Microarray Gene Expression Cancer Diagnosis Using Machine Learning Algorithms

A.Bharathi[1], A.M.Natarajan[2]
[1&2]Bannari Amman Institute of technology
[1&2]Tamilnadu (State)
[1]abkanika07@gmail.com, [2]amn@bitsathy.ac.in

Abstract:

In this paper, we use the extreme Learning Machine (ELM) for cancer classification. We propose a two step method. In our two step feature selection method, we first use a gene importance ranking and then, finding the minimum gene subset form the top-ranked genes based on the first step. We tested our two step method in cancer datasets like Lymphoma data set and SRBCT data set. The results in the Lymphoma data set and SRBCT dataset show our two-step methods is able to achieve 100% accuracy with much fewer gene combination than other published results. The results indicate that ELM produces comparable or better classification accuracies with reduced training time and implementation complexity compared to neural networks methods like Back Propagation Networks, SANN and Support Vector Machine methods. ELM also achieves better accuracy for classification of individual categories.

Key Words: Extreme learning machine, Gene expression, Support vector machine, Back Propagation networks.

1. Introduction

In the gene expression profiling-based classification area for cancer diagnosis, binary classification problems have been more extensively studied and only a small amount of work has been dine on direct classification problems. Distinguishing tumor types is an important challenge of cancer treatment. Fortunately, with the developing of the gene chip technology, the possibility of cancer classification and diagnosis at the gene expression level increases. However, the gene expression data usually have high dimension and the samples of patients are small. Some of the genes may be irrelevant to cancer classification. Thus, to obtain good results in classifications of cancer treatment, we should select discriminatory genes and get a small set of genes [2].

Recently, different kinds of feature selection methods have been used to classify cancers using gene expression data, such as t-test [3], relief algorithm [4], Z-score, and principal component analysis [5]. These methods select features which can be maximally distinguished from tens of thousand genes. Then, classification algorithms applied to these selected features. Traditional classification algorithms such as k-nearest neighbor, Naïve Bays, C4.5 have been applied broadly. Conventional neural networks usually produce lower classification accuracy than SVM [6]. Neural Network schemes such as SVM, and BPN which are susceptible to local minima and long training times. To overcome these difficulties, in this paper, we propose using a neural network training algorithm called Extreme Learning machine (ELM) [7] [8] [9] and evaluate it for microarray gene expression cancer diagnosis problems.

We have evaluated the performance of the ELM algorithm on the cancer diagnosis benchmark data sets, namely Lymphoma, and SRBCT data set. The number of categories of the Lymphoma data set is 3. The SRBCT data set [13] includes the expression data set of 2308genes. There are totally 63 training samples and 25 testing samples. For the Lymphoma and SRBCT data set, we have compared the performance of ELM with that of Lipo Wang et al. [3] separately.

In this paper, we applied a two step feature selection method and the Extreme learning machine to the problem of cancer classification. In our two step feature selection method, we first use a gene

importance ranking and then, finding the minimum gene subset form the top-ranked genes based on the first step.

Studies on the Lymphoma data set and SRBCT data set indicate that the total training time for ELM is always lower than that of SVM. However, in terms of the classification accuracy, ELM, SVM have similar performances when the top ten genes with two gene combination.

2. The two-step feature selection method

2.1 Step 1: Gene Importance Ranking

In step 1, we compute the importance ranking of each gene using an Analysis of Variance (ANOVA) method. ANOVA is a powerful statistical technique that is used to compare the means of more than two groups. One way ANOVA is a part of the ANOVA family. When we are comparing the means of more than two populations based on a single treatment factor, then it said to be one way ANOVA. The equation used for one way ANOVA is as follows: $y_{ij} = m + a_i + e_{ij,}$ where this equation indicates that the *jth* data value, from level *i,* is the sum of three components: the common value (grand mean), the level effect (the deviation of each level mean from the grand mean), and the residual.

2.2 Step 2: Finding the minimum gene subset

After selecting some top genes in the important ranking list, we attempt to classify the data set with one gene. We input each selected gene into our classifiers. If no good accuracy is obtained we go on classifying the data set with all possible 2 gene combinations within the selected genes. If still no good accuracy is obtained, we repeat this procedure with all of the 3-gene combinations and so on until we obtain a good accuracy. In this paper, we used the following classifier to test 2-gene combinations and 3-gene combinations.

2.2.1. Extreme Learning Machine (ELM)

In supervised batch learning, the learning algorithms use a finite number of input-output samples for training. Here, we consider N arbitrary distinct samples

$$(Xi, ti) \in R^n \times R^m$$

where xi is an n ×1 input vector and ti is an m ×1 target vector. If an SLFN with N~ hidden nodes can approximate these N samples with zero error, it then implies that there exist βi, ai, and bi such that

$$f_N(X_j) = \sum_{i=1}^{N} \beta_i G(a_i, b_i, X_j) = t_j,$$
$$j = 1, \ldots \ldots \ldots \ldots N.$$

The above equation written compactly as,

$$H\beta = T,$$

Where,

$$H(a1, \ldots, a_N, b_1, \ldots, b_N, x_1, \ldots, x_N)$$

$$\begin{bmatrix} G(a_1, b_1, x_1) & \cdots & G(a_N, b_N, x_1) \\ \vdots & \ddots & \vdots \\ G(a_1, b_1, x_N) & \cdots & G(a_N, b_1, x_N) \end{bmatrix}_{N X N}$$

H is called the hidden layer output matrix of the network [14]; the i^{th} column of H is the i^{th} hidden node's output vector with respect to inputs x1, x2…… xN and the jth row of H is the output vector of the hidden layer with respect to input xj. In real applications, the number of hidden nodes, ~N, will always be less than the number of training samples, N, and, hence, the training error cannot be made exactly zero but can approach a nonzero training error ϵ. The hidden node parameters ai and bi (input weights and biases or centers and impact factors) of SLFNs need not be tuned during training and may simply be assigned with random values according to any continuous sampling distribution [9], [10], [11]. Above Equation then becomes a linear system and the output weights β are estimated as

$$\hat{\beta} = H \uparrow T$$

Where H is the Moore- Penrose generalized inverse [15] of the hidden layer output matrix H. The ELM algorithm, which consists of only three steps, can then be summarized as

ELM Algorithm:

Given a training set

$N = \{(x_i, t_i) | x_i \in R^n, t_i \in R^m, i = 1, \ldots, N\}$,

The universal approximation capability of ELM has been analyzed in Huang et al. [16] using an incremental method and it has been shown that single SLFNs with randomly generated additive or RBF nodes with a wide range of activation functions can universally approximate any continuous target functions in any compact subset of the euclidean space Rn. In this paper, the activation function used in ELM is the sigmoidal function

$$g(x) = \frac{1}{1+e^{-\lambda z}}$$

2.2.2 Algorithm Description

We used five fold cross validation in the experiments because formal training and test datasets are not available for this data set. More specifically, we randomly divide data in each class into five groups. In each fold, data points in four groups are used as a training set, the data points in the remaining group is used as a test set. Hence, we have five folds of the data. The training and test sets in each fold are independent. Moreover, the experiment using data in each fold is done independently. Hence, cross validation is used here for separating the data set into several groups of training and testing sets, not for avoiding over fitting [3].

3. Results

3.1 Experiment 1- Microarray Benchmark data set Lymphoma

In the lymphoma data set there are 42 samples derived from Diffuse Large B-cell Lymphoma (DLBCL), nine samples from Follicular Lymphoma (FL), and 11 samples from Chronic Lymphocytic Leukemia (CLL). The entire data set includes the expression data of 4026 genes. In this data set, a small part of the data is missing. A k-nearest neighbor algorithm was applied to fill those missing values [10]. In the first step, we randomly divided the 62 samples into 2 parts: 31 samples for testing, 31 samples for training. We ranked the entire set of 4,026 genes according to their ANOVA in the training set. Then we picked out the 20 genes with the highest ANOVA. We applied our ELM to classify the lymphoma micro array data set. At first, we added the selected 20 genes one by one to the network according to their ANOVA ranks. That is, we first used only a two gene that is ranked 1 as the input to the network. We trained the network with the training data set and subsequently, tested the network with the test data set. The excellent performance of our ELM motivated us to search for the smallest gene subsets that can ensure highly accurate classification for the entire data set [1]. We first attempted to classify the data set using two gene and three gene tested for all possible combinations within the 20 and 10 genes (See the table 1).

Table 1 Accuracy for ELM in no. of fold is 5

No.of genes	Gene Comb.	Training Acc.	Acc.	Elapsed Time in seconds
10	2	71.85	85.71	2.094
10	3	71.60	83.33	4.813
20	2	66.28	100	8.047
20	3	64.30	100	10.823

3.1.1. Comparison with the SVM

The above procedure can be applied for SVM. Fivefold cross validations are carried out on the 62 samples using ELM, and SVM for this comparison (see the table 2 and fig 2).

Validation Accuracy: The validation accuracy of the different algorithms for the benchmark data set Lymphoma using 62 samples is presented in table 2 and fig. 2. These algorithms include ELM, and SVM. As observed form Table 2 and fig. 2, the ELM achieves a much higher accuracy than SVM. For the ELM and SVM algorithms, classification accuracy tends to grow with the number of gene combination selected. It can be noted that, for all of these selections, ELM achieves the highest classification accuracy.

Table.2 Validation accuracy for ELM, and SVM

No. of genes	Gene Combination	Validation accuracy	
		SVM	ELM
10	2	88.60	89.31
10	3	89.22	89.89
20	2	86.55	89.49
20	3	87.55	89.69

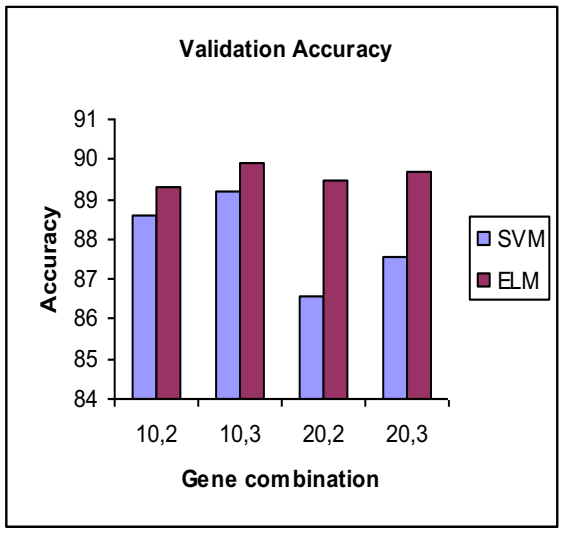

Fig.2: Validation Accuracy for ELM, and SVM

Training time: The total training time including the cross-validation time for ELM, and SVM for 5-fold cross validation for each gene combination is shown in Table 3 and fig. 3. The time given for ELM is based on Matlab 7.0 implementation. ELM takes a much smaller training time than SVM.

Table 3: Training time(s) for ELM and SVM (No.of fold: 5)

No.of genes	Gene Combination	Training time (sec)	
		ELM	SVM
10	2	2.094	22.093
10	3	4.813	64.422
20	2	8.047	97.625

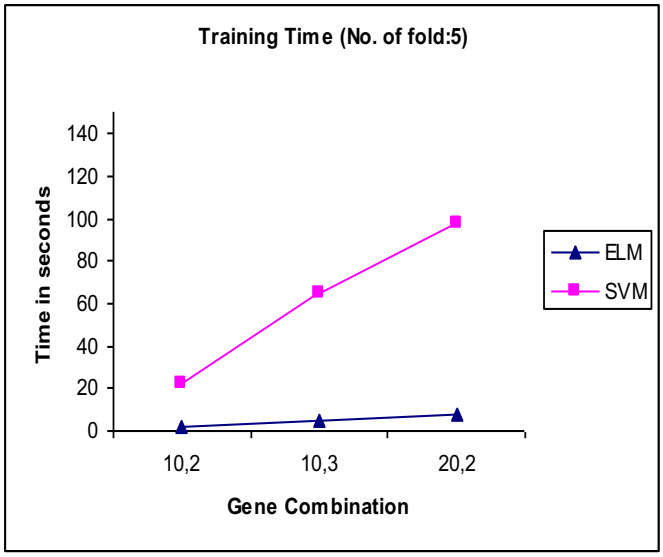

Fig.3: Comparison of Training time(s) of ELM and SVM for 5-fold

The same procedure can be applied to 10-fold cross validation for ELM and SVM for each gene combination is shown in table 4. ELM takes smaller training time than SVM.

Table 4: Training time(s) for ELM and SVM (No.of fold: 10)

No.of genes	Gene Combination	Training time (sec)	
		ELM	SVM
10	2	3.094	46.250
10	3	7.922	137.828
20	2	12.391	190.985

Testing accuracy: For a classifier, the overall classification performance is important. A good classifier is one that produces a good overall classification performance. Table 5 and fig.4 show the classification results for each gene combination for different algorithms. The figures are based on the simulation results of 5-fold cross validation of 62 samples. Table 6 based on the simulation results of 10-fold cross validation of 62 samples.

Table 6 gives the detail of this comparison.

Table 6: The comparison of classification accuracies with different feature selection methods for Lymphoma data set

Feature Selection	Classification Accuracy	Reference
Signal to Noise Ratio	94.1%	[2]
Genetic Algorithm	84.6%	[11]
Principal Component Analysis	94.2%	[12]
T-test	97.1%	[3]
ANOVA with ELM	100%	This paper

3.2 Experiment 2- Microarray Benchmark data set SRBCT

The SRBCT data set [13] includes the expression of 2308 genes. There are totally 63 training samples and 25 resting samples. We first ranked the importance of all the genes with ANOVA method, and then selected top 20 or 10 genes. Then we trained the ELM classifiers with the 63 training samples and tested with 25 testing samples. We first attempted to classify the data set using two gene and three gene tested for all possible combinations within the 20 and 10 genes (See the table 7).

Table 7 Accuracy for ELM in no. of fold is 5

No.of genes	Gene Comb.	Training Accuracy	Acc.	Elapsed Time in seconds
10	2	70.50	84.42	1.994
10	3	72.85	85.67	3.725
20	2	65.38	100	7.215
20	3	66.20	100	9.643

3.2.1. Comparison with the SVM

The above procedure can be applied for SVM. Fivefold cross validations are carried out on the 85 samples using ELM, and SVM for this comparison (see the table 8).

Validation Accuracy: The validation accuracy of the different algorithms for the benchmark data set SRBCT using 85 samples is presented in table 8 These algorithms include ELM, and SVM. As observed form Table 8. the ELM achieves a much higher accuracy than SVM. For the ELM and SVM algorithms, classification accuracy tends to grow with the number of gene combination selected. It can be noted that, for all of these selections, ELM achieves the highest classification accuracy.

Table.8 Validation accuracy for ELM, SVM and BPN

No. of genes	Gene Combination	Validation accuracy	
		SVM	ELM
10	2	87.40	88.63
10	3	88.28	89.98
20	2	85.65	90.43
20	3	88.32	90.89

Training time: The total training time including the cross-validation time for ELM and SVM for 5-fold cross validation for each gene combination is shown in Table 9. The time given for ELM is based on Matlab 7.0 implementation. ELM takes a much smaller training time than SVM.

Table 9: Training time(s) for ELM and SVM (No.of fold: 5)

No.of genes	Gene Combination	Training time (sec)	
		ELM	SVM
10	2	1.994	23.190
10	3	3.725	65.433
20	2	7.215	98.825

3.3 Overall Comparison

3.3.1 Classification accuracy

From the result on the Lymphoma data set and SRBCT data set, we find that ELM achieves better classification result (in a statistical sense) on the two categories.

3.3.2 Training Time

For the lymphoma data set and SRBCT data set, ELM takes much less total training time than SVM algorithms.

4. Conclusion

Selecting important features and building effective classifiers are both pivotal processes to cancer classification. In this paper, we selected genes with our proposed feature selection methods, and then classified the Lymphoma data set and SRBCT data set with ELM classifiers. Compared with previously published methods, we achieved higher classification accuracy with fewer genes. This method, a fast and efficient classification method for cancer diagnosis problem based on the microarray data is presented. Its performance has been compare with other methods such as the SVM algorithm. The SVM algorithm involves greater system complexities and a longer training time but ELM achieves better classification accuracy and less training time.

References

[1]. Runxuan Zhang et al., Multicategory Classification using an Extreme Learning Machine for Microarray Gene Expression Cancer Diagnosis, IEEE/ACM transactions on Computational Biology and Bioinformatics, Vol.4, No.3,2007.

[2]. T.S.Furey, et al., "Support Vector Machine Classification and Validation of Cancer Tissue Samples Using microarray Expression Data", Bioinformatics, 2000, 16.906-914.

[3]. Lipo Wang et al., Accurate Cancer Classification Using Expressions of Very Few Genes, IEEE/ACM transactions on Computational Biology and Bioinformatics, Vol.4, No.1,2007.

[4].Lipo Wang, Nina Zhou, and Feng Chu, "A General Wrapper Approach to Selection of Class-Dependent Features", IEEE Transactions on Neural Networks, 2008, Vol. 19.No.7.

[5].I.T. Joliffe, Principal Component Analysis *[M]*, Springer, New York, 2002.

[6]. J.W. Lee, J.B. Lee, M. Park, and S.H.Song, An extensive comparison of recent classification tools applied to microarray data, Computational statistics and data analysis, Vol. 48, pp.869-885, 2005.

[7]. G.B. Huang, Q.Y Zhu and C.K. Siew, Extreme Learning Machine: A New Learning Scheme of Feed forward Neural Networks, Proc. Int'l Joint Conf. Neural Networks (IJCNN '04) July 2004.

[8]. G.B Huang and C.K Siew, Extreme learning machine: RBF Network Case, Proc. Eighth Int'l Conf. Control, Automation, Robotics, and Vision (ICARCV '04), December 2004.

[9]. G.B Huang and C.K Siew, Extreme learning machine with Randomly Assigned RBF Kernels, Int'l Journal of Information Technology, Vol. 11, No.1, 2005.

[10]. Friedland S., Niknejad A., and Chihara L.,A Simultaneous reconstruction of missing data in DNA microarrays, Institute of Mathematics and its Applications preprint series, No.1948.

[11]. L.Li.C.R. Weinberg, T.A. darden, et al., Gene Selection for Sample Classification Based on Gene Expression Data, Bioinformatics, 2001. pp. 1131-1142.

[12]. D.U.Nguyen, D.M. Rocke, Tumor Classification by partial Least Squares using Microarray Gene Expression Data, Bioinformatics,2002,pp.39-45.