Minimal gene selection for classification and diagnosis prediction based on gene expression profile

Alireza Mehridehnavi[1,2], Lia Ziaei[1]

[1]Medical School, Medical Physics and Engineering, [1]Medical Image and Signal Processing Research Center,
[2]Isfahan University of Medical Sciences, Isfahan, Iran

Abstract

Background: Up to date different methods have been used in order to dimensions reduction, classification, clustering and prediction of cancers based on gene expression profiling. The aim of this study is extracting most significant genes and classifying of Diffuse Large B-cell Lymphoma (DLBCL) patients on the basis of their gene expression profiles.

Materials and Methods: We studied 40 DLBCL patients and 4026 genes. We utilized Artificial Neural Network (ANN) for classification of patients in two groups: Germinal center and Activated like. As we were faced with low number of patients (40) and numerous genes (4026), we tried to deploy one optimum network and achieve to minimum error. Moreover we used signal to noise (S/N) ratio as a main tool for dimension reduction. We tried to select suitable training data and so to train just one network instead of 26 networks. Finally, we extracted two most significant genes.

Result: In this study two most significant genes based on their S/N ratios were selected. After selection of suitable training samples, the training and testing error were 0 and 7% respectively.

Conclusion: We have shown that the use of two most significant genes based on their S/N ratios and selection of suitable training samples can lead to classify DLBCL patients with a rather good result. Actually with the aid of mentioned methods we could compensate lack of enough number of patients, improve accuracy of classifying and reduce complication of computations and so running time.

Key Words: Artificial neural network, classification, dimension reduction, diffuse large B-cell lymphoma, significant genes

Address for correspondence:
Dr. Alireza Mehridehnavi, Medical Physics and Engineering Department, Medical School, Isfahan University of Medical Sciences, Isfahan, Iran.
E-mail: mehri@med.mui.ac.ir

Received: 05.02.2012, **Accepted:** 12.08.2012

Access this article online

Quick Response Code:

Website:
www.advbiores.net

DOI:
10.4103/2277-9175.107999

INTRODUCTION

Tumors conventionally are diagnosed by morphological appearance on the base of their pathology and immunohistochemistry on the protein expression activities. The underlying genetic disorders are hidden form histological appearance of tumors.[1]

By the development of the microarray techniques the simultaneous monitoring of thousands of genes expression became an ordinary job in genetic behavior

How to cite this article: Mehridehnavi A, Ziaei L. Minimal gene selection for classification and diagnosis prediction based on gene expression profile. Adv Biomed Res 2013;2:26.

of tumors.[2] The gene expression data is very different from the data produced by all other previous methods. First, it has very high dimension usually contains thousands to tens of thousands of genes. Second, the number of samples in current data is a few due to shortage of candidates in this type of study. Third, relevant genes to cancer are narrow subsection of expressed gene spectrum. It is obvious that traditional existing classification methods were not designed to handle this kind of the data efficiently and effectively.[3]

As it has shown the gene selection should be performed prior to cancer classification and it could improve the accuracy. Feature selection helps to reduce data size and the running time.[4]

Artificial neural network is a robust tool as either clustering or classification. Supervised models are used for classification and unsupervised models are used for clustering.[5]

Diffuse Large B-cell Lymphoma (DLBCL), the most common subtype of non-Hodghin's Lymphoma, is clinically heterogeneous. Forty percent of patients have better overall survival time than the others. Alizadeh *et al.* showed that there is diversity in gene expression among tumors of DLBCL patients, apparently reflecting the variation in tumor proliferation rate, host response and differentiation state of the tumor.[6]

O'Neill *et al.* used two layers neural network for classification of DLBCL patients. Their classification accuracy was 100% and they were able to extract 34 significant genes. But they did not claim that the gene sets extracted in their procedure were the "best" gene sets.[7]

Lossos *et al.* studied 36 genes whose expression had been reported to predict survival in diffuse large B-cell lymphoma of 66 patients. They showed that measurement of the expression of 6 genes is sufficient to predict overall survival in diffuse large B-cell lymphoma.[8]

The above-mentioned study identified two molecularly distinct forms of DLBCL: Germinal center B-like DLBCL and Activated B-like DLBCL.[8]

In this study we were going to use data from these DLBCL patients to differentiate between two forms of DLBCL using supervised neural network. The goal was to find out minimum possible number of genes that the used model (ANN) would be able to classify a new expression pattern.

Moreover with regard to limited number of patients, we tried to obtain suitable training samples and so an optimum ANN.

MATERIALS AND METHODS

The data presented in first figure of earlier report http://llmpp.nih.gov/lymphoma/data. Shtml were used in this study. These data were from 40 labeled patients and corresponding 4026 genes expression levels. In Alizadeh *et al.,* study on this database and using clustering method, patients divided into two groups: Germinal center and activated like.[6] In this study, we used artificial neural network (ANN) to classify and predict data and the labels were selected according to Alizadeh *et al.,* results. The data analysis consisted of the following steps:

Reduction of data dimensions

As we dealt with huge number of genes and low number of patients, it should be better to reduce number of genes. Actually thanks to reduction of data size, we could improve accuracy of classification and so running time. In this study we used signal to noise (S/N) ratio as a main tool in order to reduce the dimensions whereas in previous work we used combination of S/N and PCA to perform this job.[9]

The S/N ratio is defined as follow:
$S/N= (\mu_A - \mu_B)/(\sigma_A + \sigma_B)$

μ and σ are mean and standard deviation per class, respectively.[10]

This ratio is just usable in two class problems.

In this study Genes were ranked based on their S/N ratios. Then various thresholds and therefore various number of ranked genes were tried.

Artificial neural network

In this study, we applied the Perceptron neural network on the presented data. If the selection of the training data is suitable, we will have minimum training error with Training just one network. As it has shown in previous works it is possible to classify the current data by the use of 14 higher eigenvectors and subsequently the use of 14 highest order components. The results have shown 93% accuracy in classification by the use of artificial neural network.[9]

As the number of classes was limited to two subtypes of classes, it seemed that this data type was classifiable by fewer number of features.[6]

In this study, at first, 14 germinal center and 12 activated like patients were chosen randomly as the training data (approximately ⅔ of total data). Therefore remained 14 patients (eight germinal center and six activated like patients) were selected as testing data. Then we tried to classify these training and testing data with the aid of Perceptron neural network. Because of low number of patients and to achieve more accurate result, we proposed one approach in order to choose suitable training and testing data as fallow:

- The total data was classified using Perceptron network.
- The distances from data to classifier were computed.
- The distances were ranked and 14 germinal center and 12 activated like samples with shortest distances were selected as the training set.

RESULTS

Different types and structures of neural network were tried on the data by varying the number of features. We found out that two of the highest rank genes are able to classify classes in some special data selections for training and testing set.

If we select samples that have been shown in Figure 1, as the training data, then network classification isn't optimal. It is noticeable, in Figure 2, the training error is zero but testing error is 57% (8 out of 14 samples were misclassified).

For solving this problem, we proposed the selection of suitable training data approaches [Figure 3]. Finally the test data was applied on the network and the test error was measured. With use two most significant genes and training the network, the training and testing errors were 0 and 7%, respectively (i.e., 1 out of 14 was misclassified). One testing error has been shown in Figure 4. If this misclassified sample put among training data, the training and testing errors will be 3.8 and 0%, respectively.

In the final experiment, the PCA was used and the various numbers of eigenvectors were examined. In one experiment we selected 10 most significant genes based on their S/N ratios and then we applied PCA to reduce these genes from 10 to 2 and then we utilized linear Perceptron neural network for classifying the patients. In another experiment, we used PCA calculated from total genes (instead of 10 genes) for reducing them from 4026 to 2. In the other hand, in the new dimension space with the use of two eigenvectors, every data is the linear combination of genes. Best result was obtained using PCAcalculated from 10 most significant genes and the error was 0%. The distribution of new two features that have been yielded using PCA calculated from total and ten most significant genes have been shown in Figures 5 and 6 respectively. As it seems in Figure 6, two groups are differentiated simpler.

DISCUSSION

In this study, best result was computed using training one network with two most significant genes based on their S/N ratios. These two genes have been shown in Table 1. Selection of training and testing samples performed on the basis of their distance to classifier line. Actually with the aid of mentioned methods we could compensate lack of enough number of patients, improve accuracy of classifying and reduce complication of computations and so running time. The combination of S/N ratio and PCA was suitable method for reduction of dimensions and a simple neural network was near perfect tool for this classification.

In previous work we used S/N ratio as well as PCA to reduce dimensions from 4026 to 14 or 10. Then in the classification step, we chose training samples (26 patients) randomly. Since there were not enough samples available, we performed a leave one out cross validation on 26 training samples. In this method 26 networks were trained. Under mentioned conditions, 10 eigenvectors and labeling of patients as Alizadeh *et al.,* study (Germinal center and activated like), the calculated accuracy was 100%. After removing PCA and using most significant genes based on S/N ratio, the accuracy of classification was 100% but in some permutations, the accuracy was 93%.[9]

In this study we used S/N ratio as a main tool to decrease number of genes from 4026 to 2. Because the one layer Perceptron network is a linear classifier and the features are two, the two class samples are differentiated from each other using a line and the weights of the network are related to the slope of this line. In the classification step, we performed selection of training samples on the basis of their distance to classifier line and applied one linear Perceptron ANN to two dimension data. The result was rather good.

With the use of only one gene, different class samples were not differentiated as good as using two most significant genes.

It should be notified that although the achieved accuracy has not improved in our second study (current paper) comparison with the first one (both of them are rather good), the complication of the computations and so the running time has decreased.

The first generation of gene expression analysis methods has been successfully applied in a variety of

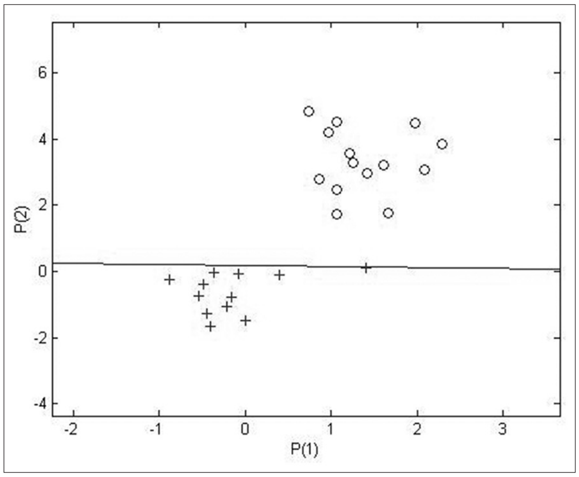

Figure 1: Random selection of train and classification

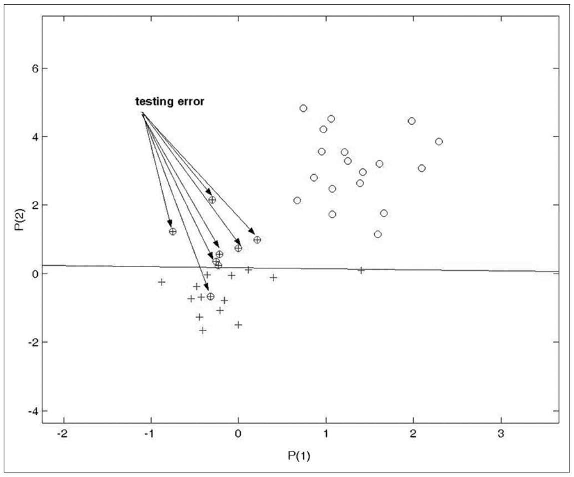

Figure 2: Random selection of training and the testing data samples that resulted to 8 number of misclassified testing sample

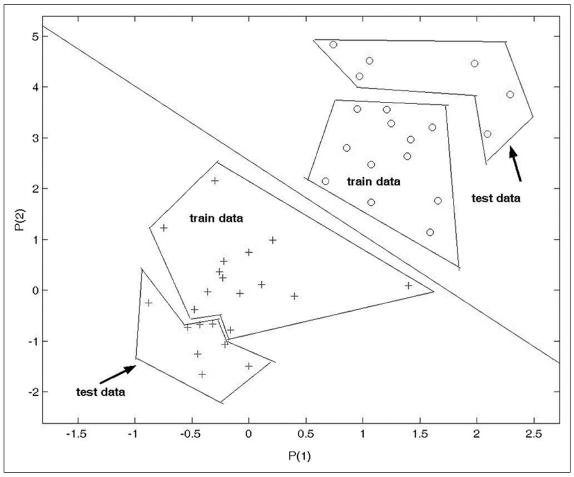

Figure 3: A typical Sample division for classified data based on the two most significant features

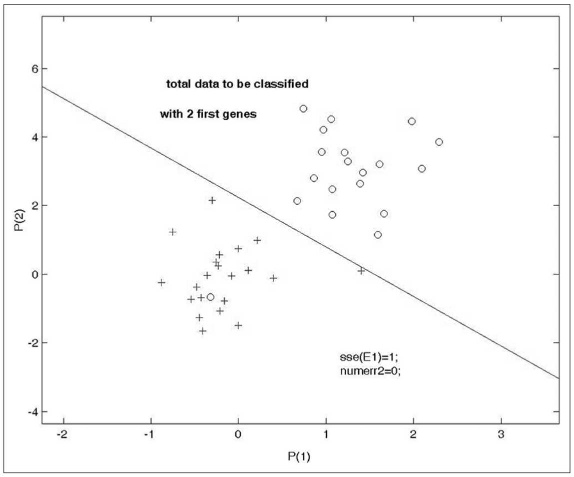

Figure 4: Classification of the whole data by neural based of two most significant features

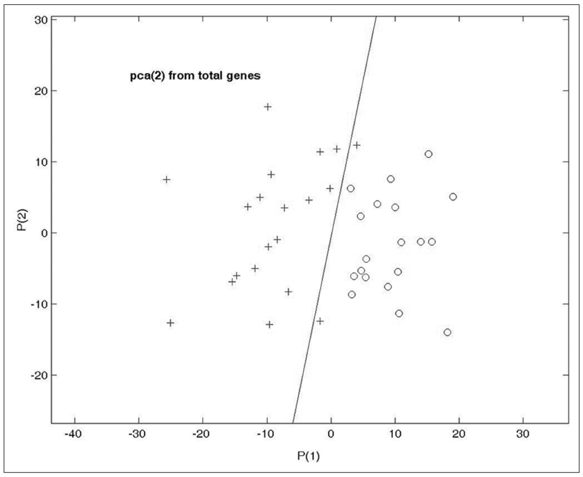

Figure 5: The distribution of whole sample space using PCA calculated from whole gene space

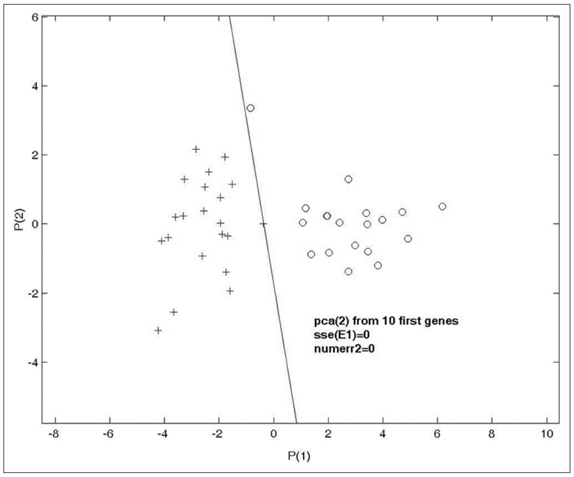

Figure 6: The distribution of two sample space using PCA of the ten most significant genes

clustering and classification settings. Alizadeh *et al.*, used hierarchical clustering for dividing patterns into two subgroups.[6]

The two genes that were extracted in our study [Table 1] have no overlap with the genes extracted in

Table 1: Two most significant genes extracted in this study

19289	(UnknownUG Hs.169565ESTs, Moderately similar to (H. sapiens); Clone=825217)
13394	(UnknownUG Hs.120716ESTs; Clone=1334260)

Table 2: Thirty four most significant genes extracted in O'Neill and song study[7]

14706	Hs.180836	18	17856	Interferon alfa/beta receptor_2
21367	Hs.134746	19	21653	Hs.1510936
13601	Similar to high mobility group	20	15656	Unknown
20397	FBPI=FUSE binding protein I	21	14393	Hs.29205
17901	*Pre_pro_orphanin	22	16631	Adenosine kinase
13097	Unknown	23	13318	Hs.122428
14560	Hs.32533	24	18330	Topoisomerase II beta
13867	Unknown	25	14983	Unknown
15664	Unknown	26	17721	IdI=inhibitor of DNA binding I
20490	Hs.122407	27	16850	PM5 protein=homology to collagenase
13650	Unknown	28	20481	Hs.37629
18252	Myosin_IC	29	17398	Receptor r_IBB ligand
16886	JAWI	30	14772	Unknown
18593	Receptor protein_ tyrosin kinase	31	19280	BENE
20759	Hs.33053	32	21603	Hs.33431
17802	Thymosin beta_4	33	19258	Tre_2
17887	A_raf=c_raf_I kinase	34	21091	Hs.199250

O'Neill and Song study [Table 2].[7]

CONCLUSION

We have shown that the use of two most significant genes based on their S/N ratios and selection of suitable training samples can lead to classify DLBCL patients with a rather good result.

In this work, it has been shown that in data types with huge number of features compared with number of samples, there is not a unique solution for problems such as microarray data classifications. Therefore, there should be more precaution in classification result announcement. As it has been shown, with the aid of simple structure networks (i.e., single layer perceptron) it is possible to classify this type of data. It is recommended to utilize simple classifiers at first and then to go towards more complicated methods. There should be some measures on degree of freedom on the models on more complicated methods.

ACKNOWLEDGMENT

The authors would like to thank Alizadeh *et al.,* for preparing Lymphoma databases. Authors are hereby acknowledge the financial and spiritual support of the Vice Chancellery of Research of the Isfahan University of Medical Sciences for this study.

REFERENCES

1. Khan J, Wei JS, Ringner M, Sall LH, Landanyi M, Wetermann F, et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat Med 2001;7:673-9.
2. Nguyen DV, Arpat AB, Wang N, Carroll RJ. DNA Microarray experiments: Biological and technological aspects. Biometrics 2002;58:701-17.
3. Lossos IS, Morgensztern D. Non-Hodgkin's lymphoma in the microarray era. Clin Lymphoma 2004;5:128-9.
4. Liu J, Iba H, Ishizuka M. Selecting informative genes with parallel genetic algorithms in tissue classification. Genome Inform 2001;12:14-23.
5. Mehridehnavi AR. Classification of different cancerous animal tissues on the basis of their 1H NMR spectra data using different types of artificial neural networks. Res Pharm Sci 2007;2:53-9.
6. Alizadeh A, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 2000;403:503-11.
7. O' Neill MC, Song LI. Neural Network analysis of lymphoma microarray data: Prognosis and diagnosis near perfect. BMC Bioinform 2003;4:13.
8. Lossos IS, Czerwinski DK, Alizadeh A, Wechser MA, Tibshirani R, Botstein D, et al. Prediction of survival in Diffuse Large B-Cell lymphma based on the expression of six genes. N Engl J Med 2004;350: 1828-37.
9. Ziaei L, Mehri AR, Salehi M. Application of artificial neural networks in classification and diagnosis prediction of a subtype of lymphoma based on gene expression profile. JRMS 2006;11:13-7.
10. Gloub TR, Slonim DK, Jamayo P, Gaasenbeek M, Huard C, Mesirov JP, et al. Molecular Classification of cancer: Class discovery and class prediction by gene expression monitoring Science 1999;286:531-7.

Source of Support: Nil, **Conflict of Interest:** None declared.