Classification and identification of differential gene expression for microarray data: improvement of the random forest method

Xiaoyan Wu, Zhenyu Wu, Kang Li*

Department of Biostatistics, Public Health College, Harbin Medical University, Harbin 150086, P.R. China

Abstract Classification and gene selection of microarray data have been important aspects of the investigation of gene expression data in biomedical researches. The analysis of gene expression data presents a new challenge for statistical methods because of its high dimensionality. Random forest has been used to deal with the problem. We present a new classifier named Recursive Random Forest which selects genes automatically and improves the accuracy of classification based on random forest. Three microarray datasets (ALL-AML Leukemia data, Colon Cancer data and Prostate cancer data) were analyzed using Recursive Random Forest. Although the genes selected from the microarray data were only a few, they were effective on cancer prediction and their biological functions have been confirmed. Especially on the ALL-AML Leukemia data, it achieved a perfect accuracy on the test set using only three genes (selected from over 7000). We also research the properties of random forest and recursive random forest on simulated experiments. Recursive random forest provides more useful information than simply using random forest for the further biological experiment, clinical diagnoses and disease therapies because of its function of gene selection, which would probably become an excellent 'tool' on sample classification and gene selection for microarray data. Source code written in R for Recursive Random Forest is available from http://yxzy.hrbmu.edu.cn/gongwei/biostatistics/.

Key words: recursive random forest; random forest; microarray data; classification; gene selection

I. INTRODUCTION

DNA microarray technology [1] allows us to investigate the expression levels of tens of thousands of different genes simultaneously and produce a large amount of gene expression data. The analysis of gene expression data mainly includes classification and feature selection which may find the smallest possible set of useful genes with equally good predictive performance. Both of these therefore have practical implications for clinical diagnoses and for disease therapies.

*To whom correspondence should be addressed.

The investigation of gene expression data presents a new challenge for statistical methods because of the large number of genes (p) and relatively small number of samples (n): $p \ll n$. Recently, much attention has been paid to combination classifiers, which provide promising approaches to resolve the problem. One of the effective combination classifiers is random forest, which was developed by Breiman [2] and composed of many classification trees.

A number of authors have applied random forest to the field of gene expression data. Some of them have used it for feature selection, while others [3-6] have compared random forest with other statistical methods for the classification of gene expression data. All the results so far have indicated that random forest provided similar or improved performance. Jiang *et al.* [7] and Díaz *et al.* [8] also tried to drop genes sequentially to improve the accuracy of classification. However, neither of these two methods considered whether the eliminated genes contributed to the classification performance. Although random forest can deal with the data which has curse of dimensionality, so much undifferentiated genes affect the classification accuracy greatly and the identification of differentiated genes.

Here, we assess the classification performance of random forest on real gene expression datasets and measure its classification capability on simulated data. In addition, we present a new algorithm based on random forest which we call 'recursive random forest' to improve the accuracy of classification and feature selection. We then analyze both microarray data and simulated datasets to evaluate the performance of recursive random forest synthetically.

II. DATASETS

A. Gene expression data

The main characters of the microarray data are

shown in Table Ⅰ. The colon and prostate datasets were from Díaz [8] and were divided into normal and abnormal (cancer). The data for acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) were downloaded from the Broad Institute Cancer Website (http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi).

TABLE Ⅰ. MAIN CHARACTERISTICS OF THE GENE EXPRESSION DATASETS

Datasets	Genes	Samples	Pathology	
			normal	abnormal
Colon	2000	62	22	40
Prostate	6033	102	50	52
Leukemia	7129	72	47	25

B. Datasets for measuring the performance of random forest.

We established the performance of random forest based on the simulated training data and classified the test data with it. Two classes of training data and test data were designed and different values of three variables were defined as follows: p representing the number of differentiated variables, θ meaning the parameter of AUC, and ρ denoting the correlation coefficient of the differentiated variables, with variance σ^2=1. The training and test datasets were all randomly chosen from a normal distribution; while undifferentiated variables were also randomly selected from a standard normal distribution, and then were added to the differentiated ones in different amount.

C. Simulated datasets for measuring the performance of Recursive Random Forest

To assess the property of Recursive Random Forest, we simulated datasets from a normal distribution. Suppose only five differentiated variables whose parameters were set according to the different AUC value (the value of the area under the curve (AUC) of the receiver operating characteristic (ROC) curve) were effective for classification under the condition that $\rho = 0$ and $\rho = 0.5$. Two thousand undifferentiated variables were added to the original simulated dataset.

III. RESULTS

A. Real gene expression data

We analyzed gene expression data via LOO to obtain a precise classification result. Table Ⅱ shows the classification results of random forest on real microarray datasets. We see good predictive results for microarray data by random forest. However, the classifier constructed by Recursive Random Forest with only a few genes achieves a better classification performance and feature selection.

TABLE Ⅱ. COMPARISON OF RESULTS BETWEEN RANDOM FOREST AND RECURSIVE RANDOM FOREST METHODS ON MICROARRAY DATA

Datasets	Before selection (Random forest)		After selection (Recursive random forest)	
	Genes	AUC*	Genes	AUC*
Colon	2000	0.868	20	0.956
Prostate	6033	0.934	8	0.965
Leukemia	7129	0.996	3	0.996

*AUC means the value of the area under the curve of the receiver operating characteristic curve.

It was proposed from these microarray data that genes helpful to predict cancer class may also be valuable for predicting cancer pathogenesis and therapy response. For example, there were only three genes in the optimal classifier for leukemia data and gene 1882 has been reported as the most important [9], [10] and is related to amyloid angiopathy and cerebral hemorrhage. The Zyxin gene 4847 that encodes a LIM domain protein important in cell adhesion in fibroblasts [11] plays an important role in distinguishing ALL from AML [12]. Other investigations of the leukemia dataset have also identified the zyxin gene which has significance in class prediction [13]. The differentiated genes chosen by Recursive Random Forest can thus classify the categories of cancer properly and we can determine their specific functions relate to disease accordingly. The selected genes which have no description may also merit further investigation.

B. The performance of Recursive Random Forest

We concluded from the primary results that the larger the true value of AUC, the better the predictive performance and the higher proportion of differentiated variables in the optimal classifier would be achieved Many datasets contain three or more differentiated variables in the optimal classifier. It can be inferred that not only do we have attractive variable selection outcomes, but also the classification accuracy is

increased. The median value of AUC is increased from o.676 to 0.790 ($\rho = 0$) and 0.892 to 0.918($\rho = 0.5$). The different distributions of AUC on condition of no undifferentiated variables, before and after variable selection are illustrated (Fig. 1 and Fig. 2).

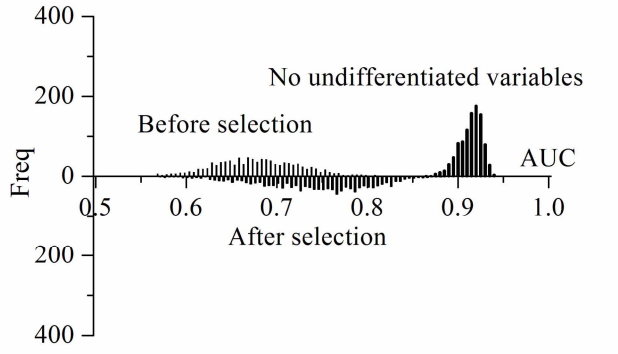

Figure 1. The distributions of AUC on condition of no undifferentiated variables (random forest), before selection (random forest) and after selection (Recursive Random Forest) ($\theta = 0.95$, $\rho = 0$).

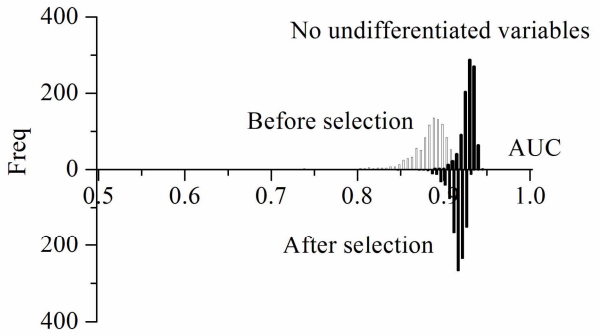

Figure 2. The distributions of AUC on condition of no undifferentiated variables (random forest), before selection (random forest) and after selection (Recursive Random Forest) ($\theta = 0.95$, $\rho = 0.5$).

C. The performance of random forest

The aim of the simulated experiments was to determine the classification ability of random forest in the existence of different numbers of undifferentiated variables when the number of differentiated variables was constant. The classification capability of random forest decreases as the number of undifferentiated variables increases for a fixed number of differentiated variables. The classification competence also corresponds to the extent of correlation of the differentiated variables. The stronger the correlation among the differentiated genes is, the better the classification performance achieved under the condition where the value of AUC is fixed. The information distribution and the correlation among differentiated variables are factors which may affect classification accuracy. We indicated the varieties of classification ability under dissimilar number of differentiated variables with the same number of undifferentiated variables (Fig. 3 and Fig. 4).

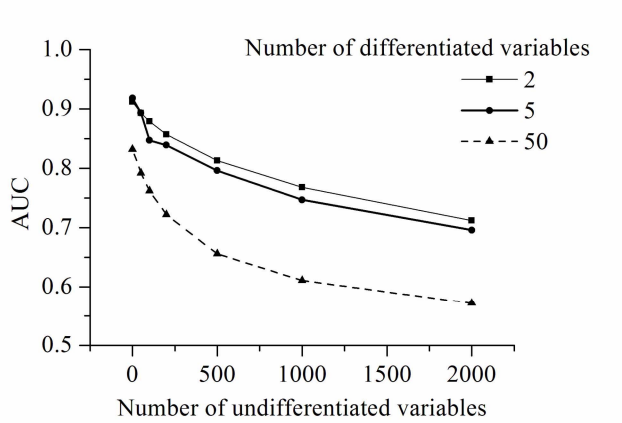

Figure 3. The performance of random forest under different (2, 5 and 50) differentiated variables ($\theta = 0.95$, $\rho = 0$).

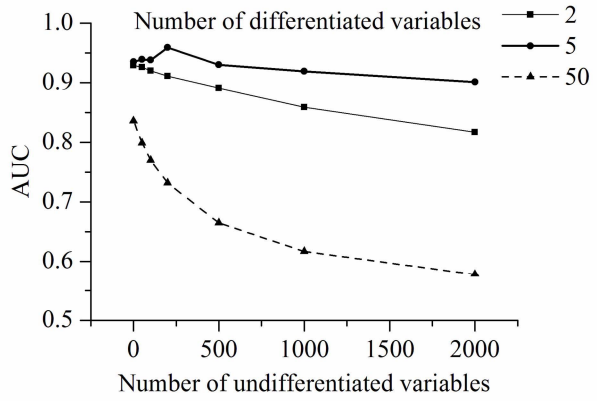

Figure 4. The performance of random forest under different (2, 5 and 50) differentiated variables ($\theta = 0.95$, $\rho = 0.5$).

IV. DISCUSSION

In this paper, we have evaluated the classification performance of random forest on gene expression datasets and simulated data. It presents good predictive ability for microarray data. The classification capability gradually decreased as the number of undifferentiated variables increased when the number of differentiated variables was fixed in the simulated datasets. This

result is strongly associated with the power of a single differentiated variable when the extent of separability of the dichotomic classifying variables in each simulated dataset is given: the fewer variables that contribute to classification, the greater the classification efficiency will be. Besides, the higher the correlation among differentiated genes, the better the predictive accuracy will be. To estimate the effect of undifferentiated variables on classification, we proposed the algorithm of recursive random forest which aims to obtain the optimal classifier together with good feature selection by ordering the variables dynamically and deleting variables according to the value of AUC.

The study of microarray data using recursive random forest indicated that we could obtain better classification performance and feature selection. The number of genes in the optimal classifier was far smaller than the original number of genes, particularly in the leukemia data there were only three genes in the classifier and their biological function had been confirmed. Although few genes were retained, the classifier improved the predictive performance greatly and the values of their AUC were all higher than 0.95. Moreover, we carried out simulated experiments in order to validate the performance of recursive random forest. It not only boosts the classification accuracy, but also retains as many as possible of the true differentiated variables in the optimal classifier. Recursive random forest distinguishes differentiated variables from all variables and helps with feature selection.

Although in our research, random forest achieved good predictive ability before gene selection (possibly because the differentiated genes contribute more than undifferentiated genes); a better classification result has been attained by recursive random forest. But gene selection is determined by variable importance through random forest, so the differentiated genes may not be the best because of random effects. Using other algorithms, such as Genetic Algorithms (GAs) may resolve this issue. Moreover, attention should be paid to the problem that the number of genes is so large that some genes picked by the classifier may have nothing to do with the classification because of random fluctuations. It would be possible to avoid this by establishing a classifier based on reliable biological information.

ACKNOWLEDGEMENTS

The research was supported by the National Natural Science Foundation of China (Grant No. 30371253).We sincerely appreciate comments on the English from Professor Chris Tyler-Smith, Yali Xue and Edgar Love.

REFERENCES

[1] D. Murphy, “Gene expression studies using microarrays: principles, problems, and prospects,” Adv. Physiol. Educ, vol. 26, pp. 256-270, 2002.

[2] L .Breiman, “Random forests,” Mach. Learn, vol. 45, pp. 5-32, 2001.

[3] A. Bureau, J. Dupuis, K. Falls, K.L. Lunetta, B. Hayward, T.P. Keith, P.V. Eerdewegh, “Identifying SNPs Predictive of Phenotype Using Random Forests,” Genet.. Epidemiol, 2005; 28:171–182.

[4] X.H. Huang, W. Pan, S. Grindle, X.Q. Han, Y.J. Chen, S.J. Park, L.W. Miller, J. Hall, “A comparative study of discriminating human heart failure etiology using gene expression profiles,” BMC Bioinformatics, vol. 6: 205, 2005.

[5] A. Cutler, J.R. Stevens, “Random Forests for Microarrays,” Methods Enzymol, vol. 411, pp. 422-432, 2002.

[6] K. Hoffmann, M.J. Firth, A.H. Beesley, N.H. Klerk, U.R. Kees, “Translating microarray data for diagnostic testing in childhood leukaemia,” BMC Cancer, vol. 6: 229, 2006.

[7] H.Y. Jiang, Y.P. Deng, H.S. Chen, Lin. Tao, Q.Y. Sha, J. Chen, C.J. Tsai, S.L. Zhang, “Joint analysis of two microarray gene-expression data sets to select lung adenocarcinoma marker genes,” BMC Bioinformatics,vol. 5: 81, 2004.

[8] R. Díaz, S. Alvarez, “Gene selection and classification of microarray data using random forest,” BMC Bioinformatics, vol. 7:3, 2006.

[9] P. Subramani, R. Sahu, S. Verma, “Feature selection using Haar wavelet power spectrum,” BMC Bioinformatics, vol. 7:432, 2006.

[10] Y. Wang, I.V.Tetko, M.A. Hall, E. Frank, A. Facius, K.F.X . Mayer, H.W. Mewes, “Gene selection from microarray data for cancer classification—a machine learning approach,” Comput. Biol. Chem, vol. 29, pp. 37–46, 2005.

[11] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, “Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring,” Science, vol. 286:531, 1999.

[12] S.G. Baker, B.S. Kramer, “Identifying genes that contribute most to good classification in microarrays," BMC Bioinformatics, vol. 7:407, 2006.

[13] P. Broberg, “Ranking genes with respect to differential expression,” Genome Biol, 3(9): preprint0007.1-0007.23, 2002.