

A Post-filtering Gene Selection Algorithm Based on Redundancy and Multi-gene Analysis

Liangliang Wang, Jinwen Ma

Department of Information Science, School of Mathematical Sciences and LMAM, Peking University, Beijing, 100871, China

jwma@math.pku.edu.cn

Abstract

This paper proposes a post-filtering gene selection algorithm to discover informative genes of a tumor. With the identified genes via some conventional statistical test methods on the microarray data, the proposed algorithm utilizes a correlation measure and an approximate Markov blanket from the statistical tests to reduce redundant ones. It further evaluates the goodness of classification from the point of view of multi-gene action instead of individual gene action. To test the effectiveness of the post-filtering gene selection algorithm, we use the support vector machine (SVM) to construct a tumor diagnosis system, (i.e., a binary classifier) based on the identified informative genes. Compared with the conventional methods, the constructed diagnosis system with the post-filtering gene selection algorithm can reach higher diagnosis accuracy on both the colon and leukemia data, with a smaller number of informative genes.

Keyword: Gene expression profile, gene selection, redundancy, informative gene, correlation measure, multi-genes, unsupervised clustering, support vector machine (SVM)

I. Introduction

The microarray technology can now be used to rapidly measure the levels of thousands of genes expressed in a biological sample (or tissue) through the process of hybridization. The analysis of these microarray data is significant to some fundamental problems in biology as well as in clinical medicine. Actually, it can improve the understanding of the diseases at a molecular level and to develop some new diagnostic methods.

Currently, microarray data or gene expression profiles have been widely used in many applications, especially on tumor diagnosis. Given a set of samples labeled “tumorous” or “normal”, the task of tumor diagnosis is just to build a binary classifier as a diagnosis system to predict the unlabelled samples. Mathematically, the microarray data can be represented by a matrix $A = (a_{ij})_{n \times m}$, where the i -th row represents the i -th gene, the j -th column represents the j -th sample, and the element a_{ij} represents the expression level of the i -th gene at the j -th sample. In comparison with the number of genes per sample, we can only collect a small number of samples because of the high expense. However, such high dimensional data could cause a series of problems, such as high computing complexity, low prediction accuracy and unexplainable biological meaning (e.g., [1]). Therefore, informative gene selection is often used as a preprocessing technique in tumor classification.

Informative gene selection, i.e., finding the genes that are discriminative among different phenotypes, has been investigated extensively in the past several years. Typically, informative genes are selected by ranking genes according to a certain criterion, such as t , F and rank sum test

statistics [2][3][4]. Usually, these conventional methods select the top k genes. But it has been shown that the genes selected by individual gene evaluations are often highly correlated. To maintain a high prediction accuracy of the classifier, we must find uncorrelated but still highly informative genes. Consequently, the Markov blanket was introduced to reduce the redundant genes from the selected ones [1][5][6]. Since the Markov blanket is purely theoretical, it was implemented by some approximates. However, these methods are sensitive to the order of genes according to their individual ranks. That is, these top genes have greater influence than other genes. As a result, it may eliminate too many genes and thus some useful information may be lost. Except individually selected genes, valuable information can also be discovered by evaluating the classification of combinations of genes [7].

To better tackle the problems mentioned above, we propose a post-filtering gene selection algorithm to select the informative genes of a tumor via a microarray data set by combining the multi-gene analysis with individual gene selection. Since non-parametric statistical tests have been shown to be efficient on informative gene selection [2][3][4], we will use statistical tests to select the genes for the redundancy analysis. The post-filtering gene selection process will reduce redundant genes by analyzing the correlation among the selected genes as well as their approximate Markov blankets based on the statistical tests. Moreover, the multi-gene analysis is conducted to keep the genes which play an important role on the classification through the whole group. As a result, the post-filtering gene selection algorithm can remove redundant genes and maintain a powerful set of informative genes for the classification. It is demonstrated by the experiments that the post-filtering gene selection algorithm can lead to a high prediction accuracy with a small number of informative genes on both the colon and leukemia data sets.

In the sequel, we propose the post-filtering gene selection algorithm in Section II. It is then experimented and compared in Section III. We further give some discussions and remarks in Section IV. Finally, we conclude briefly in Section IV.

II. Post-filtering Gene Selection Algorithm

We begin with introducing some definitions and statistical test methods that will be used in our algorithm. For a pair of two matrices or vectors X, Y of the same size, we compute their correlation coefficient r by:

$$r = \frac{\sum_i \sum_j (x_{ij} - \bar{X})(y_{ij} - \bar{Y})}{\sqrt{\sum_i \sum_j (x_{ij} - \bar{X})^2} \sqrt{\sum_i \sum_j (y_{ij} - \bar{Y})^2}} \quad (1)$$

where \bar{X} is the mean of all values in X , while \bar{Y} is the mean of all values in Y . Clearly, the correlation coefficient r lies between -1 and 1. If X and Y are completely correlated, r takes either 1 or -1; if X and Y are totally independent, r becomes zero.

The 2-sample Hotelling T^2 test makes the hypothesis testing on whether two populations have the same mean through the two samples. In other words, it checks the following hypothesis with the two samples: $H_0 : \mu_1 = \mu_2$, $H_a : \mu_1 \neq \mu_2$, where, μ_1 and μ_2 are the two means of two populations, respectively. For clarity, we give the procedure of the Hotelling T^2 test as follows.

1. We begin with the multivariate statistical testing for the homogeneity of covariance matrices by the Box's M statistic. In fact, this testing result will make the following Hotelling T^2 test different.
2. If the two covariance matrices are the same according to the above testing result, the 2-sample Hotelling T^2 test statistic is given by:

$$T^2 = \frac{n_1 n_2}{n_1 + n_2} (\bar{X}_1 - \bar{X}_2)' S^{-1} (\bar{X}_1 - \bar{X}_2), \quad (2)$$

where \bar{X}_1 and \bar{X}_2 are the two sample means, n_1 and n_2 are the sizes of the two samples, and $S = (S_1 + S_2)/(n_1 + n_2 - 2)$, where S_1 and S_2 are the two sample covariance matrices.

3. If the two covariance matrices are not the same according to the above testing result, the 2-sample Hotelling T^2 statistic is given by:

$$T^2 = (\bar{X}_1 - \bar{X}_2)' \left(\frac{S_1}{n_1(n_1 - 1)} + \frac{S_2}{n_2(n_2 - 1)} \right)^{-1} (\bar{X}_1 - \bar{X}_2). \quad (3)$$

In fact, as the sample size is large, the Hotelling's T^2 statistic approximately subjects to a Chi-square distribution; otherwise, it approximately subjects to an F distribution [8].

In statistics, the p value of a statistical test provides a measure of the strength of evidence that the sample is in favor of the null hypothesis. It is the probability of getting a result as extreme or more extreme than the one observed if the proposed null hypothesis is correct. A p value closing to zero means that the null hypothesis is false, and a difference is very likely to exist between the two populations. On the other hand, a large p closing to 1 implies that there is no detectable difference through the sample used. By the definition of p value, we can use it to measure the goodness of classification on a gene set. Actually, our informative gene selection method uses a significance level α on the statistical tests. That is, the genes are selected as the informative ones if the p values of the statistical tests are less than α .

With above preparations, we can now present the post-filtering gene selection algorithm. By conducting certain statistical tests on the gene expression profiles, we assume that a gene set with a smaller α value contains more information about the class than a gene set with a larger p value. We introduce a kind of approximate Markov blanket based on statistical tests as follows:

Definition 1 (Approximate Markov blanket) For two gene sets GS_i and GS_j ($i \neq j$), GS_i forms an approximate Markov blanket for GS_j ($i \neq j$) iff $p(GS_i) < p(GS_j)$ and $p(GS_i) < p(GS_i \cup GS_j)$.

Here $p(GS)$ is the p value of a 2-sample statistical test on the homogeneity of two types of the expression profiles throughout the gene set GS . If the gene set contains only one gene, we take a 2-sample univariate statistical test; otherwise, we take a 2-sample multivariate statistical test. The null hypothesis is that the two populations of the tumor and normal samples are identical in statistics. Our aim is to check whether the expression profiles on a gene set are different between the normal and tumor samples. If GS_i forms an approximate Markov blanket for GS_j , we consider that GS_i can provide more information about the classification than the combination of GS_i and GS_j .

We summarize the post-filtering gene selection algorithm in Table 1. There are three predefined parameters. α_0 is the significance level of the statistical test to determine the genes that will be considered by the post-filtering algorithm. The larger α_0 is, the more genes will be processed. α is also a significance level of the statistical tests which restricts the goodness of classification through the gene set. β is a threshold value of the correlation measure. When β is too small, we will remove more genes at risk of losing some useful information. On the contrary, if β is too large, our approximate Markov Blanket based on the statistical tests is not powerful enough to remove sufficiently redundant genes. Usually, we set β in $[0.5, 0.8]$. $UTest$ in the algorithm is a kind of univariate statistical test, and $MTest$ is a kind of multivariate statistical test.

Table 1: The post-filtering gene selection algorithm

1.	Input : $G, L, \alpha_0, \alpha, \beta$
2.	Output : GS_{best}
3.	Order the gene sets:
4.	if $s == 1$
5.	Calculate every $P(G_i) = UTest(G_i)$ for all $G_i \in G$
6.	if $(P(G_i) < \alpha_0)$, append G_i to G_{list} ;
7.	Order G_{list} in ascending p values;
8.	if $s > 1$ and $s \leq maxs$
9.	$G_{list} = NULL$;
10.	begin
11.	Calculate $\min P(G_i; G_j) = \min(Mtest(G_i; G_j))$ for all $G_i, G_j \in RG_{list}$
12.	remove G_i and G_j from RG_{list} , append $G_{i,j} = G_i \cup G_j$ to G_{list} ;
13.	end until($RG_{list} == NULL$);
14.	Redundancy analysis :
15.	$G_j = getFirstElement(G_{list})$;
16.	if $P(G_j) > \alpha$, append G_j to RG_{list}
17.	else
18.	begin
19.	$G_k = getNextElement(G_{list}, G_j)$;
20.	begin
21.	if $(C(G_j; G_k) > \beta)$, remove G_k from G_{list}
22.	elseif $(P_j < P_{j,k})$, remove G_k from G_{list} , append G_k to RG_{list} ;
23.	$G_k = getNextElement(G_{list}, G_j)$
24.	end until($G_k == NULL$)
25.	$G_j = getNextElement(G_{list}, G_j)$
26.	end until($G_j == NULL$);
27.	Add G_{list} to GS_{best} ;
28.	$s = s + 1$, go to order the gene sets.

The post-filtering gene selection algorithm involves $maxs$ iterations ($s = 1, \dots, maxs$). Every iteration is composed of two sequential steps. 1. Order the gene sets: for the first iteration, this step calculates the p values of a univariate statistical test (line 5) for every gene from G , selects genes whose p values are less than α_0 , and appends these genes to G_{list} , a list of gene sets, in an ascending order of their p values. In the following iterations, this step calculates a multivariate statistical test (line 11) for every pair of gene sets from RG_{list} , a list of gene sets that are temporarily removed during the previous iteration. Then, this step removes the pair with minimum p value from RG_{list} and appends them as one element to G_{list} . The process repeats until RG_{list} is empty. 2. Redundancy analysis by evaluating classification of a combination of 2^{s-1} genes: the filtering is based on a predefined threshold β of the correlation measure (line 21) and the approximate Markov blanket. Through this step, every newly selected gene set satisfies three criteria: 1). The p value of the statistical test is less than α . 2). It is not highly correlated with other gene sets selected already. 3). It does not have an approximate Markov blanket in the selected gene sets. Those genes in the sets satisfying all the three criteria will be added into GS_{best} as the result of the s -th iteration. Those gene sets violating criterion 1 or 3 will be added into RG_{list} and be processed in the $(s + 1)$ -th iteration which will evaluate larger gene sets by combining two smaller ones from RG_{list} , and select those gene sets satisfying the same three criteria. We can stop the algorithm when no more informative

genes can be found in an iteration. Alternatively, for simplicity, the algorithm stops when it reaches the user-specified maximum number of iterations.

III. Experimental Results and Comparisons

We tested the effectiveness of our post-filtering gene selection algorithm on the two real data sets through the support vector machine as follows.

Colon cancer data set: The colon cancer data set¹ contains the expression profiles of 2000 genes in 22 normal tissues and 40 tumor tissues.

Leukemia cancer data set: The leukemia cancer data set² consists of expression profiles of 7129 genes from 47 acute lymphoblastic leukemia (ALL) and 25 acute myeloid leukemia (AML) samples.

We first implemented the same preprocessing steps as those proposed in [9]: thresholding with a floor of 100 and a ceiling of 16000; and filtering by excluding genes with $max/min \leq 5$ and $max - min \leq 500$, where max and min refer to the maximum and minimum intensities for a particular gene across the 72 mRNA samples. This preprocessing left us with a data set of 3571 genes.

Before our experiments, we normalized the gene expression profiles with zero mean and unit variance in order to eliminate the possible noise. For a tumor diagnosis system, i.e., a binary classifier, through the post-filtering algorithm, we took its prediction accuracy as the evaluation metric of the algorithm. In our experiments, we used the radial basis function (RBF) as the kernel to build a SVM via a MATLAB toolbox called OSU SVM 3.0³. There were two parameters γ and C to be selected. In our experiments, we took a grid search procedure from 16×16 pairs of γ and C ($\gamma, C = 2^{-7}, 2^{-6}, \dots, 2^8$), and chose the values optimizing the performance of SVM on the training data set [10].

In our experiments, we always used the correlation coefficient as the correlation measure and selected the 2-sample Hotelling T^2 test as the multivariate statistical test in the post-filtering algorithm. However, we used several univariate statistical tests and different parameters which were indicated in Table 2-5 to conduct the 3 iterations of the algorithm. The prediction results of the different parameters and methods were shown in these tables, respectively, where the two numbers on the sides of “/” represented the prediction accuracy on the test sets and the number of selected informative genes. To compare with the conventional methods, we also listed the prediction accuracy of genes selected by the conventional methods without further filtering ($s = 0$).

Table 2. The leave-one-out cross-validation (LOOCV) results of the post-filtering gene selection algorithm on the colon data set with the rank sum test. Here and in the following, M & P means “Methods and Parameters”.

M & P	α	0.001	0.005	0.01	0.05
$\beta = 0.6$ <i>Rank sum</i>	$s = 0$	90.3%/60	91.9%/137	91.9%/188	88.7%/389
	$s = 1$	91.9%/7	91.9%/10	91.9%/10	91.9%/13
	$s = 2$	93.5%/10	95.2%/19	95.2%/22	95.2%/34
	$s = 3$	93.5%/10	96.8%/23	98.4%/28	95.2%/49

We conducted the leave-one-out cross-validation (LOOCV) [11] experiments on the colon data set. The classifier was successively trained on $n - 1$ samples and tested on the remaining one. We applied gene selection in each cross-validation trial on the training examples of that trial. The construction of a classifier was restricted to the selected informative genes using the training data. Finally, we computed the average prediction accuracy and the number of informative genes for the

¹ retrieved from <http://microarray.princeton.edu/oncology/database.html>

² retrieved from <http://www-genome.wi.mit.edu/cgi-bin/cancer/datasets.cgi>

³ available on http://eewww.eng.ohio-state.edu/~maj/osu_svm

n results as our evaluation result. In this experiment, we set $\alpha_0 = \alpha$ and used the 2-sample rank sum test. From Table 2, the best prediction accuracy achieved by using the post-filtering algorithm was 98.4%, where only one prediction error happened in the LOOCV experiment.

In order to illustrate the influence of the correlation coefficient threshold β under the same significance level, we gave the classification accuracies with different correlation coefficient thresholds. To utilize all the information of a microarray data set, we selected genes on the whole data set before conducting LOOCV in this and the following experiments. We conducted LOOCV experiments using 2-sample t tests after selecting genes on the whole colon data set. From Table 3, we can find out that our algorithm achieved better prediction accuracies when β is 0.5 or 0.6 on the colon data set.

Table 3. The influence of the threshold β on the prediction accuracy under the same significance level

M & P	β	0.3	0.4	0.5	0.6	0.7	0.8
$\alpha_0 = 0.03$ $\alpha = 0.03/315$ 2-sample t test	1	85.4%/5	91.9%/4	90.3%/8	95.2%/13	93.5%/23	93.5%/62
	2	88.7%/7	95.2%/10	96.8%/18	96.8%/27	93.5%/41	93.5%/94
	3	90.3%/11	95.2%/10	96.8%/18	96.8%/31	95.2%/53	93.5%/102

When we set $\alpha_0 = 0.015$ and used the 2-sample Kolmogorov-Smirnov (KS) test to do the conventional gene selection on the colon data set, 177 genes were selected and the corresponding prediction accuracy is 93.5%. These genes were to be further filtered by the post-filtering algorithm under several different significance levels. From the results shown in Table 4, the post-filtering algorithm can achieve better prediction accuracies by removing redundant genes. As a matter of fact, all the best prediction accuracies reached 98.4% under these significance levels shown in Table 4.

Table 4. The LOOCV results on the colon data set with the KS test.

M & P	α	0.001	0.003	0.005	0.007	0.009
$\alpha_0 = 0.015$ $\beta = 0.6$	$s = 1$	93.5%/5	93.5%/6	93.5%/6	95.2%/7	95.2%/7
	$s = 2$	96.8%/19	95.2%/20	98.4%/22	98.4%/23	98.4%/23
KS test	$s = 3$	98.4%/27	98.4%/24	98.4%/26	96.8%/27	96.8%/27

α_0 set a significance level which was only used in the first iteration of the algorithm. We fixed $\alpha = 0.005$ and used the 2-sample Kolmogorov-Smirnov test to do the experiment on the colon data set. We gave the prediction accuracies over the different α_0 in Fig.1. We calculated the average and the maximum prediction accuracies of 3 iterations of the post-filtering algorithm. Meanwhile, Figure 1 also illustrated the prediction accuracies before filtering. By comparison, best prediction accuracies were achieved when α_0 was 3-5 times of α on the colon data set, which was shown in Fig. 1.

We also conducted an experiment on the leukemia data set. Because of the large scale, we took the 4-fold cross validation [4] instead of LOOCV on this data set. From Table 5, we can find that the post-filtering algorithm reached a perfect result. The two classes can be totally separated with only 9 genes.

Fig. 1. The LOOCV results with different α_0 on the colon data set.

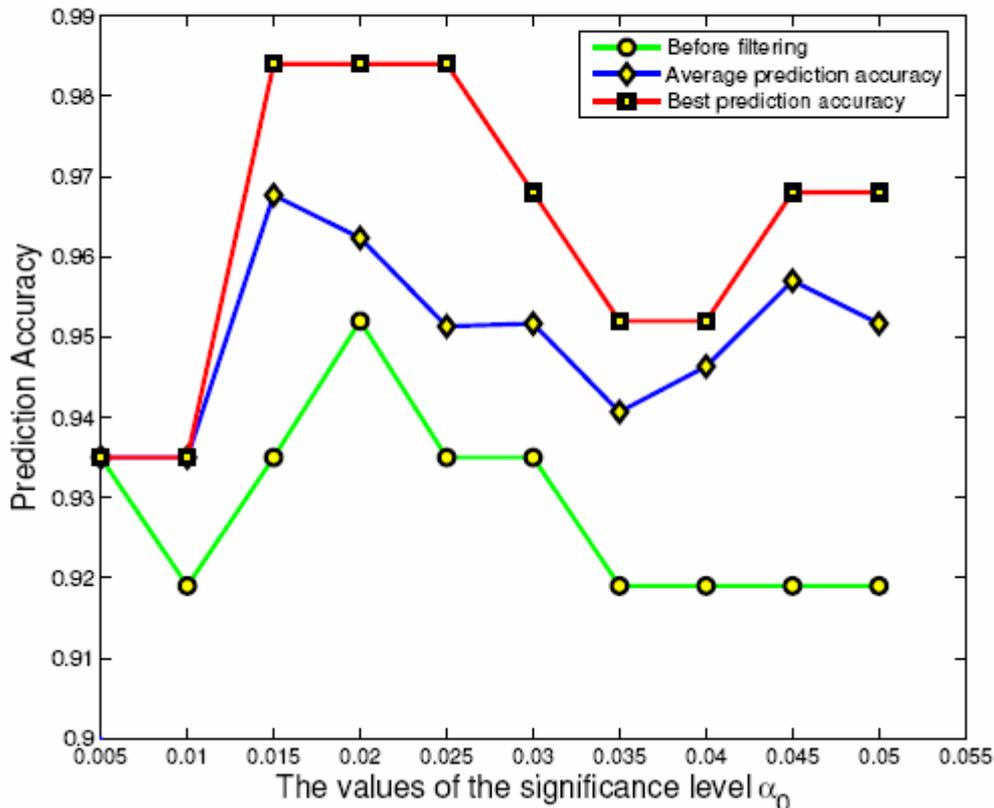


Table 5. The 4-fold cross-validation results on the leukemia data set with the rank-sum test.

M & P	α	0.0000001	0.000001	0.00001	0.0001	0.001
$\beta = 0.5$ Rank sum	$s = 0$	98.6%/53	98.6%/93	98.6%/163	95.8%/293	77.8%/516
	$s = 1$	98.6%/7	100%/9	100%/13	100%/19	100%/21
	$s = 2$	100%/9	100%/11	100%/21	100%/49	98.6%/79
	$s = 3$	100%/9	100%/11	100%/21	100%/57	98.6%/107

IV. Further Discussions and Remarks

From the above experiments, we can find that the post-filtering gene selection algorithm can lead to a better prediction accuracy on the two data sets using a smaller number of genes than the conventional methods. More iterations of the algorithm may reveal more information by evaluating the goodness of classification by multi-gene analysis. Meanwhile, we notice that the prediction accuracy of the RBF SVM through the post-filtering algorithm tends to decrease with the increasing number of genes. The reason is that the high-dimensional data may cause the problem of overfitting. Consequently, sometimes the prediction accuracy does not increase by adding more informative genes. However, in most cases, informative genes found by the post-filter algorithm can increase the prediction accuracy.

In general, we can select 0.005 or 0.01 as a good choice of the significance level α for the post-filtering algorithm. On the one hand, if α is relatively large, we tend to select genes with good classification by individual evaluations. On the other hand, if α is too small, we tend to select genes with good classification by multi-gene analysis. However, too many genes will affect the

generalization of the constructed classifier. As a result, we should seek a balance between the two aspects at selecting α . From the results of the experiments, 0.005 or 0.01 as the significance level α works rather good on both the colon and leukemia data sets. As for the selection of α_0 , we should balance the computing complexity and the required information of classification. Generally, it should be 3-5 times of α .

By using statistical tests, our post-filtering method for informative gene selection is not merely a kind of algorithm. It can be considered as a basic framework. Actually, many other statistical tests and methods can be also applied to this approach. To form an approximate Markov Blanket, we can use any statistical test with a null hypothesis that the two populations have no differences. Actually, all the 2-sample statistical tests used in the experiments, including t test, Kolmogorov-Smirnov test and rank sum test [4][12], can reach a better prediction accuracy through this post-filtering process. Recent studies have shown that the microarray data set does not follow the normal distribution and non-parametric statistical methods often achieve the better result, which is also demonstrated in our experiments. However, since the Hotelling T^2 are based on the normal distribution, we could get a better result by using more appropriate statistical tests for the microarray data set in our further study. In addition, we can seek a more powerful correlation measure to reveal more information between two gene sets.

V. Conclusions

Informative gene selection is a critical problem in the microarray data analysis and the tumor diagnosis. We have investigated this problem and proposed a post-filtering gene selection algorithm in this paper. The proposed algorithm uses a correlation measure and an approximate Markov blanket on the statistical tests to reduce redundant genes. Many statistical tests can be applied to the algorithm. Moreover, it combines the individual gene selection with multi-gene analysis. By the experiments conducted on the colon and leukemia data through the SVM for the tumor classification or diagnosis, it has been shown that this post-filtering gene selection algorithm can lead to a higher prediction accuracy on the two data sets with a smaller number of informative genes than the conventional methods.

Acknowledgments This work was supported by the Natural Science Foundation of China for project 60471054.

References

- [1] L. Yu and H. Liu, "Redundancy based feature selection for microarray data," *Proceedings of the Tenth ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2004, pp. 737-742.
- [2] T. R. Golub, D. K. Slonim, P. Tamayo, et al., "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, vol.286, pp. 531-537, 1999.
- [3] C. Ding, "Analysis of gene expression profiles: class discovery and leaf ordering," *Proceedings of the 6th Annual International Conference on Computational Molecular Biology (RECOMB'02)*, Washington, D. C., USA, April 18-21, 2002, pp. 601-680.
- [4] L. Deng, J. Ma, and J. Pei, "Rank sum method for related gene selection and its application to tumor diagnosis," *Chinese Science Bulletin*, vol.49, no.15, pp. 1652-1657, 2004.
- [5] D. Koller and M. Sahami, "Toward optimal feature selection," *Proceedings of the Thirteenth International Conference on Machine Learning*, 1996, pp. 284-292.

- [6] E. P. Xing, M. I. Jordan, and R. M. Karp, "Feature selection for high-dimensional genomic microarray data," *Proceedings of the 18th International Conference of Machine Learning (ICML'01)*, Massachusetts, USA, June 28-July 1, 2001, pp. 601-608.
- [7] T. Bo and I. Jonassen, "New feature subset selection procedures for classification of expression profiles," *Genome Biol.*, vol.3, pp. RESEARCH0017.1--0017.11, 2002.
- [8] K. Fang, "Applied Multivariate Statistical Analysis" (in Chinese), Shanghai: East China Normal University Press, 1989, pp. 105-141.
- [9] D. S. Dudoit, J. Fridyand, and T. P. Speed, "Comparison of discrimination methods for the classification of tumor using gene expression data," *Univ. of California, Dept. of Statistics, Tech Report*, no.576, 2000.
- [10] C. Hsu, C. Chang, and C. Lin, "A Practical Guide to Support Vector Classification," *National Taiwan University, Dept. of Computer Science and Information Engineering* posted in the web: www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf.
- [11] A. Ben-Dor, N. Friedman, and Z. Yakhini, "Scoring Genes for Relevance," *Agilent Technical Report*, no. AGL-2000-13, 2000.
- [12] M. Hollander and D. A. Wolfe, "Nonparametric statistical method," New York: Wiley, 1999.



Liangliang Wang received the Bachelor of Engineering degree at Department of Computer Science of Zhengzhou University in 2002. She received the Master of Engineering degree at Department of Information Science, School of Mathematical Sciences of Peking University in 2005. Recently, she is studying for a PhD degree at Department of Epidemiology and Biostatistics of McGill University, Canada. Currently her main research interests include gene selection and Bayesian approach for design and analysis of studies involving non-gold standard diagnostic tests.



Jinwen Ma received the Master of Science degree in applied mathematics from Xi'an Jiaotong University in 1988 and the Ph. D. degree in probability theory and statistics from Nankai University in 1992. From July 1992 to November 1999, he was a Lecturer or Associate professor at Department of Mathematics, Shantou University. He has been a full professor at Institute of Mathematic, Shantou University since December, 1999. In September 2001, has was transferred to the Department of Information Science at the School of Mathematical Sciences, Peking University. During 1995 and 2003, he also visited several times at Department of Computer Science & Engineering, the Chinese University of Hong Kong as a Research Associate or Fellow. He has published over 70 academic papers on neural networks, pattern recognition, bioinformatics, and information theory.