Gene Expression Data Classification Using Artificial Neural Network Ensembles Based on Samples Filtering

Wutao Chen,Huijuan Lu,Mingyi Wang
College of Information Engineering
China Jiliang University
Hangzhou 310018, China
Email:hjlu@hz.cn

Cheng Fang
College of Civil Engineering and Architecture
Zhejiang University of Technology
Hangzhou 310023, China
Email:85534444@163.com

Abstract—**Bioinformatics analysis based on microarray technology is facing serious challenges, due to the extremely high dimensionality of the gene expression data comparing to the typical small number of available samples. Single artificial neural network was unstable and inaccurate for classification. In this paper we introduce classifying gene expression data using artificial neural network ensembles based on samples filtering. Simulation tests were carried out to verify the proposed strategy using Leukemia data sets, and the test results were compared with those of single artificial neural network, bagging artificial neural network ensembles and support vector machine. The results indicated that our method is more stable and more accurate.**

Keywords-classification; artificial neural network ensembles; samples filtering

I. INTRODUCTION

With the development of microarray technology, DNA microarrays have provided a great opportunity to measure the expression levels of thousands of genes simultaneously[1]. It provided a powerful tool for the research of gene function and was important to classification, diagnosis and case studies of diseases, such as cancers. In recent years, artificial neural network (ANN) has been widely used in gene expression data processing owing to its capability in identifying the complicated relationships in large data sets, and achieved a great success. However, the determination of network topology is a non-polynomial time problem. Due to the lack of theoretical methodology, in practical applications the performance of ANN often depends on the user's experience and could not achieve the desired result[2]. Hansen[3] proposed neural network ensembles methods in 1990. It improved the generalization ability of classifier significantly with small computational cost.

Diversity in neural network ensemble was the key to improve the generalization ability of classifier[4]. In fact this diversity was difficult to obtain. The wrongly labeled samples were relatively stable; in other words, a sample was likely to be wrongly labeled in other networks if it was wrongly labeled in one network. The generalization ability was improved while the wrongly labeled samples were separated from the training data set to construct one more network. Based on this idea, in this paper we introduce neural network ensembles using samples filtering. First, we performed feature selection with *t*-test. Then we trained *k* artificial neural networks with samples filtering algorithm by copying the wrongly labeled samples into a temporary data set. Third, we trained one more network with samples in the temporary data set. Finally, we predicted the labels of every sample based on the voting strategy.

II. MATERIALS AND METHODS

A. Gene expression data

Gene expression data[5] can be obtained through DNA microarray hybridization experiments. The format of data usually was a matrix after data was pre-processed. Each row in the matrix corresponds to one particular gene and each column to a sample. Observation of each sample can be seen as a vector $X_i(x_{i1}, x_{i2}, \ldots, x_{in}, y_i)$, in which $x_{ij}, j = 1,2,...,n$ is the expression level of gene j in sample X_i, and $y_i \in \{1,2,\ldots L\}$ is the class label of sample X_i. Gene expression classification is aimed at constructing a classifier and outputting a class label for each sample input.

B. Data sets

In order to make the results comparable, we adopted open data sets leukemia[6] for simulation tests. Leukemia data sets were divided into two types of samples: acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML). The training data set consists of 38 bone marrow samples (27 ALL and 11 AML), over 7129 probes from 6817 human genes. The other 34 samples were provided as testing data, including 20 ALL and 14 AML.

C. Feature selection

A typical gene expression dataset is a high-dimensional dataset with a small number of available samples. Generally, the gene number ranges from 2000 to 20000, far exceeding the sample size which is no more than several

978-0-7695-3816-7/09 $26.00 © 2009 IEEE
DOI 10.1109/AICI.2009.441

hundred. In fact, there is no need for such a multi-gene analysis in biology. First we need to perform feature selection before further processing that is, choosing genes which have expression levels of high diversity in different types of samples. Among the various feature selection methods, such as SNR[6], *t*-test[7], Fisher[7] and information gain[8], we chose *t*-test which has been proved to be an effective method in the binary-classification problem. In our case, it was given as

$$t(k) = \frac{\mu_k^{ALL} - \mu_k^{AML}}{\sqrt{\sigma_k^{ALL} / N_{ALL} + \sigma_k^{AML} / N_{AML}}}, \quad (1)$$

where $\mu_k^{ALL/AML}$ is the mean, $N_{ALL/AML}$ is the size and $\sigma_k^{ALL/AML}$ is the standard deviation of all samples with the class label $_{ALL/AML}$. We chose the first and the last $d/2$ $t(k), k = 1,2,...,n$ after *t(k)* was sorted in descending order.

D. Samples filtering algorithm

Due to the specificity of some samples on the training data set, prediction accuracy of the classifier cannot reach 100%. That is, the error cannot be a small enough value when training a single artificial neural network for all the samples. Therefore, the samples which were wrongly predicted at the training phase should be analyzed separately. First, we trained *k* artificial neural networks and filtered the wrongly labeled samples when training each artificial neural network. Second, we trained one more artificial neural network with those wrongly labeled samples. At last, we integrated these *k*+1 artificial neural networks into an ensemble classifier. The algorithm was described as follows:

Train ()
1 allocates memory for *WRONG_LABELED_SAMPLES* and *NETS*;
2 **for** *i*←1 to *k*
3 err_samples=train *net*(*i*) and return samples which were wrongly labeled;
4 **put** err_samples **into** *WRONG_LABELED_SAMPLES*;
5 **put** *net*(*i*) **into** *NETS*;
6 **end for**
7 train *net*(*k*+1) with *WRONG_LABELED_SAMPLES* and **put** it **into** *NETS*;

Here *WRONG_LABELED_SAMPLES* represents all the samples which were wrongly predicted during training an artificial network; *NETS* is the model of the ensemble classifier.

We defined the confidence level as follows:

$$Conf = \frac{MAX_1^L\{vote(j)\}}{k}, \quad (2)$$

where *vote(j)* is the number of votes that class *j* obtained, and *j*=1,2…,*L* is the set of class labels. Firstly, we predicted the class label based on the voting strategy with the former *k* neural networks. The prediction was adopted if *Conf* is larger than a threshold value; otherwise, the class label of the sample was predicted with the last network. The algorithm is described as follows:

Predict (sample, conf)
1 **for** the former *k* nets in ensemble classifier *NETS*
2 get the prediction class label of sample, vote for this class;
3 **end for**
4 **if** the maximum number of votes in one class is larger than *k*conf*
5 print the prediction class label;
6 **else** get the prediction class label through the last network;

III. EXPERIMENTS AND RESULTS

Simulations were carried out on the open data sets Leukemia. 2, 4, 8, 16, 32 genes were chosen respectively to classify genes. There were three layers in each network: input layer, hidden layer, and output layer. The Sigmoid function was adopted as the active function between the hidden layer and the output layer:

$$f(x) = \frac{2}{1+e^{-2x}} - 1 \quad (3).$$

In our tests, parameters of the classifier were set as follows: training epochs: *epochs*=100, permitted maximal error: *error*=0.01, learning rule: η=0.7, reduced parameter of learning rule: δ=0.9, momentum factor: *m*=0.3. The ensemble classifier was composed of 10 artificial neural networks and the confidence level *conf*=0.9. The numbers of cells in the hidden layer were all the same in every network, limited from 3 to 50. For each different number of cells in the hidden layer, the program ran 10 times for an average. The highest average accuracy rates are shown in Table 1.

TABLE 1. PERFORMANCE COMPARISON OF INDEPENDENT TESTS

d	BP	SVM	Bagging	Filtering
2	0.8235	0.8824	0.7824	0.8971
4	0.9706	0.9706	0.9706	0.9706
8	0.9118	0.8235	0.8589	0.9441
16	0.9706	0.9412	0.9676	0.9412
32	0.9706	0.9706	0.9706	0.9706

From Table 1, we could observe that the accuracies of the four classification methods rise with fluctuation with the increasing numbers of genes. When 32 genes were chosen, the filtering algorithm also reached the highest accuracy, the same as in the other methods. We found that the accuracy in the bagging algorithm was not improved significantly, which is similar to those reported in Dong[9]. The reason lies in that the data set itself was relatively small and 36.8% of the samples in the original training data set may never be selected[10]. We compared the stabilities and prediction

accuracies of BP, SVM, bagging and filtering algorithms using 8 genes. Results are shown in Fig.1.

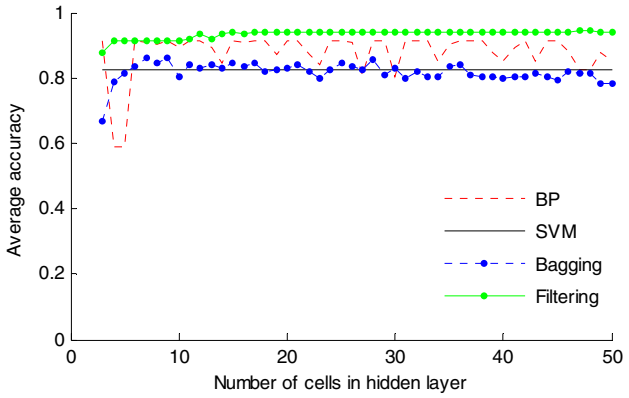

Figure 1. Comparison of accuracy rates and stability of four methods.

In Fig.1, regardless of the number of cells in the hidden layer, the filtering algorithm obtained better stability and prediction accuracy than BP and Bagging. Similar to the results shown in Table 1, the bagging algorithm did not obtain good prediction accuracy.

To further verify the effectiveness of the algorithm, we carried out 10-fold cross-validation on the training data set. The experimental results are shown in Table 2.

TABLE 2. PERFORMANCE COMPARISON OF 10-FOLD CROSS-VALIDATION TESTS

d	BP	SVM	Bagging	Filtering
2	0.8	0.7895	0.8	0.8
4	0.9333	0.9211	0.9333	0.9333
8	0.9333	0.9737	0.9667	0.9667
16	0.9333	0.9474	0.9333	0.9667
32	1.0	0.9474	1.0	1.0

It can be seen from Table 2 that ensemble artificial neural networks outperformed single BP artificial neural network, and the filtering algorithm we proposed achieved better results.

IV. CONCLUSIONS

In this paper, an artificial neural network ensemble algorithm based on samples filtering is proposed. Different from the idea of re-sampling, we focus on the wrongly predicted samples. Some issues are considered as follows: a single neural network is unstable; accuracy of the classifier is not high when training data set contains noise; the accuracy of bagging algorithm is difficult to be improved under a small number of samples. Simulations have been carried out and some results are shown as follows: integrated classifier can improve the generalization ability of prediction; the number of genes has significant impact on results; comparison of BP, SVM, bagging and filtering shows that our method has better stability and higher prediction accuracy.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China (No.60842009, No.10602055).

REFERENCES

[1] Lander, E.S., “Array of hope”, Nature Genetics, 1999, vol.21(supp l), pp.3-4.

[2] Dudoit, S., Fridlyand, J., Speed, T.P., “Comparison of discrimination methods for the classification of tumors using gene expression data”, Journal of the American Statistical Association, 2002,vol. 97, pp.77-87.

[3] Hansen, L.K., Salamon P., “Neural Network Ensembles”, IEEE Transactions on Pattern Analysis and machine Intelligence, 1990, vol.12, pp.993-1001.

[4] Ulf, J., Tuve, L., Lars, N. “The Importance of Diversity in Neural Network Ensembles-An Empirical Investigation”, Proceedings of International Joint Conference on Neural Networks. 2007, Orlando, Florida, USA.pp.12-17.

[5] Zheng, C.H., Huang, Kong, D.S., Zhen, X., Zhao, X.M., “Gene Expression Data Classification Using Consensus Independent Component Analysis”, Genomics, Proteomics & Bioinformatics, 2008,vol. 6, pp.74-78.

[6] Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield,C.D., Lander, E.S., “Class Discovery and Class Prediction by Gene Expression Monitoring”, Science, 1999, vol. 286, pp.531-537.

[7] Model, F., Adorján, P., Olek, A., Piepenbrock, C., “Feature Selection for DNA Methylation Based Cancer Classification”, Bioinformatics, 2001,vol. 17, pp.157-164.

[8] Wang, Y.H., Makedon, F.S., James C.F., Pearlman, J., “HykGene: a hybrid approach for selecting marker genes for phenotype classification using microarray gene expression data”, Bioinformatics, 2005,vol. 21, pp.1530-1537.

[9] Dong, Y.S., Han, K.S. “A comparison of several ensemble methods for text categorization”, 2004 IEEE International Conference on Service Computing. 2004, pp.419-422.

[10] Brieman, L., “Bagging Predictors. Machine Learning”, 1996, vol. 24, pp.123-140.