Journal of Convergence Information Technology
Volume 5, Number 6, August 2010

# Gene selection using random forest and proximity differences criterion on DNA microarray data

Qifeng Zhou[*Corresponding author], Wencai Hong, Linkai Luo, Fan Yang
*Department of Automation, Xiamen University, Xiamen 361005, China*
*zhouqf@xmu.edu.cn, 369414951@163.com, luolk@xmu.edu.cn*
doi: 10.4156/jcit.vol5.issue6.17

## *Abstract*

*Selection of relevant genes for sample classification is a common task in most gene expression studies. As a powerful classification approach, random forest has been applied in this field, and it shows excellent performance compared with other classification methods. The measure of variable importance is the key of gene selection using random forest. However, the existing methods just consider the original variable importance measure based on the OOB error. In this paper, we proposed a new variable importance measure based on the difference of proximity matrix, and used it for gene selection on DNA microarray data. Compared with the existing variable importance analysis of random forest, the new method is more sensitive to information gene and yields small sets of genes while preserving predictive accuracy.*

**Keywords:** *Gene selection, Random forest, Variable importance, Proximity matrix*

## 1. Introduction

Gene selection is an important aspect of microarray data analysis, and has been a central issue in recent years [1-4]. Normally, a microarray data set contains a large number of genes (usually several thousands or more) and a relatively small number of samples (ten to several hundreds). Among all the genes, many are irrelevant, insignificant or redundant to the discriminate problem under investigation. Researchers try to identify the smallest possible set of genes which are most relevant to sample classification, for example, those differentiate between normal and cancerous tissue samples.

Gene selection can improve the predictive accuracy of classifiers by using only discriminative genes. It also saves computational costs by reducing dimensionality. Moreover, if it is possible to identify a small subset of biologically relevant genes, it may provide insights into understanding the underlying mechanism of a specific biological phenomenon.

Many gene selection algorithms have been proposed for microarray data analysis over the past few years. However, most of the studies are related to binary (two-class) gene selection problems [5-7], and only a few involve multiclass gene selection and classification [8-10]. As the multiclass problem is intrinsically more difficult and presents more challenges, it is worthy of further investigation.

In binary feature selection problems, SVM-RFE is one of most popular algorithm and widely used for gene selection [1]. Just like SVM itself, SVM-RFE was originally designed to solve binary gene selection problems. Several groups have extended it to solve multiclass problems using one-versus-all techniques. However, the genes selected from one binary gene selection problem may reduce the classification performance in other binary problems. In contrast, random forest is a classification algorithm well suited for microarray data [11]. It is natural to resolve multiclass problems and shows excellent performance even when most predictive variables are noise. It can be used when the number of variables is much larger than the number of observations. Moreover, random forest can return measures of variable importance, which has been used for gene selection [12].

Random forest can also give measures of samples proximity. In this paper, we take the proximity as a special kernel measure, and use the differences of samples proximity instead of Out-of-bag (OOB) error as the criterion to select the information genes. Compared with the original variable importance analysis of random forest, the new method is more sensitive to information gene and yields small sets of genes while preserving predictive accuracy.

## 2. Background

### 2.1 Random forest (RF)

Random forest is a machine learning algorithm for classification or regression developed by Leo Breiman [13]. It uses an ensemble of unpruned (grown fully) trees. Each of the classification trees is built using a bootstrap sample of the data and at each split the candidate set of variables is a random subset of the variables. The predictions are made by majority vote of the trees (in classification) or averaging their outputs (in regression). Thus, random forest uses Bagging (bootstrap aggregation), a successful approach for combining unstable learners [14], and random variable selection for tree building. The idea is maintaining the low-bias trees while reducing their correlation with each other. Breiman has shown that an upper bound on the generalization error of random forest is given by $r(1-s^2)/s^2$, where *r* is a measure of the correlation between the trees, and s is a measure of their strength[13]. The two sources of randomness, random inputs and random features result in a good generalization capability and make random forest accurate classifiers in different domains [15, 16]. Using random feature selection to split each node yields error rates that compare even favorably to Adaboost[17], but are more robust with respect to noise[18].

Random forest has excellent performance in classification tasks, comparable to support vector machines. It can be used both for two-class and multiclass problems, and there is little need to fine-tune parameters to achieve excellent performance. It is also an ideal tool for DNA microarray data analysis. Random forest can be used when there are many more variables than observations. It has good predictive performance even when most predictive variables are noise, and therefore it does not require a pre-selection of genes. It can handle a mixture of categorical and continuous predictors.

Random forest also provides additional features that increase its utility for gene selection and classification modeling in microarray data:

**Out-of-bag (OOB) prediction**: Since each tree in the forest is grown on a bootstrap sample of the data, the data left out of the bootstrap sample, the "out-of-bag" (OOB) data, can be used as a legitimate test set for that tree. On average, one-third of the training data will be "out-of-bag" for a given tree [19]. Consequently, we can use these OOB predictions to estimate the error rate of the full ensemble. This is similar to a cross-validation performance estimate, but at a much lower computational cost.

**Variable importance measure**: Random forest returns several measures of variable importance. The widely used one is based on the variable random permute and OOB error estimation. It is computed as follows: first compute the OOB error rate (or MSE) of each tree, and also compute the same for OOB data with one variable permuted; take the difference between these. The measure is the mean difference (over all trees) divided by the standard error of these differences. Variable importance can be used for variable ranking and model interpretation.

**Proximity matrix computing**: Putting all data in a tree, they will reach some leaf nodes respectively. So the frequency that two cases reach the same leaf node in the whole forest can be used to estimation their proximity (or similarity) degree. The all samples' proximity constructs a matrix. Proximity matrix is symmetric with 1 on the diagonal and in 0 to 1 off the diagonal. It is useful for similarity or difference measure between cases and can be taken as a special kernel measure for gene selection and classification.

### 2.2 Gene selection method based on variable importance measure of OOB error

Variable reduction based on Random Forest's variable importance measure is a potential way to optimize the Random Forest algorithm. Ramón and Sara proposed a method of gene selection in classification problems based on random forest [12]. The algorithm first ranks the variables (genes) according their importance measure. Then, iteratively fit random forests, at each iteration building a new forest after discarding those variables (genes) with the smallest variable importance; the selected set of genes is the one that yields the smallest OOB error rate. The removal of irrelevant variables may improve the performance of the algorithm upon retraining and may help improve the interpretability of the model. Compared with other classification methods, such as DLDA, SVM, random forest shows good performance and can yield small sets of genes. In this paper we call the method OOB _VI.

The measure of variable importance is the key of gene selection using random forest. However, the existing methods just consider the original measure of variable importance based on the OOB error. So it is necessary to compare the performance of different measure and find a better one.

## 3. Methods

Beside OOB error estimate, some other measures of variable importance are available. Such as computing the total decrease in node impurities from splitting on the variable, averaged over all trees. For classification, the node impurity is measured by the Gini index. From our experiments, the performance of those two measures (based on OOB error and Gini index) is almost the same, so we no long consider the Gini index. In this paper we proposed a new variable importance measure based on the difference of proximity matrix.

### 3.1 Variable importance measure based on proximity matrix difference

Consider *n* training data pairs: $S = \{(x_{i,} y_i)\},\ i = 1,...,n$, where $x_i \in \mathbb{R}^l$ is a feature vector representing the *i*-th sample, and $y_i$ is the class label of $x_i$. For a binary problem, $y_i \in \{-1,1\}$, while for *k*-class(*k*>2) problem, $y_i \in \{1,2,...,k\}$. The variable importance of proximity difference is computed as follows:

(1) Train a random forest and return a proximity matrix of all data: $P=\{p_{ij}, i, j = 1,..n\}$;

(2) Compute the proximity ratio between inner-class and the inter-class:

$$C = Ps / Pd \tag{1}$$

where

$$Ps = \sum_{i,j=1}^{l} p_{ij} \quad (if\ y_i = y_j)$$

$$P_d = \sum_{i,j=1}^{l} p_{ij} \quad (if\ y_i \neq y_j);$$

(3) Compute the same for proximity matrix with some variable permuted and get the ratios $C_i$ , $i = 1,...l$;

(4) The final measure is the difference between these $C - C_i$ , $i = 1,...l$.

Proximity matrix can be taken as a special measure that reflects the distribution of samples in a "proximity space". When a variable that contributes to samples' distribution is "noised up" (like permuted), the inner-class proximity will decrease while the inter-class proximity will increase, so the ratio $C_i$ should noticeably degrade. On the other hand, if a variable is irrelevant or redundant, "noising" it up should have little effect on the performance.

### 3.2 Performances analysis of proximity matrix

A good variable importance measure must be robust to the effects of model parameters and sensitive to the change of variable. In random forest the main parameters are *ntree*, the number of trees, *mtry*, and the size of the variable subset. Generally, *ntree* is fixed as 1000 or 2000, and *mtry* with the default value $\sqrt{l}$ ($l$ is the number of variable) performs the best. Some empirically study results show the robustness of random forest to changes in its parameters [12, 20].

#### 3.2.1 Robustness of proximity matrix to two sources of randomness

Beside the parameters mentioned above, we should analyze the effect of the two sources of randomness (random inputs and random features) on proximity matrix. Here, we design the following statistic variable:

Define the difference of two proximity matrixes:

$$D_2 = P_1 - P_2 = \frac{1}{n^2}\sum_{i=1,j=1}^{n} |p1(i,j) - p2(i,j)| \tag{2}$$

where $p(i, j)$ is the matrix element in *i*-th row and *j*-th column.

Define the difference of *m* proximity matrixes:

$$D_m = \frac{1}{m-1}\sum_{i=1}^{m}(P_i - \bar{P}) \tag{3}$$

where $\bar{P}$ is the mean of *m* proximity matrixes: $\bar{P} = \frac{1}{m}\sum_{i=1}^{m}|P_i|$.

The statistic variable $D_m$ is computed on 4 public microarray data sets. Figure.1 shows the results.

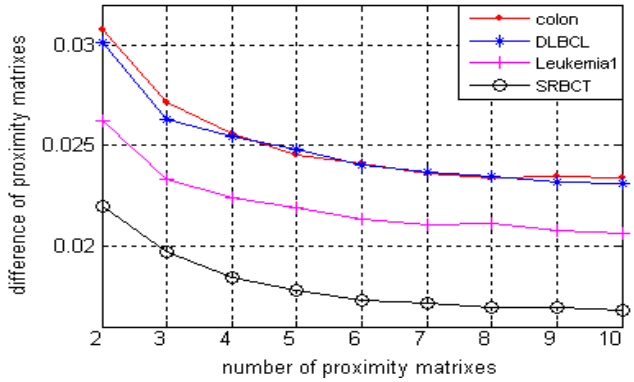



**Figure 1.**Robustness of proximity matrix, describes the average difference of proximity matrix repeated modeling in 10 times on 4 microarray data sets: colon, DLBCL, Leukemia1 and SRBCT, (the description of the data sets as Table 1).

From Figure1, we can see those statistics $D_m$ are very small on all 4 data sets. That is, random forest's two sources of randomness have little effect on proximity matrix.

#### 3.2.2 Sensitiveness of proximity matrix to the change of variable

OOB error estimate sometime is not very sensitive in variable importance measure, especially in multiclass microarray data set with a large amount of redundant or irrelevant variables. Moreover, since the number of microarray sample is very small, the OOB error estimation is not entirely reliable. For example, consider the Leukemia1 data set. It has 72 samples and 5328 genes, three subclasses. Establish a random forest, set ntree=2000, mtry=72, get the OOB error and the proximity ratio, permute the gene #2374, then get the new ones. The total OOB error keeps unchanged (is still 3.69%), but the proximity ratio increased about 3.228. So the proximity matrix is more sensitive to variable's change than OOB error.

### 3.3 Gene selection method under proximity differences criterion

Using the variable importance measure of proximity difference we propose a new gene selection algorithm (Hereinafter, we call it Prox_VI):

(1) Partition the data for *n*-fold cross-validation (CV).

(2) On each CV training set, train a random forest model on all variables and use the variable importance measure of proximity difference to rank them. Record the CV test set predictions.

(3) According the variable ranking to remove the least important *m* variables and retrain the model, predict the CV test set. Repeat removal of *m* variables until there are 1 or 2 left.

(4) Aggregate results from all *n* CV partitions and compute the error rate at each step of reduction.
(5) Repeat steps (1)-(4) to "smooth out" the variability, comput the average test error rate.
(6) Choose the smallest set of genes whose error rate is the smallest in all the fitted random forest.

As [20] mentioned, our methods is also non-recursive, that is, on each training run of CV, the variable importance is calculated just once, at the beginning, to avoid the overfitting resulting from recalculated variable importance.

## 4. Experiments and discussion

We tested our proposed new algorithm on eight microarray data sets. The specifications of the data sets are listed in Table 1.

**Table 1**. Description of the data sets

| Data set | Samples | Genes | Classes | Reference |
|---|---|---|---|---|
| Colon | 62 | 2000 | 2 | [21] |
| Leukemia | 72 | 3571 | 2 | [22] |
| DLBCL | 77 | 5469 | 2 | [23] |
| Prostate Tumor | 102 | 10509 | 2 | [23] |
| Leukemia1 | 72 | 5328 | 3 | [23] |
| SRBCT | 83 | 2308 | 4 | [23] |
| Brain_Tumor1 | 90 | 5920 | 5 | [23] |
| 9_Tumors | 60 | 5726 | 9 | [23] |

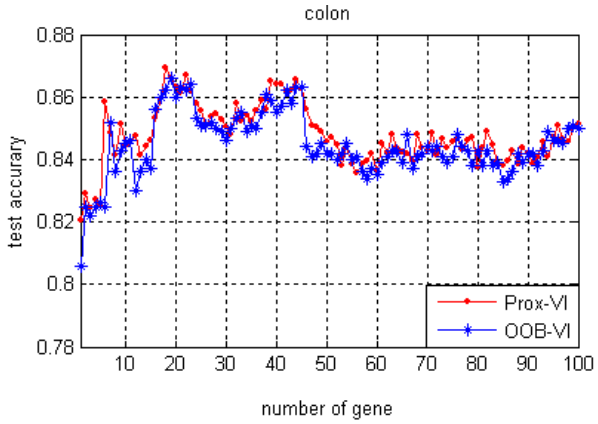
We compared the predictive performance of Prox_VI and OOB_VI with:

a. Select the most important 100 genes using random forest variable importance with OOB error and proximity difference respectively; compare the predictive performance in *K*-nearest neighbor (KNN) algorithm. The results are shown in Figure 2.

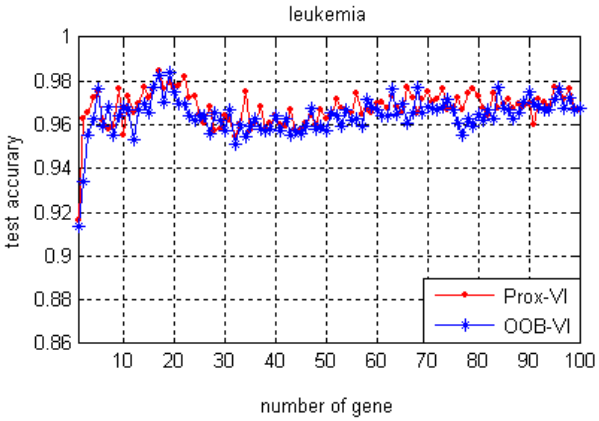
b. Select the smallest gene set using the two types of variable importance and compare their predictive performance in random forest. The results are shown in Table2.

Experiments are carried out in R language V2.9.0 with Random Forest Package and matlab 7.0.

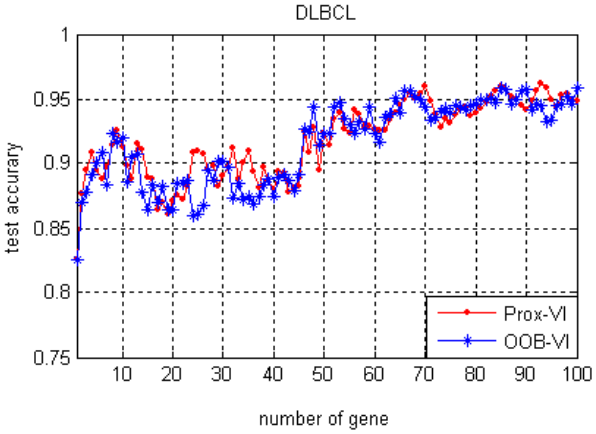
In the first experiment, we select the gene set S= {1 , 2 , 3 , … , 99 , 100} according to the variable importance ranking with two methods respectively. In order to evaluate the performance of two methods objectively, we take KNN as the "standard classifier". The KNN algorithm is simple and sensitive to the selected features. It is usual to use the Euclidean distance, while in our experiment on microarray datasets, the Manhattan distance shows better performance, and so we choose it as the KNN measure. The number of neighbors *k* is chosen by cross-validation (here $k \in \{1,3,5,7,9\}$ , and randomly repeat 20 times.



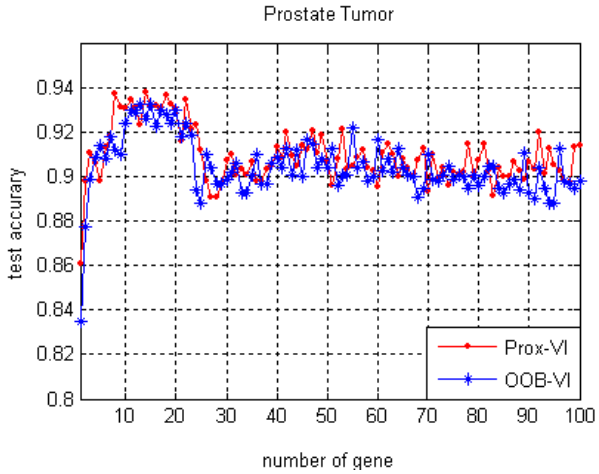

(a) Performance curves of the two methods in colon data set.



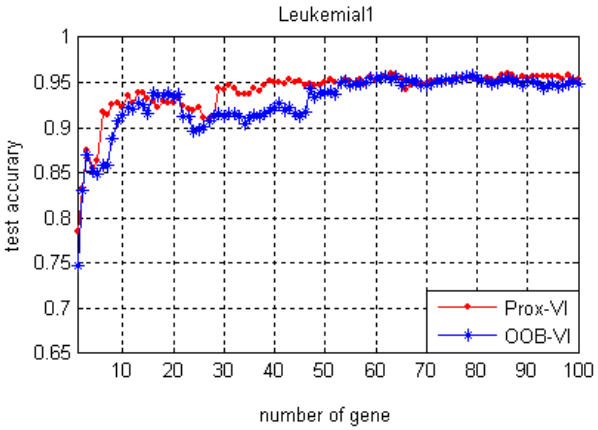

(b) Performance curves of the two methods in Leukemia data set.



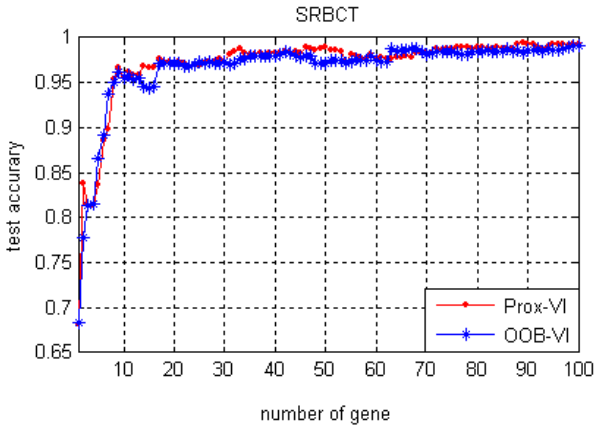

(c) Performance curves of the two methods in DLBCL data set.



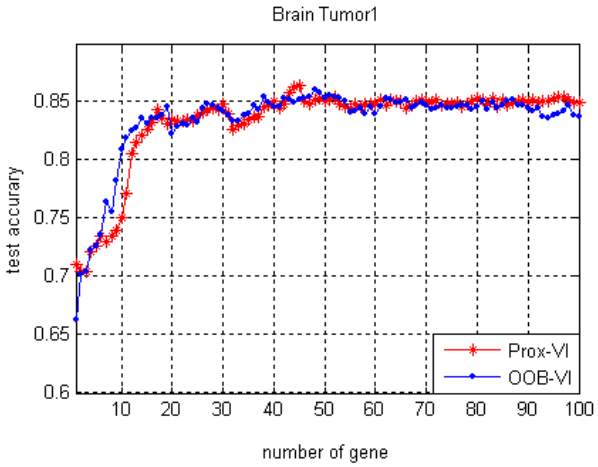

(d) Performance curves of the two methods in Prostate Tumor data set.



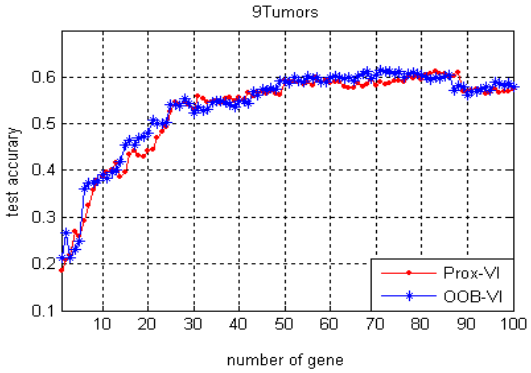

(e) Performance curves of the two methods in Leukemia1 data set.




(f) Performance curves of the two methods in SRBCT data set.




(g) Performance curves of the two methods in Brain_Tumor1 data set.




(h) Performance curves of the two methods in 9_Tumors data set.

**Figure 2.** 100 most important genes selected by two methods (*ntree*=2000, *mtry*=$\sqrt{I}$ ) and their performance compare in KNN.

From Figure 2, we can see that the most important 100 gene sets, on the whole, the KNN predictive performance for 9_Tumors data set, our new variable importance measure is worse than OOB measure, for colon and DLBCL data sets the two methods are comparable, and for other 5 microarray data sets our new method performed better. Although the difference of predictive accuracy is not remarkable, for microarray data set it is still significant. The results show that our new method can be taken as a tool for variable importance measure.

In the second experiment, we partition the data for 5-fold cross validation, and remove the least important 20 percent variables in each iterative. As [11], we use stratified *k*-fold cross-validation, that is, a data set is randomly partitioned into *k* equally sized folds such that the class distribution in each fold is approximately the same as that in the entire data set. In contrast, regular cross-validation randomly partitions the data set into *k*-folds without considering class distributions. Kohavi reported that stratified cross-validation has smaller bias and variance than regular cross-validation [24]. To obtain a more reliable estimate, the stratified 5-fold cross-validation process was repeated 20 times using different partitions of the data. The parameters of random forest are ntree=2000, mtry=$\sqrt{I}$ .

From Table 2, we can see that both of the gene selection procedures yield very small sets of genes while preserving predictive accuracy. Our new method has slightly better predictive performance on most data sets.

The two experiments demonstrate the robustness of new variable importance measure is slightly better than that of original OOB measure.

From Table 2, we can see that both of the gene selection procedures yield very small sets of genes while preserving predictive accuracy. Our new method has slightly better predictive performance on mostly data set. The two experiments demonstrate the robustness of new variable importance measure is slightly better than that of original OOB measure.

**Table 2.** Compare of predictive performance and selected smallest gene set using the two types of variable importance in random forest.

| **Data set** | **Prox_VI** | | **OOB_VI** | |
|---|---|---|---|---|
| | Accuracy and standard deviations (% ) | #Genes | Accuracy and standard deviations (% ) | #Genes |
| Colon | 0.883 ± 0.021 | 3 | 0.875 ± 0.022 | 7 |
| Leukemia | 0.987 ± 0.009 | 14 | 0.983 ± 0.011 | 14 |
| DLBCL | 0.946 ± 0.021 | 21 | 0.951 ± 0.012 | 26 |
| Prostate Tumor | 0.967 ± 0.013 | 17 | 0.953 ± 0.016 | 17 |
| Leukemia1 | 0.981 ± 0.006 | 21 | 0.997 ± 0.007 | 26 |
| SRBCT | 0.997 ± 0.003 | 42 | 0.996 ± 0.004 | 42 |
| Brain_Tumor1 | 0.936 ± 0.007 | 86 | 0.933 ± 0.010 | 44 |
| 9_Tumors | 0.880 ± 0.037 | 130 | 0.886 ± 0.026 | 162 |

"# Genes": number of genes selected on the original data set.

## 5. Conclusion

In this study, we proposed a new variable importance measure based on the difference of proximity matrix. The measure is steady to random forest model and sensitive to the change of variables. Based on the new measure we introduced a gene selection algorithm, which shows promising performance compared with other gene selection algorithms. The experiment results clearly indicate that the proposed method can be profitably used with microarray data, and the new variable importance measure can be taken as a supplement for random forest.

## 6. Conflict of interest

The authors declare that there is no actual or potential conflict of interest including any financial, personal or other relationships with other people or organizations within that could inappropriately influence this work.

## 7. Acknowledgements

This research was supported by a grant from Natural Science Foundation of Fujian Province of China (No. 2009J05153). The authors would like to thank all those who took part in this study.

## 8. References

[1] Isabelle Guyon, Jason Weston, Stephen Barnhill, Vladimir Vapnik, et al., "Gene selection for cancer classification using support vector machines", Machine Learning, vol.46, pp.389–422, 2002.

[2] Lee JW, Lee JB, Park M, Song SH., "An extensive evaluation of recent classification tools applied to microarray data", Computation Statistics and Data Analysis, vol.48, pp.869-885, 2005.

[3] Yeung KY, Bumgarner RE, Raftery AE, "Bayesian model averaging: development of an improved multi-class gene selection and classification tool for microarray data", Bioinformatics, vol.21, pp.2394-2402, 2005.

[4] Jirapech-Umpai T, Aitken S, "Feature selection and classification for microarray data analysis: Evolutionary methods for identifying predictive genes". BMC Bioinformatics, vol.6, pp.148, 2005.

[5] Qiu Yihui, Mi Hong. "Application of Feature Extraction method in customer churn prediction based on Random Forest and Transduction", Journal of Convergenence Information Technology, Volume 5, Number 3 ,
pp73-78, 2010.

[6] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, et al, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring", Science, vol.286, pp.531–537, 1999.

[7] Zhengjun Cheng Yuntao Zhang, "Classification Models of Estrogen Receptor-B Ligands Based on PSO-Adaboost-SVM", Journal of Convergenence Information Technology, Volume 5, Number 2 , pp67-83, 2010.

[8] Sridhar Ramaswamy, Pablo Tamayo, Ryan Rifkin, Sayan Mukherjee, Chen-Hsiang Yeang, Michael Angelo, et al. "Multiclass cancer diagnosis using tumor gene expression signatures", In Proceedings of the National Academy of Sciences of the United States of America, vol.98, pp.15149-15154, 2001.

[9] Chai, H. and Domeniconi, C., "An evaluation of gene selection methods for multi-class microarray data classification", In Proceedings of the Second European Workshop on Data Mining and Text Mining in Bioinformatics, pp.3-10, 2004.

[10] Xin Zhou and David P. "Tuck MSVM-RFE: extensions of SVM-RFE for multiclass gene selection on DNA microarray data", Bioinformatics, vol.23, pp.1106–1114, 2007.

[11] Pang H, Lin A, Holford M, Enerson BE, Lu B, Lawton MP, Floyd E, Zhao H, "Pathway analysis using random forests classification and regression". BMC Bioinformatics, vol.22, pp.2028-2036, 2006.
[12] Ramón Díaz-Uriarte, Sara Alvarez de Andrés, "Gene selection and classification of microarray data using random forest", BMC Bioinformatics, vol.7, pp.1471-2105, 2006.
[13] Breiman L, "Random forests", Machine Learning, vol.45, pp 5-32, 2001.
[14] Breiman L, "Bagging predictors", Machine Learning, vol.24, pp.123-140, 1996.
[15] X. Huang, W. Pan, S. Grindle, X. Han, Y. Chen, S.J. Park, et.al. "A comparative study of discriminating human heart failure etiology using gene expression profiles", BMC Bioinformatics,vol.6, 2005.
[16] B.L. Wu, T. Abbott, D. Fishman, W. McMurray, G. Mor, K. Stone, et.al. "Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data", BMC Bioinformatics vol.19, pp.1636–1643,2003.
[17] Freund, Y. and Schapire, R., "Experiments with a new boosting algorithm", in Machine learning: Proceedings of the thirteenth international conference, pp.148–156,1996.
[18] T.G. Dietterich, "An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting and randomization", Machine Learning, vol.40, pp.1–19, 2000.
[19] Hastie, T., Tibshirani, R., Friedman, J., "The Elements of Statistical Learning: Data Mining, Inference, and Prediction", Springer-Verlag ,New York, 2001.
[20] Svetnik V, Liaw A, Tong C, Wang T: Application of Breiman's random forest to modeling structure-activity relationships of pharmaceutical molecules, in: Multiple Classier Systems: Proceedings of the fifth international workshop, 2004, MCS, Cagliari, Italy. Lecture Notes in Computer Science, Springer, vol.3077, pp.334-343, 2004.
[21] Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, Levine AJ, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays", In Proceedings of the National Academy of Sciences of the United States of America, vol.96, pp.6745-6750, 1999.
[22] Showe Laboratory, Available:http://showelab.wistar.upenn.edu.
[23] A. Statnikov, I. Tsamardinos, Y. Dosbayev, C.F. Aliferis, "GEMS: A System for Automated Cancer Diagnosis and Biomarker Discovery from Microarray Gene Expression Data", International Journal of Medical Informatics, vol.74, pp.491-503, 2005.
[24] Kohavi, R., "A study of cross-validation and bootstrap for accuracy estimation and model selection". In Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, Morgan X. Zhou and D P. TuckKaufmann Publishers, San Francisco, pp. 1137–1145, 1995.