International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:4, No:2, 2010

An SVM based Classification Method for Cancer Data using Minimum Microarray Gene Expressions

R. Mallika, and V. Saravanan

***Abstract*—**This paper gives a novel method for improving classification performance for cancer classification with very few microarray Gene expression data. The method employs classification with individual gene ranking and gene subset ranking. For selection and classification, the proposed method uses the same classifier. The method is applied to three publicly available cancer gene expression datasets from Lymphoma, Liver and Leukaemia datasets. Three different classifiers namely Support vector machines-one against all (SVM-OAA), K nearest neighbour (KNN) and Linear Discriminant analysis (LDA) were tested and the results indicate the improvement in performance of SVM-OAA classifier with satisfactory results on all the three datasets when compared with the other two classifiers.

***Keywords*—**Support vector machines-one against all, cancer classification, Linear Discriminant analysis, K nearest neighbour, microarray gene expression, gene pair ranking.

I. Introduction

MICROARRAY technology provides a tool for estimating expressions of thousands of genes simultaneously [1].Many supervised learning methods have been proposed with the steps as follows: Firstly, three-fourth of the samples of data is used to train the Classifier and secondly, the trained classifier is used to predict or test the one-fourth of the samples. The challenges in such a problems were discussed in [7] as 1.Scarce of training data, 2.High dimensionality. The answer to the challenge lies in predicting cancer by using a small subset of important genes from wide collection of gene expression data. With thousands of genes and small amount of samples ranking the genes according to their importance in contributing to classifier's prediction strength is a crucial problem [9].

In 1999, Golub et al [4] gave a classification method for Leukaemia by scaling the sum of deviations of positive and negative classes. A good Feature selection method applied prior to classification will produce better and promising prediction accuracy. Feature selection is to select K dimensions out original d dimensions that can best represent original dataset.

R. Mallika is with the Department of Computer Science, Sri Ramakrishna College of Arts and Science for Women, Coimbatore, India (e-mail: mallikapanneer@hotmail.com).

V. Saravanan is with the Department of Computer Applications, Karunya University, Coimbatore, India (e-mail: tvsaran@hotmail.com).

In 2003, Tibshirani [11] successfully classified the lymphoma data set with only 48 genes by using a statistical method called nearest shrunken centroids and used 43 genes for SRBCT data. Lipo Wang [7] in 2007 proposed an algorithm in finding out minimum number of gene up to 3 genes with best classification accuracy using C-SVM and Fuzzy Neural Networks (FNN). Li-Yeh Chuang [8] in 2009 proposed a two stage feature selection method using various cancer datasets. In the first stage all genes were ranked and in the second stage fixed number of gene subsets were ranked using particle swam optimisation and those gene subsets were classified.

Tzu-Tsung Wong et al [13] summarised the classification method using two major options and proposed the method as two stage classification method for microarray data. The paper proposed Gene selection mechanism with individual gene ranking or gene subset ranking, subsequently the selection of a classification tool with or without dimensionality reduction procedures.

This paper proposes an effective classification method with the pre-processing the gene expression data, ranking, selection and classification. In the pre-processing step individual genes are ranked and the top ranked genes were selected and gene pair (gene subset) ranking were performed. During the selection step, top ranked gene pairs were used to train the classifier. The gene pairs that achieved 100% training accuracy were selected and used to retrain the classifier and the learning results of the resultant classifier were observed.

II. Methodology

A. Gene Ranking -ANOVA p-values

Analysis of Variance (ANOVA) is a technique, which is frequently used in the analysis of microarray data, e.g. to assess the significance of treatment effects, and to select interesting genes based on P-values. [5]. The ANOVA test is known to be robust and assumes that all sample populations are normally distributed with equal variance and all observations are mutually independent.

The approach chosen in this paper is the one-way ANOVA that performs an analysis on comparing two or more groups (classes) for each gene and returns a single p-value that is significant if one or more groups are different from others.

The most significantly varying genes have the smallest p-values. Of all the information presented in the ANOVA table, if the p value for the F- ratio is less than the critical value (α), then the effect is said to be significant. In this paper the α value is set at 0.05, any value less than this will result in significant effects, while any value greater than this value will result in non-significant effects. The very small p-value indicates that differences between the column means are highly significant.

The probability of the F-value arising from two identical distributions gives us a measure of the significance of the between-sample variation as compared to the within-sample variation. Small p-values indicate a low probability of the between-sample variation being due to sampling of the within-sample distribution, small p-values indicate interesting genes. This paper uses the p-values for individual gene ranking and pair wise gene ranking.

B. Classifiers

1. Support Vector Machines-One against All (SVM- OAA)

SVMs are the most modern method applied to classify gene expression data, which works by separating space into two regions by a straight line or hyper plane in higher dimensions. SVMs were formulated for binary classification (2 classes) but cannot naturally extend to more than two classes. SVMs are able to find the optimal hyper plane that minimizes the boundaries between patterns [10]. SVMs are power tools used widely to classify gene expression data [3][14]. How to effectively extend SVM for a multi-class classification is still an ongoing research issue [1]. This paper gives effective methodology to classify a multi-class problem .To extend SVM for multi-class classification, SVMs were designed with SVM *one-against-one*, SVM *one-against-all*. This paper efficiently uses SVM with heavy tailed RBF.

The SVM-OAA constructs 'n' binary SVM classifier with the i^{th} class separating from all other classes. Each binary SVM classifier creates a decision boundary, which can separate the group it represents from the remaining groups.

2. K- Nearest Neighbour (KNN)

K-Nearest neighbour is the simplest method for deciding the class to which a sample belongs and a popular nonparametric method. KNN classifies a new object based on attributes and training samples. To classify an unclassified vector X , the KNN algorithm ranks the neighbours of X amongst a given set of N data (Xi Ci),i=1,2…N and uses the class labels Cj (j=1,2..k) of the K most similar neighbours to predict the class of the new vector X.

The classes of these neighbours are weighted using the similarity between X and each of its neighbours measured by the Euclidean distance metric. Then X is assigned the class label with the greatest number of votes among K nearest class labels.

3. Linear Discriminant Analysis (LDA)

LDA otherwise known as FLDA (Fishers Linear Discriminant Analysis calculates a straight line or hyperplane that separates 2 known classes. Unlike SVM where the hyper plane is chosen to minimize the misclassification errors, LDA chooses hyper plane to minimize within class variance on either side of the line and minimize the between-class variance. Then the side of the line or hyper plane determines the class of the unknown sample. The disadvantage in using LDA is that it can perform classification well only for linearly separable data.

III. Results and Discussions

The proposed methodology was applied to the publicly available cancer database namely Liver, Lymphoma and Leukaemia cancer databases. The lymphoma dataset with 4026 genes and 62 samples with 3 classes namely DLBCL, CLL and FL, the subtype of Lymphoma cancer. Liver dataset (http://genome-www.stanford.edu/hcc) has 2 classes HCC and non-tumour liver for 1648 genes for 156 observations in which 82 are from HCC and 74 from non-tumour livers. Unlike the Lymphoma dataset, which were with the prediction of subtypes in Lymphoma cancer, the Liver dataset were from the cancerous and non-cancerous tissue. The problem is to predict whether the samples are from cancerous or non-cancerous tissue. The Leukaemia dataset contains 7129 genes and 2 classes, with the gene expression data ALL, and AML, subtypes of Leukaemia.

This section reports the experimental results of all the datasets exhibiting the SVM, KNN and LDA classifiers. The dataset was with few missing data. The K-Nearest neighbour algorithm as used by [12] with k=3 was used and the missing data were filled. Half of the samples were picked randomly for training and all the samples for testing. For the training dataset with m x n dimensions ANOVA p-value was calculated for each gene and the top ranked genes were selected .All possible combinations of the top n genes were generated. For n number of top genes all possible combinations are n (n+1)/2. All these combinations were ranked again using ANOVA p-values. Repeated random selection of samples was performed to aim for maximum gene pairs that can achieve no errors in training. This means that the method filters the best pairs for classification and prediction. The gene pairs that were above the threshold value were used to train the classifier.

The performance of the classifier was validated using cross validation (CV) technique with 5-folds. For 5 folds, the samples in the training dataset was randomly divided into 5 equal parts, classification was performed for 5 runs, using the 4 parts as training and the other as testing. Each time the classifier was trained, a different test set was used, so that over 5 runs of the classifier, all the samples were used as test set. The gene pairs that achieved 100% 5-fold CV accuracy were selected and used for classification using the three classifiers SVM, KNN and LDA. In all the cases SVM one-against-all was superior to all other methods were well proven. Fig. 1 illustrates the procedure used.

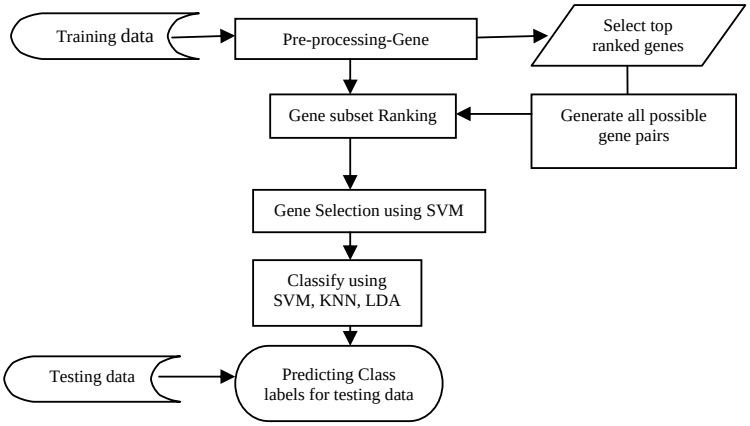


Fig. 1 The procedure of the method used in this paper

The procedure was repeated for the top 10, 20, 30, 50 and 100 genes. It was observed that the 5 fold CV accuracy varied from 80.65% to 100% for the lymphoma data and for the Liver data ranging from 95.67% to 100% and Leukaemia data with a range of 89.54% to 100% for the top 30 genes. All these results were achieved using the SVM classifier. It should be noted that all the individual genes after ranking were given numbers in ascending order. Table I depicts the results of the lymphoma dataset showing the average misclassification error (average error rate) and 5 fold Cross Validation accuracy in training using top selected 30 genes. The averaged 5 fold accuracy for all the three datasets are shown in Table II and Fig. 2 plots the average error rate in training using top 10,20,30,50 and 100 genes for all the three datasets. It is clear from the graph the error rate increases corresponding to the number of genes trained. Hence the training was performed up to top 100 genes.

TABLE I
LYMPHOMA DATASET RESULTS DEPICTING AVERAGED CV ACCURACY AND ERROR RATE

Gene Pairs	Average Misclassification Error	5-fold CV Accuracy in Training
1,5	0.10	90.32
1,6	0.06	93.55
2,6	0.13	87.10
3,6	0.16	83.87
4,5	0.00	100.00
4,6	0.00	100.00
4,8	0.03	96.77
4,21	0.06	93.55
5,7	0.10	90.32
:	:	:
21,23	0.06	93.55
21,24	0.06	93.55
21,27	0.10	90.32
21,30	0.16	83.87
Averaged Training Accuracy:		94.00%
Average error rate:		0.060%

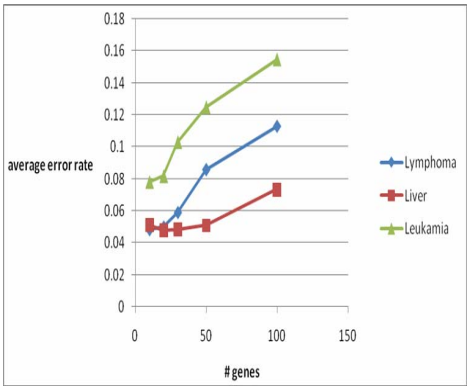
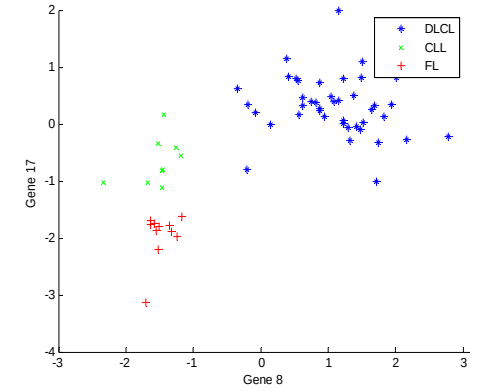


Fig. 2 Average error graph of the top selected genes for all three dataset

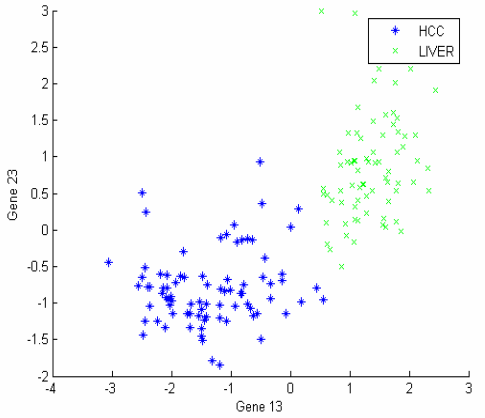
TABLE II
AVERAGED 5 FOLD CV ACCURACY RESULTS OF ALL THREE DATASETS FOR THE TOP SELECTED 30 GENES

Dataset	**Averaged 5fold CV accuracy**
Lymphoma data	94%
Liver data	95.67%
Leukaemia data	89.54%

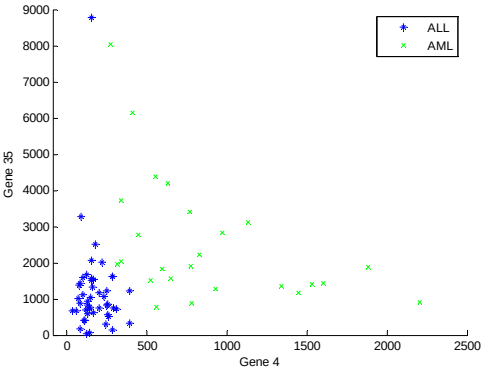
To visualise the gene expression values, graphs were plotted for few selected gene pairs which showed 100% CV accuracy results in training. Fig. 3(a, b, c) shows the graph plotted for all the three datasets used in this paper. From the Plot in Fig. 3 (a) the gene pairs (3, 17) for lymphoma dataset a clear separation of all the three subtypes DLCL, CLL and FL are shown. Similarly Fig. 3 (b) showing a separation for (13,23) gene pairs and Fig. 3(c) for Leukaemia data the plots are given for the gene pairs (4,35). In all the plots, all the different subtypes are clear and boundaries can be drawn.



(a) Lymphoma



(b) Liver



(c) Leukaemia

Fig. 3 Plots showing separation and boundaries for all the three datasets

The gene pairs that achieved 100% CV accuracy were used for retraining the SVM-OAA classifier. Since no training is required for KNN and LDA, all the gene pairs were classified and the results were noted. In all the cases, it was noted that SVM-OAA obtained a maximum testing accuracy of up to 100% for the Lymphoma dataset using the top ranked 10, 20, 30 and 50 individual genes. Hence the paper analysis the results using the top 50 genes and the results are analysed. Table III shows the maximum testing accuracy and averaged prediction accuracy for all gene pairs for all the three classifiers for top 50 ranked individual genes.

TABLE III
MAXIMUM TESTING ACCURACY RESULTS OF ALL THE THREE CLASSIFIERS FOR ALL DATASETS

	Lymphoma	Liver	Leukaemia
SVM-OAA	100%	97.44%	95.83%
KNN	98.39%	96.15%	95.83%
LDA	80.65%	94.23%	87.50%

It is clear that of all the three classifiers SVM-OAA performed the best for lymphoma data, but for the liver and leukaemia data the KNN and SVM-OAA classifiers both produced the same accuracy. The average testing accuracy results in Table IV proves for better performance of SVM-OAA classifier than the other two classifiers.

For predictive performance evaluation of all three classifiers, the average testing accuracy of all the three datasets using top selected genes are shown in Table IV. It is inferred that the performance of SVM-OAA classifier shows a marginal improvement of 0.33% compared to KNN classification and considerable improvement of 12.48% with LDA classifier. Table V reports the results of the proposed method and the best result among those contained in previous papers on cancer classification and feature ranking, selection with various classifiers. The results compared have been using different validation techniques and different partition of data for training and testing. The results indicate that the proposed method is competitive with other recent classification methods for the datasets used in this paper. It should be noted that the proposed method in this paper used just 2 genes to best predict the types of cancer.

TABLE V
RESULT COMPARISON OF BEST AVERAGED TESTING ACCURACY WITH PREVIOUS WORKS

	Lymphoma	Leukaemia
Proposed method	100%	95.83%
Li et al (KNN)[8]	94%	94%
Dudoit et al (1nn,LDA,CART)[2]	95%	95%
Elena (SVM) [3]	93%	100%

TABLE IV
AVERAGED TESTING ACCURACY COMPARISON

		10	20	30	50	100	Average
Lymphoma	SVM-OAA	100%	97.32%	96.01%	93.76%	91.66%	95.75%
	KNN	97.31%	96.63%	93.97%	93.13%	91%	94.41%
	LDA	66.67%	95.00%	71.59%	65.17%	58.85%	71.46%
Liver	SVM-OAA	97.44%	96.96%	95.19%	87.34%	86.69%	92.72%
	KNN	96.79%	97.28%	93.91%	84.12%	87.73%	91.97%
	LDA	96.79%	96.47%	92.31%	87.33%	84%	91.38%
Leukaemia	SVM-OAA	95.83%	88.61%	91.39%	88.66%	83.72%	89.64%
	KNN	95.83%	92.36%	90.28%	88.77%	86.41%	90.73%
	LDA	88.89%	78.30%	78.54%	70.72%	72.66%	77.82%

REFERENCES

[1] Chih-wei Hsu and chih jen Lin.2002, A Comparison of methods for multiclass Support vector machines, IEEE transactions on neural networks.
[2] Dudoit, S., Fridlyand, J., & T. P. Speed.2002. Comparison of discrimination methods for the classification of tumor using gene expression data. Journal of the American Statistical Association, 97, 77–87.
[3] Elena Marchiori, Michele Sebag.2005. Bayesian learning with local support vector Machines for cancer classification with gene expression data", Evo Workshops PP.74-83.
[4] T. R. Golub, D. K. Slonim and P. Tamayo et al.1999. Molecular classification of cancer:class discovery and prediction by gene expression, monitoring Science, 286:531-7.
[5] Juan Liu, Hitoshi Iba.2001. Selecting Informative Genes with Parallel Genetic algorithms in Tissue Classification, Genome Informatics 12: 14–23.
[6] Li, Dardern, Weinberg, Levine, and Pedersen.2001. Gene assessment and sample classification for gene expression data using genetic algorithm/k-nearestneighbour method.Combinatorial Chemistry and High Throughput Screening,4(8):727-739.
[7] Lipo Wang, Feng Chu, and Wei Xie .2007. Accurate Cancer Classification Using Expressions of Very Few Genes, IEEE/ACM Transactions on computational biology and bioinformatics, vol. 4, no. 1, January-march.
[8] Li-Yeh Chuang,Chao-Hsuan Ke,Hsueh-Wei Chang,Cheng-hong Yang 2009. A two-stage Feature selection method for gene expression data, OMICS A journal of Integrative Biology, Volume.13, number 2.
[9] Mao Yong,Zhou Xiao-bo,PI Dao-ying,Sun You-xian et al.2005.Parameters selection in gene selection using Gaussian Kernel support vector machines by genetic algorithm, Journal of Zhejiang University SCIENCE 6B(10):961-973.
[10] Mingjun Song, Sanguthevar Rajasekaran 2007. A greedy correlation-incorporated SVM-based Algorithm for gene selection, 21st International Conference on Advanced Information Networking and Applications workshop, IEEE.
[11] R. Tibshirani, T. Hastie, B. Narasimhan, G. Chu .2003. Class Prediction by Nearest Shrunken Centroids with Applications to DNA Microarrays, Statistical Science, vol. 18, pp. 104-117.
[12] O.Troyanskaya.2001. Missing values estimation methods for DNA Microarrays, Bioinformatics, vol.17, pp. 520-525.
[13] Tzu-Tsung Wong, Ching-Han Hsu.2008.Two-stage classification methods for microarray data, Science Direct, Expert Systems with Applications 34(2008) 375-383.
[14] Yoonkyung Lee, Cheo Koo Lee .2003. Classification of multiple cancer types by Multicategory support vector machines using gene expression data, Vol. 19 no. 9, Bioinformatics.