Journal of Clinical & Medical Genomics

Song et al., J Clin Med Genomics 2016, 3:2
DOI: 10.4172/2472-128X.1000134

Research Article Open Access

Design and Analysis of Ensemble Classifier for Gene Expression Data of Cancer

Nianfeng Song, Kun Wang, Menglu Xu, Xiaolu Xie, Gan Chen and Ying Wang*

Department of Automation, Xiamen University, Xiamen, Fujian, 361005, China

Abstract

Gene expression levels are important for disease, such as, Cancer diagnosis. This paper proposed a SVM-based ensemble classifier to classify the control and cancer groups based on gene expression levels from microarray data. A combinational Recursive Feature Elimination in conjunction with the Adaboost algorithm was developed to select significant features and design the proper classifier. The method is applied to microarray data of cancer patients, and the results show improvements on the success rate. By AUC calculation, the SVM-based ensemble classifier shows predominate performance. Furthermore, the characteristics and different effect issues to classification performance is discussed. If a single SVM can obtain satisfactory classification performance, an ensemble SVM is hardly capable to improve it. Otherwise, an ensemble of SVM is superior to the best single SVM. We also investigated the effect of kernel functions, feature selections and type of classifiers on the classification.

Keywords: SVM; Ensemble methods; ROC; Microarray; Gene expression

Introduction

Gene expression patterns are characteristic for disease diagnosis. By now, many classification or prediction methods have been developed in machine learning community and many of them have been applied to cancer classification [1-3] based on gene expression levels from microarray data. But a great challenge would be raised by using the traditional learning algorithms, because the high dimensionality of microarray data may bring some disadvantages, such as over-fitting, poor performance and low efficiency. To alleviate this so-called 'high-dimensional small-sample' problem, several comprehensively comparative and improved methods have been proposed recently [4-6]. Feature selection [7-9], ensemble decision trees [10-12] and ensemble neural network [13-15] seem to be effective and sound solutions. Although a lot of researchers have done a lot of explorations on cancer classification, few people have focused on the combinational ensemble method with support vector machine to this problem or researched how the features affect the performance of the classifiers.

In this paper, we attempt to introduce a combinational Recursive Feature Elimination [16] (RFE) in conjunction with the Adaboost algorithm [17] used the Support Vector Machine (SVM) [18,19] as its learning algorithm to remarkably improve the accuracy and robustness of sample classification. Combining feature selection with the classifier can avail of more information of samples and eliminate the noisy features for classification. By using ensemble support vector machine, we can more effectively combine these features and improve the stability and robustness of answers. What's more, we also explore how do the different kinds of the feature selection methods affect the performance of the classifiers and how many features need to be selected to get the best performance of the classifiers. Finally, our method is compared with three different ensemble methods based on decision trees.

Materials and Method

Experimental process

In this paper, the experimental process is shown in the Figure 1. When we obtain the gene expression data, such as microarray data, from normal and cancer patients, preprocess and normalization should be implemented before further analysis. After that, features were

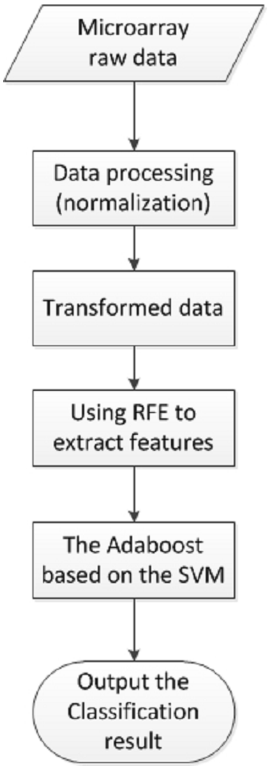
Microarray raw data
Data processing (normalization)
Transformed data
Using RFE to extract features
The Adaboost based on the SVM
Output the Classification result

Figure 1: The experimental process.

***Corresponding author:** Ying Wang, Ph.D., Department of Automation, Xiamen University, Xiamen, Fujian, 361005, China, Tel: 86-18959200966; E-mail: wangying@xmu.edu.cn

Received December 13, 2015; **Accepted** December 24, 2015; **Published** December 31, 2015

Citation: Song N, Wang K, Xu M, Xie X, Chen G, et al. (2015) Design and Analysis of Ensemble Classifier for Gene Expression Data of Cancer. J Clin Med Genomics 3: 134. doi:10.4172/2472-128X.1000134

selected with RFE algorithm. Based on selected features, an ensemble classifier with SVM as its learning algorithm is trained and constructed. Finally, through the competitive ensemble method, the robustness is improved greatly. Here, we simply use majority voting to combine the results in the Adaboost algorithm. All the implementations of the framework were implemented in MATLAB.

Datasets description

In this study, two microarry datasets of gene expressions from different groups were adopted. These two datasets have different characteristics (one can be linearly separated and the other one can't). The first data set was from cancer patients with two variants of leukemia (acute myeloid leukemia-AML and acute lymphoblastic leukemia-ALL) [20]. The data has two subsets: the training set is used to select genes and adjust the weights of the classifiers, and an independent test set is used to estimate the performance of the Classifier. The training set consists of 38 bone marrow samples (27 ALL and 11 AML), and the test set has 34 samples with 20 ALL and 14 AML. All samples have 7129 features, corresponding to some normalized gene expression value extracted from the micro-array image.

The second data set was from cancerous or normal breast tissues [21]. The data set has 295 samples, 8141 features. The data has two kinds of patients. The first class has 217 samples and the other class only has 78 samples, so this data set is unbalanced. To get better performance of the classifier, we extract 61 samples from the first class and 65 samples from the second class as training set. In the same way, 27 samples from the first class and the 26 samples from the second class were extracted as test set.

	SVM (RBF)	SVM (RBF)-En	SVM (linear)
The best success rate	90.566%	94.3396%	96.2264%

1) SVM (RBF) represents the SVM uses the RBF as the kernel function.

2) SVM (RBF)-En represents the Ensemble method use the SVM as learning algorithm, which is used RBF as the kernel function.

3) SVM (linear) represents the kernel function in the SVM is linear.

Table 1: The success rate after features extracted.

Results and Discussion

Classification results of the designed classifier

Applying the SVM and the ensemble method based on the SVM to the breast dataset. The results are shown in Table 1. It represents the best success rate from different classify algorithms after selecting features. Although, the success rate of the SVM (kernel function is linear) is better than the SVM (the kernel function is RBF), the former need more features and time to run the program. The success rate of the SVM-RBF is only 90.566%, but the success rate of the ensemble method is 94.3396% and fewer features are required. So we could obtain the conclusion that the ensemble method based on the SVM could improve the performance of the classifiers. In Figures 2 and 3, it is easy to find that when the number of genes is 34, the success rate of the training set and the test set get the best. These genes are called as mark genes which are mostly associated with the classification.

Comparison with adaboost with decision-trees

A receiver operation characteristics (ROC) curve is a two-dimensional depiction of classifier performance. To compare classifiers, the method used here is to calculate the area under the ROC curve, abbreviated AUC [22]. By calculating AUC of Figures 4-7, which represent the ROC curves about the Adaboost based on the SVM (the

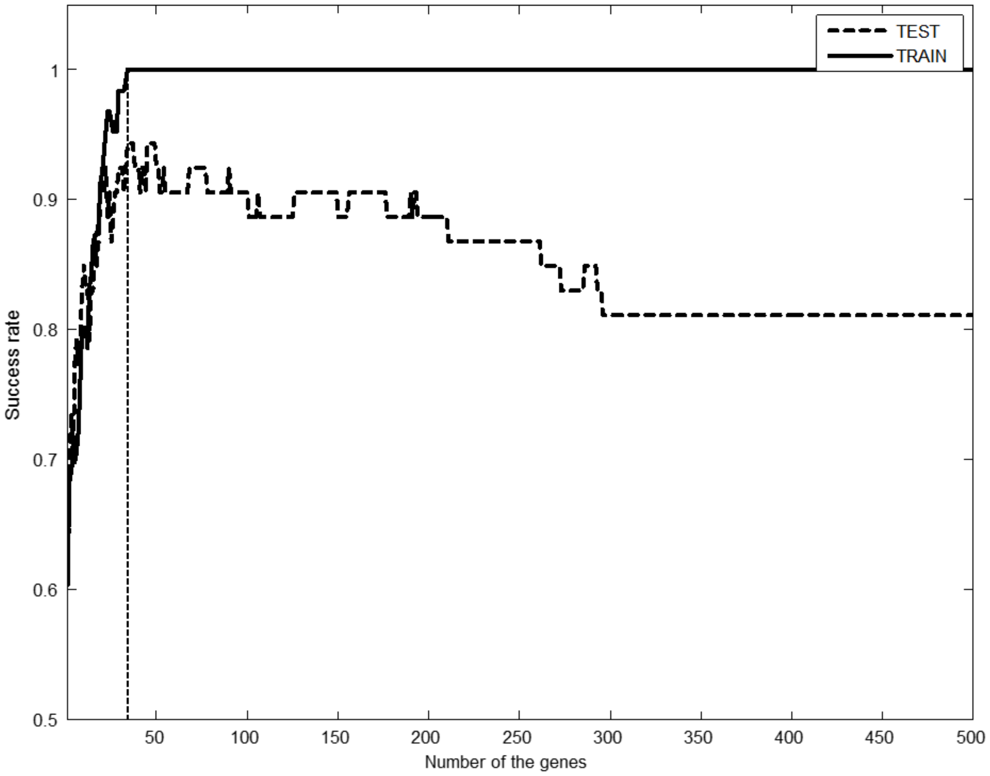

Figure 2: The success rate of the SVM method with the different number of the genes (Breast data).

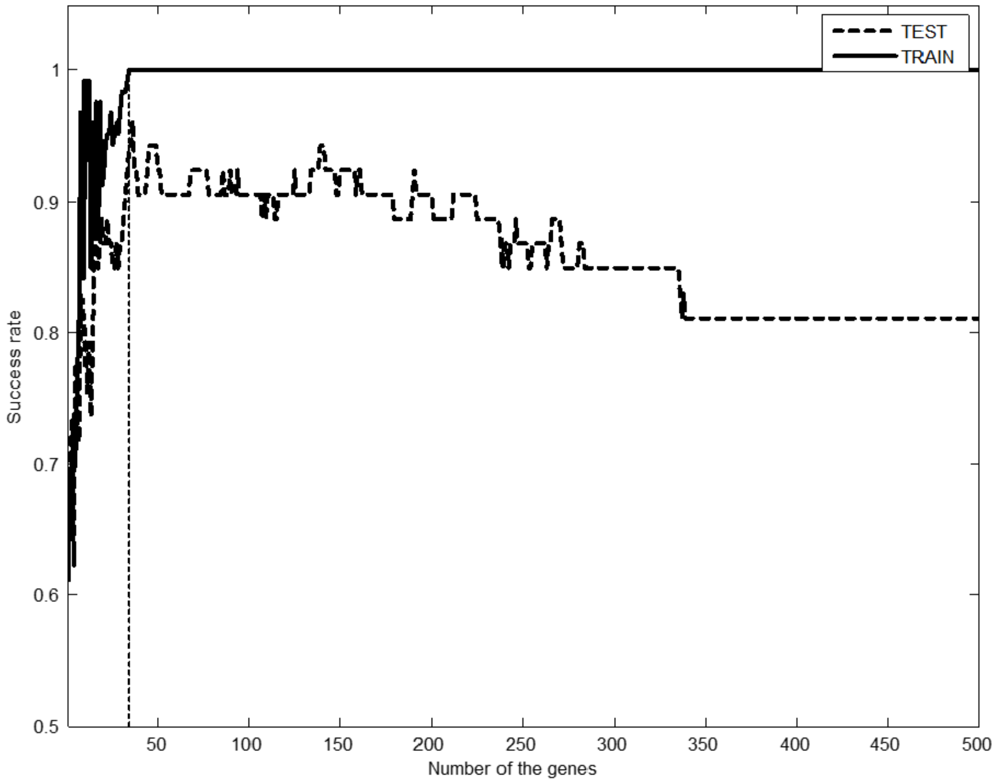

Figure 3: The success rate of the ensemble method based on the SVM with the different number of the genes (Breast data).

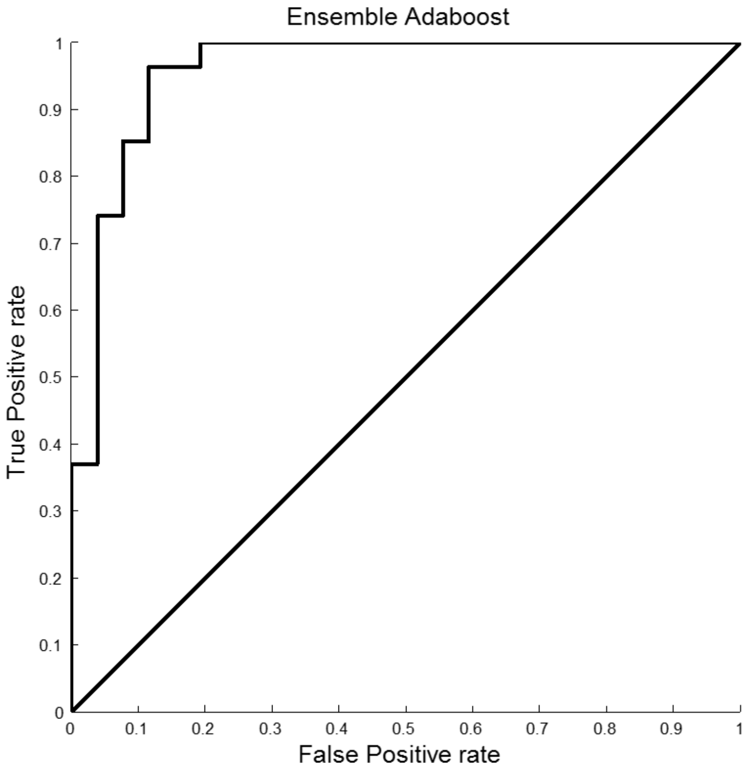

Figure 4: SVM (RBF)-Ada ROC.

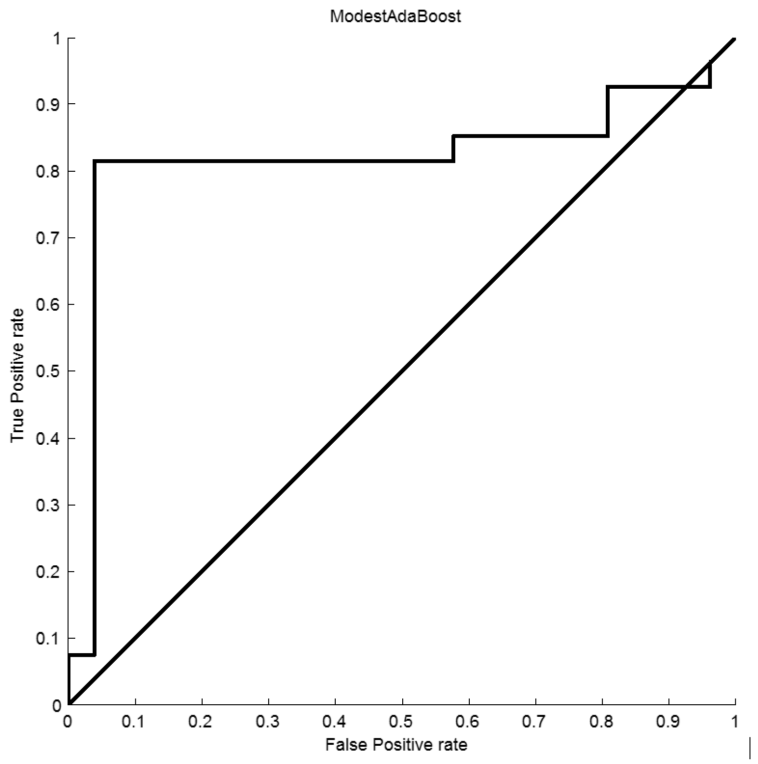

Figure 5: MAB ROC.

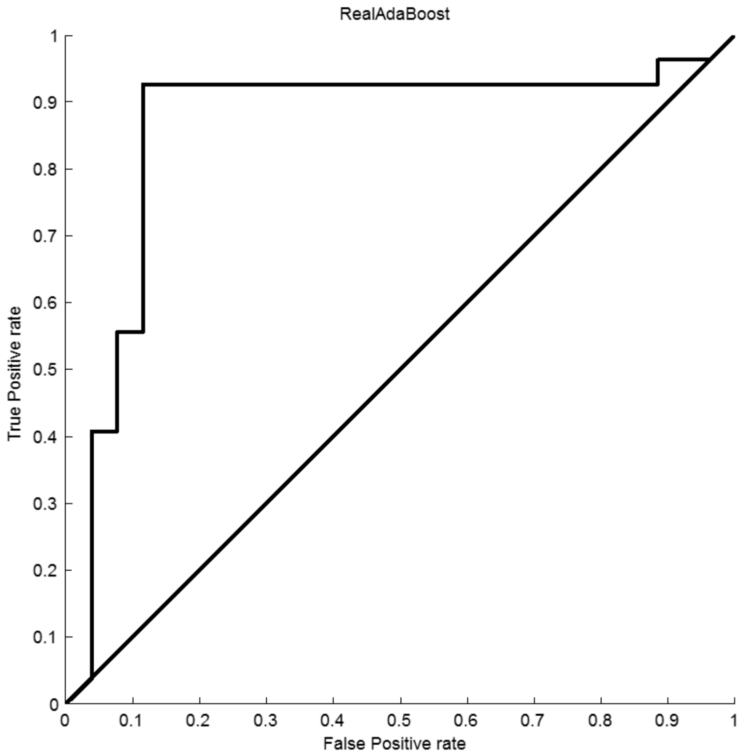

Figure 6: RAB ROC.

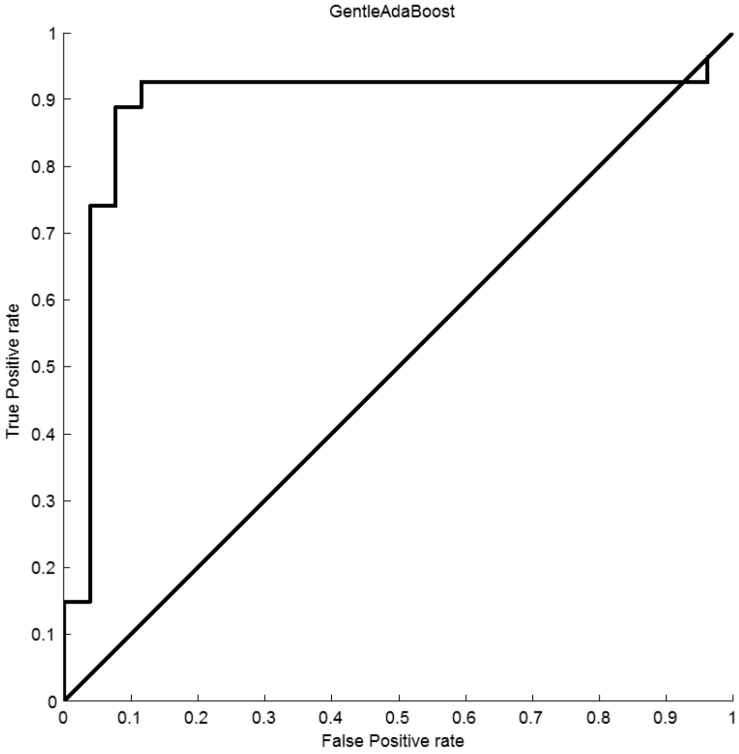

Figure 7: GAB ROC.

kernel funtion is RBF) and three kinds of the Adaboost algorithm which use the decision-trees as its learning algorithm (RAB, GAB, MAB). It is obvious that the AUC of three Adaboost algorithms based on the decision-trees are smaller than the Adaboost algorithm based on the SVM.

Effect of different issues on the performance of classifier

The importance of selecting the kernel function: Firstly, we apply the SVM (kernel function is RBF) to the leukemia data set, but the result is very poor with the success rate standing at only 58.8235%. If we set the type of the kernel function as linear, the success rate has been improved greatly with its success rate being 82.3529%. For the breast data set, the situation is the same. When we apply the SVM algorithm to classify the gene expression datasets, the kernel function is important for the classification results (Table 2).

The importance of selecting the features: Firstly, we apply the SVM or the SVM-Ensemble to the leukemia data set, and their results are shown in Table 3. It is easy to find that before selecting features, the success rate on the test set is very low. However, with significant features, whether the type of the kernel function is suitable or not, the success rate has been improved a great deal. Therefore whether to select features is a fatal factor in our experiments. What on earth causes the difference? Because the feature dimension in the leukemia data is 7129 and the breast data is 8141 which are more than the sample dimensions in the data sets, this will easily lead over-fitting. Besides, these features maybe include noisy, which could also have an impact on the classifiers.

From the experiments, we find that the ensemble method is ineffective on the leukemia data. The reason is that when we use the SVM (linear) only, we have got a very good result which is 91.1765%. If the SVM is done well on the data, then the ensemble will lose efficacy. That is to say, the ensemble based on the SVM does not always improve the performance of the classifiers as expected.

Data	SVM (kernel) function	The number of support vectors	The success rate
Leukemia	SVM (linear)	28	82.3529%
Data	SVM (RBF)	38	58.8235%
Breast	SVM (linear)	114	83.0189%
Data	SVM (RBF)	124	679245%

Table 2: The success rate of the SVM which is based on different kernel functions.

	SVM (linear)	SVM (RBF)	SVM (linear) -En[1]	SVM (RBF) -En
NONRFE	**82.3529%**	58.8235%	**58.8235%**	58.8235%
RFE	**91.1765%**	88.2353%	**91.1765%**	85.2941%

1) SVM (linear)-En represents the Ensemble method use the SVM as learning algorithm, whose kernel function is linear.
2) SVM (RBF)-En represents the Ensemble method use the SVM as learning algorithm, which is used RBF as the kernel function.
3) NONRFE represents don't extract the features.
4) RFE represents using the RFE method to extract the features.

Table 3: The success rate before and after feature selection (Leukemia data).

The number of features: How many features extracted could produce the best classifier? In order to study the problem more comprehensively, we conduct the experiments respectively on the SVM, Adaboost based on the SVM and the three Adaboost algorithms based on decision-trees. The results are shown in Table 4. For the SVM (RBF) and SVM (RBF)-Ensemble, their success rate reach to the highest level, standing at 90,566% and 94.3396% respectively, when there are 32 features. AdaBoost based on decision-trees was employed

The number of the features	SVM (linear)%	SVM (RBF)%	SVM (RBF)-En%	RAB[1]%	GAB[2]%	MAB[3]%
8146 (All)	67.9245	67.9245	67.9245	88.6792	81.1321	84.9057
4096	90.566	67.9245	67.9245	86.7925	92.4528	90.566
2048	94.3396	67.9245	67.9245	86.7925	88.6792	90.566
1024	94.3396	73.5849	73.5849	88.6792	86.7925	86.7925
512	96.2264	81.1321	81.1321	88.6792	83.0189	**92.4528**
256	**96.2264**	86.7925	86.7925	86.7925	88.6792	90.566
128	92.4528	90.566	90.5660	90.566	**94.3396**	84.9057
64	90.566	90.566	90.5660	90.566	88.6792	84.9057
32	86.7925	**90.566**	**94.3396**	86.7925	86.7925	86.7925
16	86.7925	86.7925	86.7925	**92.4528**	86.7925	83.0189
8	75.4717	79.2453	83.0189	84.9057	83.0189	86.7925
4	71.6981	69.8113	54.7170	86.7925	86.7925	79.2453
2	67.4603	67.9245	67.9245	83.0189	77.3585	77.3585
1	62.2642	60.3774	60.3773	71.6981	66.0377	66.0377

1) RAB represents the Real Adaboost Algorithm
2) GAB represents the Gentle Adaboost Algorithm
3) MAB represents the Modest Adaboost Algorithm

Table 4: The success rate of the different classifiers based on the different features.

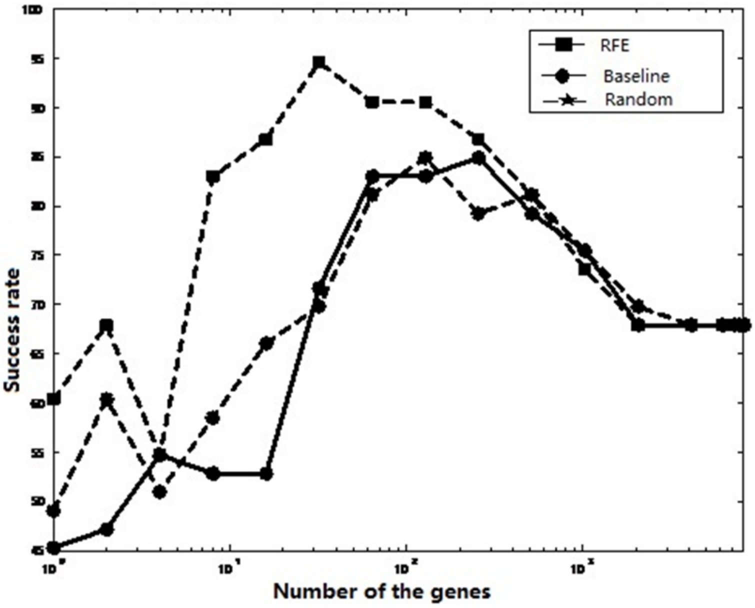

Figure 8: The different feature selection methods based on the SVM-ensemble classifier (Breast data).

and the results show that when the number of the features is 16, the RAB obtains the best success rate which is 92.4528%, and when the number of the features is 128, the best success rate of the GAB is 94.3396%, while to obtain the same success rate as GAB, the MAB need 512 features [23].

To the problem proposed in the beginning of this section, by the data in Table 4, we could solve it now. When you use the same data set, and adopt the RFE to extract features in an experiment. To get the best performance, you should know that different classifiers need different dimensions of the features. Because we only use the success rate as the criterion to assess the performance of the classifiers, we could not affirm the features that we selected will also display better on the Reject rate, Extremal margin, and Median margin [20] which are often used as the criterions to assess the performance of the classifiers. We will do this work in the future.

When you are given a particular classification technique, it is conceivable to select the best subset of features satisfying a given "model selection" criterion by exhaustive enumeration of all subsets of features. But in our paper, we do not adopt it because we apply it to some algorithms, and find that the results yield not significantly different from the method we used in the Table 4.

The different feature selection methods: In the last subsection, we have presented the feature selection is so important to classify the gene data. If we change the feature selection, what will the performance of the classifier be? We compare the RFE with the feature selection method which is proposed in Golub et al [20], we call this method as Baseline and we adopt random number. In this experiment, we only adopt the SVM-Ensemble classifier.

The ranking criterion of the features used by Golub *et al* is defined as Equation 1:

$$rank(i) = (\mu_i(+) - \mu_i(-)) / (\sigma_i(+) + \sigma_i(-)) \qquad (1)$$

where μ_i and σ_i are the mean and standard deviation of the gene expression values of gene i for all the samples. Large positive rank (i) values indicate strong correlation with class (+) whereas large negative rank (i) values indicate strong correlation with class (-). In Figure 8, when the number of genes is small, the success rate between different feature selection methods is obviously difference. But when the number of genes becomes large, the success rate is same. Because the more features, the more redundancy. By comparing the curves in Figure 8, we could conclude that the RFE performs better than the Baseline and the Random at the problem of the gene classification.

Discussion and Conclusion

In this paper, we have applied the feature selection to improve the Adaboost algorithm. In this algorithm, genes are selected by the RFE method. As a result, the obtained gene subset has good discriminative capability for classification. By observing these experimental results on two public microarray datasets, we get these conclusions:

1. The ensemble method improves the performance of the SVM classifiers at some extent.
2. Selecting the feature subset and how to extract the features have fatal effect on the problem of the gene classification.
3. By the ROC curve, we find the performances of the Adaboost based on the SVM is better than the Adaboost based on the decision-trees.

Although, we get some good results on the breast data set with the ensemble method based on the SVM. But when we apply the algorithm on the leukemia data set, the results are so bad. If the performance of the SVM is better on some data, then the ensemble will be useless. However, what lead ensemble method based on the SVM to be ineffective, this will be done in the future.

Acknowledgment

The research is partly supported by the National Natural Science Foundation of China (61203282) and National University students Innovation Training Project of Xiamen University (2015, PI:Gan Chen).

References

1. Boulesteix AL, Strobl C, Augustin T, Daumer M (2008) Evaluating microarray-based classifiers: an overview. Cancer Informatics 6: 77-97.
2. Dupuy A, Simon RM (2007) Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. Journal of the National Cancer Institute 99: 147-157.
3. Larrañaga P, Calvo B, Santana R, Bielza C, Galdiano J, et al. (2006) Machine learning in bioinformatics. Briefings in bioinformatics 7: 86-112.
4. Dudoit S, Fridlyand J, Speed TP (2002) Comparison of discrimination methods for the classification of tumors using gene expression data. Journal of the American statistical association 97: 77-87.
5. Liu H, Li J, Wong L (2002) A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns. Genome informatics 13: 51-60.
6. Jäger J, Sengupta R, Ruzzo WL (2003) Improved gene selection for classification of microarrays. In Pacific Symposium on Biocomputing 8: 53-64.
7. Bonev B, Escolano F, Cazorla M (2008) Feature selection, mutual information, and the classification of high-dimensional patterns. Pattern Analysis and Applications 11: 309-319.
8. Liu H, Sun J, Liu L, Zhang H (2009) Feature selection with dynamic mutual information. Pattern Recognition 42: 1330-1339.
9. Liu H, Liu L, Zhang H (2010) Ensemble gene selection for cancer classification. Pattern Recognition 43: 2763-2772.
10. Schapire RE, Singer Y (1999) Improved boosting algorithms using confidence-rated predictions. Machine learning 37: 297-336.
11. Friedman J, Hastie T, Tibshirani R (2000) Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). The annals of statistics 28: 337-407.
12. Vezhnevets A, Vezhnevets V (2005) Modest AdaBoost-teaching AdaBoost to generalize better. In Graphicon 12: 987-997.
13. Cho, Sung B, Won HH (2007) Cancer classification using ensemble of neural networks with multiple significant gene subsets. Applied Intelligence 26: 243-250.
14. O'Neill, Michael C, Song L (2003) Neural network analysis of lymphoma microarray data: prognosis and diagnosis near-perfect. BMC bioinformatics 4: 13.
15. Khan J, Wei JS, Ringner M, Saal LH, Ladanyi M, et al. (2001) Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nature medicine 7: 673-679.
16. Guyon I, Weston J, Barnhill S, Vapnik V (2002) Gene selection for cancer classification using support vector machines. Machine learning 46: 389-422.
17. Freund Y, Schapire RE (1997) A decision-theoretic generalization of on-line learning and an application to boosting. Journal of computer and system sciences 55: 119-39.
18. Boser BE, Guyon IM, Vapnik VN (1992) A training algorithm for optimal margin classifiers. In Proceedings of the fifth annual workshop on Computational learning theory. ACM.
19. Joachims T (2000) Estimating the generalization performance of a SVM efficiently. Universität Dortmund.
20. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, et al. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286: 531-537.
21. Chuang HY, Lee E, Liu YT, Lee D, Ideker T(2007) Network-based classification of breast cancer metastasis. Molecular systems biology 3: 140.
22. Guyon I, Makhoul J, Schwartz R, Vapnik V (1998) What size test set gives good error rate estimates?. Pattern Analysis and Machine Intelligence, IEEE Transactions on 20: 52-64.
23. Fawcett T (2006) An introduction to ROC analysis. Pattern recognition letters 27: 861-874.