results obtained are used to plot an ROC curve in the Figure 1 (c). The results have shown the performance of SVM to be better than that of Logistic Regression and Neural net (see Figure 1a and 1b). The area under the ROC curve shows significantly better results for SVM with a radial basis function indicating good sensitivity and specificity.

Further, it is worth noticing that five out of 10 discriminating genes are present in the genes which were selected by using information gain **[7]** as a method for feature selection, four are present in the list selected using MPE **[8]**. M63391 DES gene has been discovered to be down regulated in colon cancer patient samples **[9]**. This has also been verified by biological experiments **[13]**. HSP D1 whose gene accession number is M22382 is expressed in both primary tumor and lymph node metastasis **[14]**.

Conclusion:

As reported (3,6) in earlier literature and as shown by our study, the potential of applying machine learning techniques is very high for classification of malignancy in tumors on the basis of variation in gene expression. We also demonstrate the superior prediction accuracy of SVM over neural net and logistic regression classifiers in case of the colon cancer data set. An important point is the question regarding what the significant features or patterns mean from a biological perspective. We can point out the genes, which give best prediction accuracy in case of classification, correlating them to their biological significance with respect to the disease. Developing more sophisticated methods of feature selection coupled with SVM would yield more insights into defining a better binary classification model for this biological problem.

References:

[01] D. A. Rew, *Euro. J. Surg. Oncol.*, 27: 504 (2001)
[02] H. Liu *et al.*, *Genomic Informatics*, 13: 51 (2002)
[03] T. S. Furey *et al., Bioinformatics*, 906 (2000) [PMID: 11120680]
[04] L. Li *et al.*, *Bioinformatics*, 17: 1131 (2001)
[05] G. Sun *et al.*, *Neurocomputing*, 69: 387 (2006)
[06] Z. Chen et al., *Artif Intell Med.,* 161 (2007) [PMID: 17851055]
[07] K. J. Kim and S. B. Cho, *Neurocomputing*, 61: 361 (2004)
[08] P. Mahata and K. Mahata, *Biomedical Informatics*, 40: 775 (2007) [PMID: 17950675].
[09] U. Alon *et al.*, *Proc. Natl. Acad. Sci. USA*, 6745 (1999) [PMID: 10359783]
[10] http://datam.i2r.a-star.edu.sg/datastes/krbd/ColonTumor/ColonTumor.html
[11] http://www.knime.org/downloads.html
[12] http://www.csie.ntu.edu.tw/~cjlin/libsvm
[13] G. M. Groisman *et al.*, *Histopathology*, 48: 43 (2006) [PMID: 16487365].
[14] F. Cappello *et al.*, *BMC Cancer*, 5 (2005) [PMID: 16253146]

Edited by P. Kangueane
Citation: Alladi *et al.*, Bioinformation 3(3): 130-133 (2008)

ISSN 0973-2063 (online) 0973-2063 (print)
Bioinformation 3(2): 130-133 (2008)

Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Prediction Model**

Supplementary material

Equation 1

$$Tf[i] = \frac{|\mu^{+}_{i} - \mu^{-}_{i}|}{\sqrt{(\sigma^{+}_{i})^2/n_{+} + (\sigma^{-}_{i})^2/n_{-}}}$$

where μ^{+}_{i} and μ^{-}_{i} are the mean expression values for classes 1 and 0 respectively, $(\sigma^{+}_{i})^2$ and $(\sigma^{-}_{i})^2$ are the variances for classes 1 and 0 respectively and n_{+} and n_{-} are the class specific number of input samples.

ISSN 0973-2063 (online) 0973-2063 (print)
Bioinformation 3(2): 130-133 (2008)