Biomedical Data Mining Using RBF Neural Networks

Feng Chu and Lipo Wang

School of Electrical and Electronic Engineering
Nanyang Technological University
Block S1, 50 Nanyang Avenue, Singapore 639798

E-mail: elpwang@ntu.edu.sg

URL: http://www.ntu.edu.sg/home/elpwang

INTRODUCTION

Accurate diagnosis of cancers is of great importance for doctors to choose a proper treatment. Furthermore, it also plays a key role in the searching for the pathology of cancers and drug discovery. Recently, this problem attracts great attention in the context of microarray technology. Here, we apply radial basis function (RBF) neural networks to this pattern recognition problem. Our experimental results in some well-known microarray data sets indicate that our method can obtain a very high accuracy with a small number of genes.

BACKGROUND

Microarray is also called gene chip or DNA chip. It is a newly appeared biotechnology that allows biomedical researchers monitor thousands of genes simultaneously (Schena *et al.*, 1995). Before the appearance of microarrays, a traditional molecular biology experiment usually works on only one gene or several genes, which makes it difficult to have a “whole picture” of an entire genome. With the help of microarrays, researchers are able to monitor, analyze and compare expression profiles of thousands of genes in one experiment.

On account of their features, microarrays have been used in various tasks such as gene discovery, disease diagnosis, and drug discovery. Since the end of the last century, cancer classification based on gene expression profiles has attracted great attention in both the biological and the engineering fields. Compared with traditional cancer diagnostic methods based mainly on the morphological appearances of tumors, the method using gene expression profiles is more objective, accurate, and reliable. More importantly, some types of cancers have subtypes with very similar appearances that are very hard to be classified by traditional methods. It has been proven that gene expression has a good capability to clarify this previously muddy problem.

Thus, to develop accurate and efficient classifiers based on gene expression becomes a problem of both theoretical and practical importance. Recent approaches on this problem include artificial neural networks (Khan *et al.*, 2001), support vector machines (Guyon *et al.*, 2002), k-nearest neighbor (Olshen and Jain 2002), nearest shrunken centroids (Tibshirani *et al.*, 2002), and so on.

A solution to this problem is to find out a group of important genes that contribute most to differentiate cancer subtypes. In the meantime, we should also provide proper algorithms that are able to make correct prediction based on the expression profiles of those genes. Such work will benefit early diagnosis of cancers. In addition, it will help doctors choose proper treatment. Furthermore, it also throws light on the relationship between the cancers and those important genes.

From the point of view of machine learning and statistical learning, cancer classification using gene expression profiles is a challenging problem. The reason lies in the following two points. First, typical gene expression data sets usually contain very few samples (from several to several tens for each type of cancers). In other words, the training data are scarce. Second, such data sets usually contain a large number of genes, for example, several thousands. That is, the data are high dimensional. Therefore, this is a special pattern recognition problem with relatively small number of patterns and very high dimensionality. To provide such a problem with a good solution, appropriate algorithms should be designed.

In fact, a number of different approaches such as k-nearest neighbor (Olshen and Jain 2002), support vector machines (Guyon *et al.*,2002), artificial neural networks (Khan *et al.*, 2001) and some statistical methods have been applied to this problem since 1995. Among these approaches, some obtained very good results. For example, Khan *et al.* (2001) classified small round blue cell tumors (SRBCTs) with 100% accuracy by using 96 genes. Tibshirani *et al.* (2002) successfully classified SRBCTs with 100% accuracy by using only 43 genes. They also classified 3 different subtypes of lymphoma with 100% accuracy by using 48 genes. (Tibshirani *et al.*, 2003)

However, there are still a lot of things can be done to improve present algorithms. In this work, we use and compare 2 gene selection schemes, i.e., principal components analysis (PCA) (Simon, 1999) and a t-test-based method (Tusher *et al.*, 2001). After that, we introduce an RBF neural network (Fu and Wang, 2003) as the classification algorithm.

MAIN THRUST OF THE CHAPTER

After a comparative study of gene selection methods, a detailed description of the RBF neural network and some experimental results are presented in this section.

MICROARRAY DATA SETS

We analyze 3 well-known gene expression data sets, i.e., the SRBCT data set (Khan *et al.*, 2001), the lymphoma data set (Alizadeh *et al.*, 2000), and the leukemia data set (Golub *et al.*, 1999).

The lymphoma data set (http://llmpp.nih.gov/lymphoma) (Alizadeh *et al.*, 2000) contains 4026 “well measured” clones belonging to 62 samples. These samples belong to following types of lymphoid malignancies: diffuse large B-cell lymphoma

(DLBCL, 42 samples), follicular lymphoma (FL, 9 samples) and chronicle lymphocytic leukaemia (CLL, 11 samples). In this data set, a small part of data is missing. A k-nearest neighbor algorithm was used to fill those missing values (Troyanskaya *et al.*, 2001).

The SRBCT data set (http://research.nhgri.nih.gov/microarray/Supplement/) (Khan *et al.*, 2001) contains the expression data of 2308 genes. There are totally 63 training samples and 25 testing samples. 5 of the testing samples are not SRBCTs. The 63 training samples contain 23 Ewing family of tumors (EWS), 20 rhabdomyosarcoma (RMS), 12 neuroblastoma (NB), and 8 Burkitt lymphomas (BL). And the 20 testing samples contain 6 EWS, 5 RMS, 6 NB, and 3 BL.

The leukemia data set (http://www-genome.wi.mit.edu/cgi-\\bin /cancer/publications) (Golub *et al.*, 1999) has two types of leukemia, i.e., acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL). Among these samples, 38 of them are for training; the other 34 blind samples are for testing. The entire leukemia data set contains the expression data of 7,129 genes. Different with the cDNA microarray data, the leukemia data are oligonucleotide microarray data. Because such expression data are raw data, we need to normalize them to reduce the systemic bias induced during experiments. We follow the normalization procedure used by Dudoit (2002). Three preprocessing steps were applied: (a) thresholding with floor of 100 and ceiling of 16000; (b) filtering, exclusion of genes with *max*/*min*<5 or (*max*-*min*)<500. *max* and *min* refer to the maximum and the minimum of the gene expression values, respectively; and (c) base 10 logarithmic transformation. There are 3571 genes survived after these three steps. After that, the data were standardized across experiments, i.e., minus the mean and divided by the standard deviation of each experiment.

METHODS FOR GENE SELECTION

As mentioned in the former part, the gene expression data are very high-dimensional. The dimension of input patterns is determined by the number of genes used. In a typical microarray experiment, usually several thousands of genes take part in. Therefore, the dimension of patterns is several thousands. However, only a small number of the genes contribute to correct classification; some others even act as “noise”. Gene selection can eliminate the influence of such “noise”. Furthermore, the fewer the genes used, the lower the computational burden to the classifier. Finally, once a smaller subset of genes is identified as relevant to a particular cancer, it helps biomedical researchers focus on these genes that contribute to the development of the cancer. The process of gene selection is ranking genes' discriminative ability first and then retaining the genes with high ranks.

As a critical step for classification, gene selection has been studied intensively in recent years. There are two main approaches, one is principal component analysis (PCA) (Simon, 1999), perhaps the most widely used method; the other is a t-test-based approach which has been more and more widely accepted. In the important papers (Alizadeh *et al.*, 2000; Khan *et al.*, 2001), PCA was used. The basic idea of PCA is to find the most “informative” genes that contain most of the information in the data set. Another approach is based on t-test that is able to measure the difference between two groups. Thomas *et al.* (2001) recommended this method. Tusher *et al.* (2001) and Pan (2002) also proposed their method based on t-test, respectively. Besides these two main methods, there are also some other methods. For example, a method called Markov blanket was proposed by Xing *et al.* (2001). Li *et al.* (2001) applied another method which combined genetic algorithm and K-nearest neighbor.

PCA (Simon, 1999) aims at reducing the input dimension by transforming the input space into a new space described by principal components (PCs). All the PCs are orthogonal and they are ordered according to the absolute value of their eigenvalues. The k-th PC is the vector with the k-th largest eigenvalue. By leaving out the vectors with small eigenvalues, the input space's dimension is reduced.

In fact, the PCs indicate the directions with largest variations of input vectors. Because PCA chooses vectors with largest eigenvalues, it covers directions with largest variations of vectors. In the directions determined by the vectors with small eigenvalues, the variations of vectors are very small. In a word, PCA intends to capture the most informative directions (Simon, 1999).

We tested PCA in the lymphoma data set (Alizadeh *et al.*, 2000). We obtained 62 PCs from the 4026 genes in the data set by using PCA. Then, we ranked those PCs according to their eigenvalues (absolute values). Finally, we used our RBF neural network that will be introduced in the latter part to classify the lymphoma data set.

At first, we randomly divided the 62 samples into 2 parts, 31 samples for training and the other 31 samples for testing. We then input the 62 PCs one by one to the RBF network according to their eigenvalue ranks starting with the PC ranked 1. That is, we first used only a single PC that is ranked 1 as the input to the RBF network. We trained the network with the training data and subsequently tested the network with the testing data. We repeated this process with the top two PCs, then the top three PCs, and so on. Figure 1 shows the testing error. From this result, we found that the RBF network can not reach 100% accuracy. The best testing accuracy is 93.55% that happened when 36 or 61 PCs were input to the classifier. The classification result using the t-test-based gene selection method will be shown in the next section, which is much better than PCA approach.

The t-test-based gene selection measures the difference of genes’ distribution using a t-test based scoring scheme, i.e., t-score (TS). After that, only the genes with the highest TSs are to be put into our classifier. The TS of gene *i* is defined as follows (Tusher *et al.*, 2001):

$$TS_i = \max\left\{\left|\frac{\bar{x}_{ik} - \bar{x}_i}{d_k s_i}\right|, k = 1,2,...K\right\}$$

$$\bar{x}_{ik} = \sum\nolimits_{j \in C_k} \bar{x}_{ij} / n_k$$

$$\bar{x}_i = \sum_{j=1}^{n} x_{ij} / n$$

where:

$$s_i^2 = \frac{1}{n-K}\sum_k \sum\nolimits_{j \in C_k} \left(x_{ij} - \bar{x}_{ik}\right)^2$$

$$d_k = \sqrt{1/n_k + 1/n}$$

There are K classes. $\max\{y_k, k = 1,2,..K\}$ is the maximum of all $y_k, k = 1,2,..K$. C_k refers to class k that includes n_k samples. x_{ij} is the expression value of gene i in sample j. $\bar{x}_{ik}$ is the mean expression value in class k for gene i . n is the total number of samples. $\bar{x}_i$ is the general mean expression value for gene

i. s_i is the pooled within-class standard deviation for gene i. Actually, the t-score used here is a t-statistics between a specific class and the overall centroid of all the classes.

To compare the t-test-based method with PCA, we also applied it to the lymphoma data set with the same procedure as what we did by using PCA. This method obtained 100% accuracy with only the top 6 genes. The results are shown in Figure 1. This comparison indicated that the t-test-based method was much better than PCA in this problem.

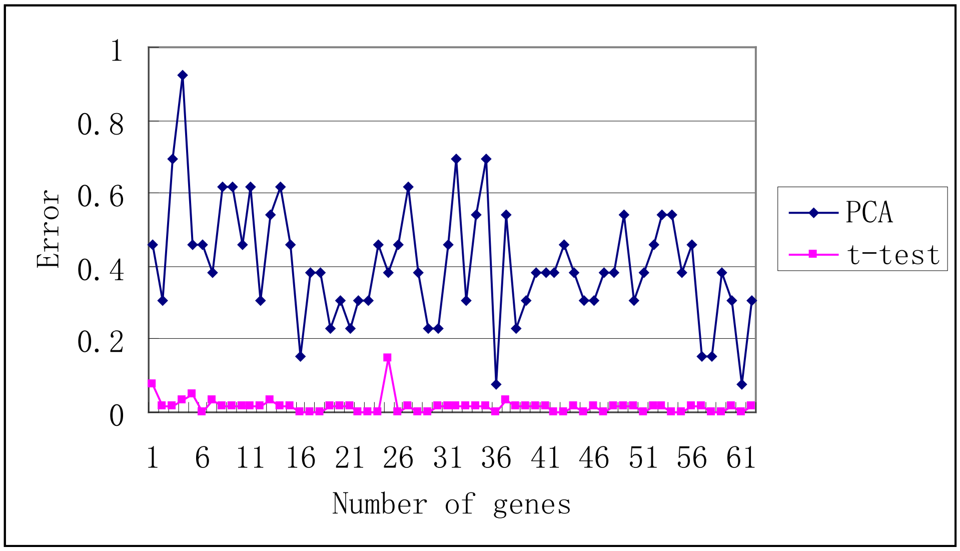

Fig.1 Classification results of using PCA and the t-test-based method as gene selection methods.

AN RBF NEURAL NETWORK

An RBF neural network (Haykin, 1999) has three layers. The first layer is an input layer; the second layer is a hidden layer that includes some radial basis functions, also known as hidden kernels; the third layer is an output layer. An RBF neural network can be regarded as a mapping of the input domain X onto the output domain Y. Mathematically, an RBF neural network can be described as follows:

$$y_m(x) = \sum_{i=1}^{N} w_{mi} G(\|x - t_i\|) + b_m \qquad , i=1,2,\ldots,N;\ m=1,2,\ldots M$$

Here $\|\cdot\|$ stands for the Euclidean norm. M is the number of outputs. N is the number of hidden kernels. $y_m(x)$ is the output m corresponding to the input x. t_i is the position of kernel i. w_{mi} is the weight between the kernel i and the output m. b_m is the bias on the output m. $G(\|x - t_i\|)$ is the kernel function. Usually, an RBF neural network uses Gaussian kernel functions as follows:

$$G(\|x - t_i\|) = \exp(-\frac{\|x - t_i\|^2}{2\sigma_i{}^2})$$

where σ_i is the radius of the kernel *i*.

The main steps to construct an RBF neural network include: (a) determining the positions of all the kernels (t_i); (b) determining the radius of each kernel (σ_i); and (c) calculating the weights between each kernel and each output node.

In this paper, we use a novel RBF neural network proposed by Fu and Wang (Fu and Wang, 2003), which allows for large overlaps of hidden kernels belonging to the same class.

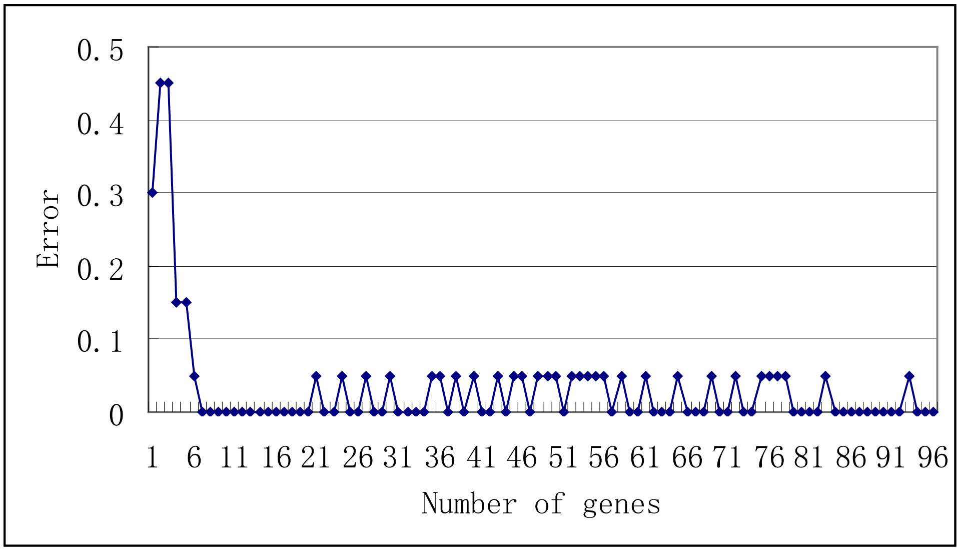

Fig. 2 The testing result in the SRBCT data set

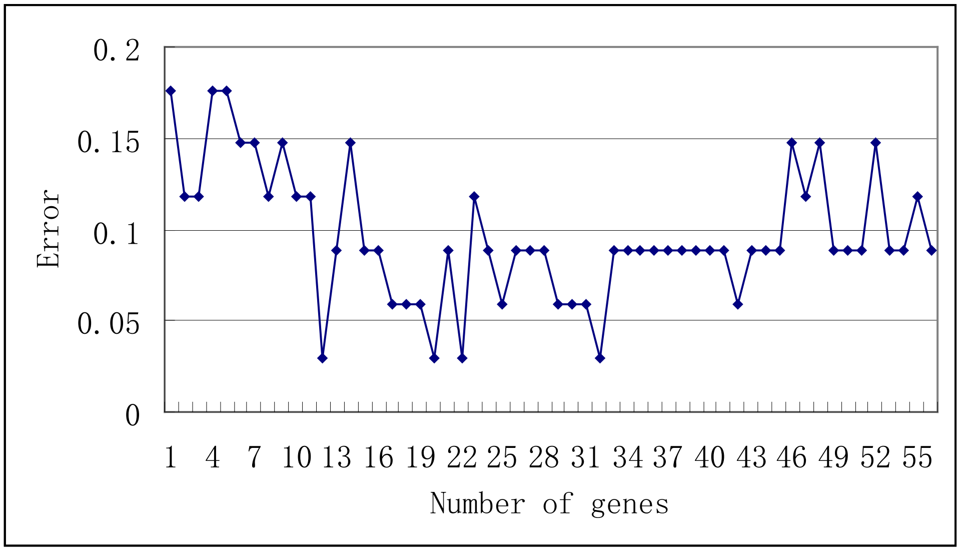

Fig.3. The testing result in the leukemia data set

RESULTS

In the SRBCT data set, we first ranked the entire 2308 genes according to their TSs (Tusher *et al.*, 2001). Then we picked out 96 genes with the highest TSs. We applied our RBF neural network to classify the SRBCT data set. The SRBCT data set contains 63 samples for training and 20 blind samples for testing. We input the selected 96 genes one by one to the RBF network according to their TS ranks starting with the gene ranked 1. We repeated this process with the top 2 genes, then the top 3 genes, and so on. Figure 2 shows the testing errors with respect to the number of genes. The testing error decreased to 0 when the top 7 genes were input into the RBF network.

In the leukemia data set, we chose 56 genes with the highest TSs (Tusher *et al.*, 2001). We followed the same procedure as in the SRBCT data set. We did classification with 1 gene, then 2 genes, then 3 genes and so on. Our RBF neural network got an accuracy of 97.06%, i.e. one error in all 34 samples, when 12, 20, 22, 32 genes were input, respectively.

FUTURE TRENDS

Until now, the focus of work is investigating the information with statistical importance in microarray data sets. In the near future, we will try to incorporate more biological knowledge into our algorithm, especially the correlations of genes.

In addition, with more and more microarray data sets produced in laboratories around the world, we will try to mine multi-data-set with our RBF neural network, i.e., we will try to process the combined data sets. Such an attempt will hopefully bring us a much broader and deeper insight into those data sets.

CONCLUSION

Through our experiments, we conclude that the t-test-based gene selection method is an appropriate feature selection/dimension reduction approach, which can find more important genes than PCA can.

The results in the SRBCT data set and the leukemia data set proved the effectiveness of our RBF neural network. In the SRBCT data set, it obtained 100% accuracy with only 7 genes. In the leukemia data set, it made only 1 error with 12, 20, 22, and 32 genes, respectively. In view of this, we also conclude that our RBF neural network outperforms almost all the previously published methods in terms of accuracy and the number of genes required.

REFERENCES

Alizadeh, A. A. *et al*. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403, 503-511.

Dudoit, S., Fridlyand, J. and Speed, J. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of American Statistics Association* 97, 77-87.

Fu, X., Wang, L. (2003). Data dimensionality reduction with application to simplifying RBF neural network structure and improving classification performance. *IEEE Trans. Syst., Man, Cybernetics. Part B: Cybernetics*. 33, 399-409.

Golub, T. R. *et al.* (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286, 531-537.

Guyon, I., Weston, J., Barnhill, S., Vapnik, V. (2002). Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning*, 46, 389-422.

Khan, J. M., *et al*. (2001). Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Medicine*, 7, 673-679.

Li, L., Weinberg, C. R., Darden, T. A., and Pedersen, L. G. (2001). Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method. *Bioinformatics*, 17, 1131-1142.

Olshen, A. B. and Jain, A. N. (2002). Deriving quantitative conclusions from microarray expression data. *Bioinformatics*, 18, 961-970.

Pan, W. (2002). A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics*, 18, 546-554.

Schena, M., Shalon, D., Davis, R.W. and Brown, P.O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270, 467-470.

Haykin, S. (1999). *Neural network, a comprehensive foundation*, 2nd Edition. Prentice-Hall, Inc. New Jersey, U.S.A.

Thomas, J. G., Olsen, J. M., Tapscott, S. J. and Zhao, L. P. (2001). An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles. *Genome Research*, 11, 1227-1236.

Tibshirani, R., Hastie T., Narashiman and Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl. Acad. Sci. USA*, 99, 6567-6572.

Tibshirani, R., Hastie, T., Narasimhan, B., Chu, G. (2003) Class predicition by nearest shrunken centroids with applications to DNA microarrays. *Statistical Science*, 18, 104-117.

Troyanskaya, O., Cantor, M, Sherlock, G. *et al*. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17, 520-525.

Tusher, V.G., Tibshirani, R. and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. USA*, 98, 5116-5121.

Xing, E. P., Jordan, M. I., and Karp, R. M. (2001). Feature selection for high-dimensional genomic microarray data. *Proceedings of the Eighteenth International Conference on Machine Learning.*

Terms and Definitions

Feature Extraction: Feature extraction is the process to obtain a group of features with the characters we need from the original data set. It usually uses a transform (e.g. principal component analysis) to obtain a group of features at one time of computation.

Feature Selection: Feature selection is the process to select some features we need from all the original features. It usually measures the character (e.g. t-test score) of each feature first, then, chooses some features we need.

Microarray: A Microarray is also called a gene chip or a DNA chip. It is a newly appeared biotechnology that allows biomedical researchers monitor thousands of genes simultaneously.

Gene Expression Profile: Through microarray chips, an image that describes to what extent genes are expressed can be obtained. It usually uses red to indicate the high expression level and uses green to indicate the low expression level. This image is also called a gene expression profile.

Principal Components Analysis: Principal components analysis transforms one vector space into a new space described by principal components (PCs). All the PCs are orthogonal to each other and they are ordered according to the absolute value of their eigenvalues. By leaving out the vectors with small eigenvalues, the dimension of the original vector space is reduced.

Radial Basis Function (RBF) Neural Network: An RBF neural network is a kind of artificial neural network. It usually has three layers, i.e., an input layer, a hidden layer, and an output layer. The hidden layer of an RBF neural network contains some radial basis functions, such as Gaussian functions or polynomial functions, to transform input vector space into a new non-linear space. An RBF neural network has the universal approximation ability, i.e., it can approximate any function to any accuracy, as long as there are enough hidden neurons.

T-Test: T-test is a kind of statistical method that measures how large the difference is between two groups of samples.

Training a neural network: Training a neural network means using some known data to build the structure and tune the parameters of this network. The goal of training is to make the network represent a mapping or a regression we need.

Testing a neural network: To know whether a trained neural network is the mapping or the regression we need, we test this network with some data that have not been used in the training process. This procedure is called testing a neural network.