Simple Rule-based Ensemble Classifiers for Cancer DNA Microarray Data Classification

Hualong Yu[1*]
[1*]School of Computer Science and Engineering
Jiangsu University of Science and Technology
Zhenjiang, People's Republic of China
yuhualong@just.edu.cn

Sen Xu[2]
[2]School of Information Engineering
Yancheng Institute of Technology
Yancheng, People's Republic of China
xusen@ycit.cn

***Abstract*—DNA microarray, which is one of the most important molecular biology technologies in post-genomic era, has been widely applied in medical field, especially for cancer classification. However, it is difficult to acquire excellent classification accuracy by using traditional classification approaches due to microarray datasets are extremely asymmetric in dimensionality. In recent years, ensemble classifiers which may obtain better classification accuracy and robustness have attracted more interests in this field but it is more time-consuming. Therefore, this paper proposed a novel ensemble classification method named as SREC(Simple Rule-based Ensemble Classifiers). Firstly, the classification contribution of each gene is evaluated by a novel strategy and the corresponding classification rule is extracted. Then we rank all genes to select some important ones. At last, the rules of the selected genes are assembled by weighted-voting to make decision for testing samples. It has been demonstrated the proposed method may improve classification accuracy with lower time-complexity than traditional classification methods.**

Keywords-DNA microarray; ensemble classifiers; cancer; feature selection

I. INTRODUCTION

The advent of DNA microarray has provided the ability to measure the expression levels of thousands of genes simultaneously in a single experiment and made it possible to provide diagnosis for disease, especially for cancer, at molecular level [1-2]. To date, it has been one hot topic in bioinformatics and attracted more and more researchers from different fields, such as biology, medical, computer science and even statistics etc. They have proposed lots of methods and tools to analyze microarray data according to their domain knowledge. The approaches and tools mainly focus on: 1.extracting feature genes of a particular disease to help doctors to improve clinical diagnostic accuracy [1,3]; 2. clustering microarray data to find new subtypes of a particular disease so that improving effect of clinical treatment [2]; 3.constructing classification model to making accurate diagnosis for diseases [4-8].

It is noteworthy that many conventional classification approaches, such as support vector machine (SVM) [4], multilayer perception [5], K nearest neighbors classifier [6], C4.5 decision tree [7], linear bayesian classifier [8] etc., have been used to classify cancer microarray data. However, most of them produce poor recognition rate and robustness owing to the characteristic of high-dimensional and small-sample for microarray datasets. Therefore, in recent several years, more and more researchers paid their attention on ensemble classifiers for the purpose of acquiring more accurate and robust performance [9-12]. However, it is generally more time-consuming for ensemble classifiers, especially for those selective ensemble methods.

In this paper, we proposed a novel ensemble classification method named as SREC (Simple Rule-based Ensemble Classifiers). Firstly, we evaluate the classification contribution for each gene by a novel strategy which could guarantee to obtain classification rule of each gene synchronously. Then, we extract some excellent feature genes, i.e. the corresponding classification rules to construct classification committee. At last, the classification committee will make decision for testing samples according to the results of weighted-voting. The experimental results on two benchmark cancer microarray datasets have demonstrated the proposed method may acquire better classification accuracy with fewer base classifiers than previous works. Meanwhile, the proposed method holds lower computational complexity, even lower than some complex single classifiers.

The remainder of this paper is organized as follows. The Section II describes the methods in detailed. Experimental results and discussions are presented in Section III. At last, we address conclusions in Section IV.

II. METHODS

A. Feature Gene Selection(Simple Rule Extraction)

Different from other data, DNA microarray data are often extremely asymmetric in dimensionality, such as thousands or even tens of thousands of genes and a few hundreds of samples or less. Such extreme asymmetry between the dimensionality of genes and samples can lead inaccurate diagnosis of disease in clinic. Therefore, it has been shown that selecting a small set of marker genes can lead to improved classification accuracy[1-3].

In recent years, various feature gene selection methods have been proposed. Most of them have been approved helpful for improving predictive accuracy of disease and providing useful information for medical experts. All of these feature gene selection methods may be grouped into two teams: filter, which is also called gene ranking approach and wrapper,

978-1-4244-9763-8/11/$26.00 ©2011 IEEE

which is also entitled as gene subset selection approach [13]. In filter approach, each gene is evaluated individually and assigned a score reflecting its correlation with the class according to certain criteria. Genes are then ranked by their scores and some top-ranked ones are selected. The most famous filter feature gene selection method is SNR (signal-noise ratio) [2]. In the wrapper approach, a search is conducted in the space of genes, evaluating the goodness of each found gene subset by the estimation of the accuracy percentage of the specific classifier to be used, training the classifier only with the extracted genes. Compared with filter approach, wrapper approach generally obtains one gene subset with better classification performance but much more computational cost.

In this paper, we proposed a novel filter feature gene selection method which evaluates the score and synchronously obtains classification rule for each gene.

Firstly, all genes in training dataset are normalized in the range of [0, 1] using formula as below:

$$g_{ij}' = \frac{g_{ij} - g_{i\min}}{g_{i\max} - g_{i\min}} \tag{1}$$

where g_{ij} and g_{ij}' are the original and normalized expression value for the *i*th gene of the *j*th sample, respectively. While $g_{i\min}$ and $g_{i\max}$ are minimal and maximal value in all samples for the *i*th gene, respectively.

After normalization, the score of each gene will be evaluated according to its contribution to classification and meanwhile the corresponding classification rule will be acquired. For a specific gene, we scan it from 0 to 1 to find all adjacent different classes sample-pairs, i.e., there are no any other samples between the two samples of the same sample-pair and meanwhile the two samples must come from different classes. For example, for a specific gene, supposing there are six samples in training dataset, three are from class 1, the other three belong to class 2, the expression values in class 1 are 0.1, 0.3 and 0.6, while in class 2, the expression values of three samples are 0.4, 0.5 and 0.9, respectively. We can find three sample-pairs, i.e., {0.3,0.4}, {0.5,0.6} and {0.6,0.9}.

Then we use the mean value of each sample-pair as the classification boundary for two classes to computing the number of misclassified samples in training dataset. The least number of misclassified samples is used as the score of the gene and the corresponding mean value may be regarded as the classification rule. We still describe it with the example above, the score is 1 and the classification boundary is 0.35.

At last, all genes are ranked in ascending order based on their scores (the smaller the score is, the more important the gene is) and some top-ranked ones are extracted as feature genes.

B. *Construction of Ensemble Classifiers*

Generally, ensemble classifiers employ "majority voting" as the strategy to select the class most favors by the base classifiers as the label of a testing sample. Majority voting has some advantages in that it does not require any previous knowledge nor does it require any complex computation to decide. However, a poor classifier can affect the result of the ensemble in majority voting because it gives the same weight to all classifiers. Meanwhile, it is possible to arising equal number of members voting for several different classes. Therefore, we adopt weighted-voting in our study. Weighted-voting gives each base classifier different weights and the weights of the classifiers are determined by the accuracy on the training dataset. Owing to each base classifier is constructed on only one gene in this study, we defined the weight of each base classifier as 1/*Score*, where *Score* is the estimated score acquired in Section II. When making decision for the class label of a testing sample, weighted-voting is defined as below:

$$C_{ensemble} = \arg\max_{1 \le i \le 2} \{ \sum_{j=1}^{n} W_j S_i(C_j) \} \tag{2}$$

where *n* is number of classifiers; W_j is the weight of the *j*th classifier and $S_i(C_j)$ is 1 if the output of the *j*th classifier C_j equals class *i* otherwise 0.

Furthermore, it is noteworthy that we don't need to train base classifiers by traditional classification methods due to some simple classification rules have been extracted in the process of feature selection, which guarantee low computation -al complexity of the proposed method.

The detailed description of SREC method is listed as follows:

```
Algorithm: SREC
Input: training dataset $S_{train}$; testing dataset $S_{test}$; number of feature genes $n$
Output: $R$ which is the classification accuracy of testing dataset
Training Process:
1. Data normalization of training dataset $S_{train}$ by formula (1);
2. Scan each gene to getting all sample-pairs;
3. Compute score of each gene and extract the corresponding classification rules;
4. Rank all genes in ascending order based on their scores and select $n$ feature genes;
5. Compute weight for each selected feature gene by their scores and then use them to construct classification committee.
Testing Process:
1. Data normalization of testing dataset $S_{test}$ using formula (1) and the information provided by $S_{train}$;
2. Extract $n$ feature genes matching with training dataset;
3. Put each sample in $S_{test}$ into classification committee and compute its class by extracted classification rules and formula (2);
4. Output $R$ which is the classification accuracy of testing dataset.
```

III. RESULUTS & DISCUSSION

A. *Dataset & Experimental Settings*

In this paper, we use two benchmark cancer microarray datasets to testing the performance of SREC: Acute Leukemia dataset [2] and Colon Cancer dataset [1], respectively. The detailed information for both datasets is listed in Table I.

Acute Leukemia dataset includes 72 samples, where 47 samples are ALL (acute lymphoblastic leukemia) and another

25 samples belong to AML (acute myeloid leukemia). Each sample has 7129 genes [2].

Colon Cancer dataset contains 62 samples collected from colon cancer patients. Among them, 40 tumor biopsies are from tumors (labeled as "negative") and 22 normal (labeled as "positive") biopsies are from healthy parts of the colons of the same patients. Two thousand out of around 6500 genes were selected based on the confidence in the measured expression levels. More detail information about Colon dataset can be found in reference [1].

Raw data for both datasets may be downloaded from http://datam.i2r.a-star.edu.sg/datasets/krbd/.

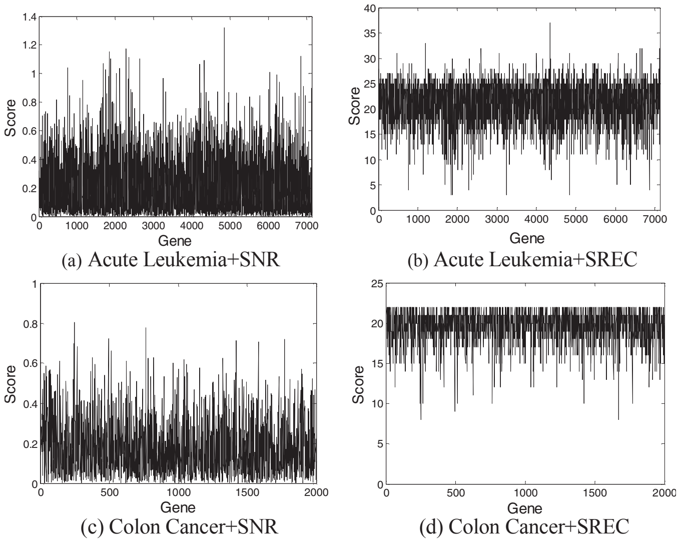
TABLE I. DATASETS USED IN THIS PAPER

Dataset	Number of		
	Samples	Genes	classes
Acute Leukemia	72	7129	2
Colon Cancer	62	2000	2

We ran experiments on a personal computer (Intel Pentium processor/ dual core 2.26 GHz/4G RAM) and all codes are written with Matlab 7.0. Meanwhile, in order to confirm the effectiveness of the proposed method, we compared it with SNR feature selection strategy+SVM classifier, where the SVM toolbox was used, the parameter of RBF kernel function σ and the penalty factor *C* are set to be 5 and 100, respectively.

B. *Extracted Feature Genes*

To explore the effectiveness of the feature gene selection strategy of SREC and its difference with SNR, we described their scores distribution in two datasets, respectively, as shown in Figure 1.



(a) Acute Leukemia+SNR (b) Acute Leukemia+SREC

(c) Colon Cancer+SNR (d) Colon Cancer+SREC

Figure 1. Scores Distribution of SNR & SREC in Two Datasets

It is noteworthy that in SNR, the important genes correspond to high scores, while in SREC, the genes are more inportant when their scores are low. From Figure 1, it is not difficult to find a fact that no matter SNR or SREC, there are only a few important genes for classification.

We also extracted top-10 genes of SNR and SREC in the two datasets, respectively. They are listed as below:

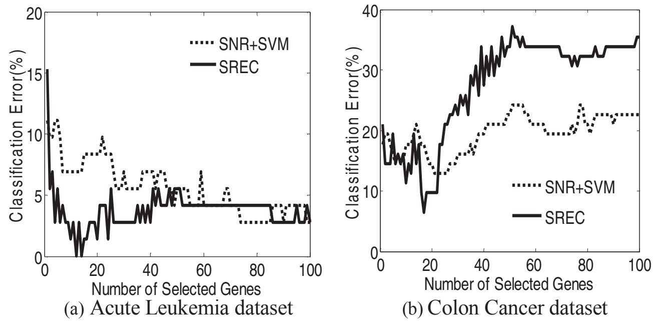
TABLE II. TOP-10 GENES OF SNR AND SREC IN THE TWO DATASETS

	Acute Leukemia Dataset
SNR	**4847**,**2288**,**1834**,**6855**,**2354**,**1882**,2642,4328,**1685**,4196
SREC	**1834**,**1882**,3252,**4847**,760,**2288**,**6855**,**1685**,**2354**,6041
	Colon Cancer Dataset
SNR	**249**,**493**,**765**,**1423**,1772,1582,**245**,377,**267**,822
SREC	**249**,1671,**493**,**245**,**267**,513,**765**,**1423**,1771,625

From Table II, it has been shown SREC may find many same feature genes as SNR, 7 and 6 top-10 genes are same in Acute Leukemia dataset and Colon Cancer dataset, respectively, which indicates that the feature selection strategy of SREC is effective and feasible. We consider these genes are closely related with cancer and expect these finds may provide useful information for medical experts.

C. *Evaluation of Classification Performance*

In addition, we investigated the correlation between number of selected feature genes (number of base classifiers) and the performance of SREC (classification accuracy). In order to confirm its effectiveness, SNR+SVM is used for comparison. The number of selected feature genes (number of base classifiers for SREC) varies from 1 to 100 and LOOCV (Leave-One-Out Cross Validation) classification accuracy is applied to evaluate performance. The relationship between the number of feature genes and classification error is shown in Figure 2.



(a) Acute Leukemia dataset (b) Colon Cancer dataset

Figure 2. Relationship Between Number of Selected Genes (Number of Base Classifiers) and Classification Error for SREC and SNR+SVM in Two Microarray Datasets

From Figure 2, we observe:

a) The proposed method is effective because it has produced less classification errors than traditional SNR+SVM classification method (0% : 2.8% in Acute Leukemia dataset and 6.5% : 12.9% in Colon Cancer dataset).

b) SREC may obtain excellent classification accuracy with only a few base classifiers (10-20), which means the proposed method need less storage space than other classification approaches.

c) No matter little or excessive base classifiers are put into classification committee of SREC, it will not obtain excellent performance. Little means deficiency of some important classification information and excess signifies addition of some noise.

We also compared the performance of SREC with that of many other ensemble classifiers. They are listed in Table III.

TABLE III. COMPARISION OF VARIOUS CLASSIFIERS

Method	Classification Accuracy (Number of Base Classifiers)	
	Acute Leukemia Dataset	Colon Cancer Dataset
SREC	100%(12)	93.5%(17)
SNR+SVM	97.2%(73)	87.1%(22)
BagBoosting[9]	95.9%(100)	83.9%(100)
enSVM[10]	97.2%(25)	88.7%(25)
EA+Bagging[11]	None	88.9%(42)
FS[12]	None	91.9%(100)

From Table III, we find that the SREC is better than many previous works due to it not only produced higher classification accuracy, but also saved much storage space, which confirms its effectiveness and feasibility again.

D. Computaional Complexity

Computational complexity is one of important aspects to assess an algorithm. Therefore, we tested the time complexity of SREC and SNR+SVM, respectively. Number of feature genes is assigned as 20 and the LOOCV is still selected as classification performance estimation method. The running time of SREC and SNR+SVM in both datasets is shown in Table IV.

TABLE IV. RUNNING TIME OF SREC AND SNR+SVM IN BOTH DATASETS

Method	Running Time (second)	
	Acute Leukemia Dataset	Colon Cancer Dataset
SNR+SVM	218.86	31.28
SREC	159.25	22.17

Table IV presents that the time complexity of SREC is lower than that of SNR+SVM, 159.25s : 218.86s in Acute Leukemia Dataset and 22.17s : 31.28s in Colon Cancer dataset. We believe the reason is classification rules are extracted together with feature selection for SREC, but SNR+SVM method needs to train classifier after generation of the feature gene subset.

IV. CONCLUSIONS

In this paper, we proposed a novel ensemble classification method named as simple rule-based ensemble classifiers (SREC) and used it for cancer microarray data classification. Experimental results indicate the proposed method is effective and feasible because it has produced less classification errors than many other classifiers. Meanwhile, it has some other advantages such as low time-complexity and storage space etc. However, the proposed method can only be used for binary-class problem, but not be appropriate for multiclass datasets. Our future work is to improve this method for extending its applications.

ACKNOWLEDGMENT

This paper is sponsored by Ph.D Foundation of Jiangsu University of Science and Technology under Grant No. 35301002. We also wish to express our acknowledgment to Professor Jing Zhao in Harbin Engineering University for her invaluable suggestion.

REFERENCES

[1] U. Alon, N.Barkai, D.A. Notterman, K. Gish, S. Ybarra, D. Mark, and A.J. Levine, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by Oligonucleotide array", Proceedings of the National Academy of Sciences, vol.96, no.12, pp.6745–6750,1999.

[2] T.R. Golub, D.K. Slonim, and P.Tamayo, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring", Science, vol.286, no.5439, pp.531-537,1999.

[3] H.L. Yu, G.C. Gu, H.B. Liu, J. Shen, and J. Zhao, "A modified ant colony optimization algorithm for tumor marker gene selection", Genomis, Proteomics & Bioinformatics, vol.7, no.4, pp.200-208, 2009.

[4] T.S. Furey, N. Cristianini, and N. Duffy, "Support vector machine classification and validation of cancer tissue samples using microarray expression data" Bioinformatics, vol.16, no.10, pp.906-914, 2000.

[5] K.J. Kim, and S.B. Cho, "Predication of colon cancer using an evolutionary neural network", Neurocomputing, vol.61, pp.361-379, 2004.

[6] L. Li, C.R. Weinberg, and T.A. Darden, "Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method", Bioinformatics, vol.17, no.12, pp.1131-1142, 2001.

[7] J.T. Horng, L.C. Wu, and B.J. Liu, "An expert system to classify microarray gene expression data using gene selection by decision tree", Expert System with Applications, vol.36, no.5, pp.9072-9081, 2009.

[8] M.H. Asyali, "Gene expression profile class predication using linear Bayesian classifiers", Computers in Biology and Medicine, vol.37, no.12, pp.1690-1699, 2007.

[9] M. Dettling, "Bagboosting for tumor classification with gene expression data", Bioinformatics, vol.20, no.18, pp.3583-3593, 2004.

[10] Y.H. Peng, "A novel ensemble machine learning for robust microarray data classification", Computers in Biology and Medicine, vol.36, no.6, pp.553-573, 2006.

[11] K.J. Kim, and S.B. Cho, "An Evolutionary Algorithm Approach to Optimal Ensemble Classifiers for DNA Microarray Data Analysis", IEEE Trans on Evolutionary Computation, vol.12, no.3, pp.377-388, 2008.

[12] H.L. Yu, G.C. Gu, H.B. Liu, and J. Shen, "Feature Subspace Ensemble Classifiers for Microarray Data", ICIC Express Letters, vol.4, no.1, pp.143-148, 2010.

[13] I. Inza, P. Larranaga, and R. Blanco, "Filter versus wrapper gene selection approaches in DNA microarray domains", Artificial Intelligence in Medicine, vol.31, no.2, pp.91-103, 2004.