Significance of Gene Ranking for Classification of Microarray Samples

Chaolin Zhang, Xuesong Lu, and Xuegong Zhang

Abstract—Many methods for classification and gene selection with microarray data have been developed. These methods usually give a ranking of genes. Evaluating the statistical significance of the gene ranking is important for understanding the results and for further biological investigations, but this question has not been well addressed for machine learning methods in existing works. Here, we address this problem by formulating it in the framework of hypothesis testing and propose a solution based on resampling. The proposed r-test methods convert gene ranking results into position p-values to evaluate the significance of genes. The methods are tested on three real microarray data sets and three simulation data sets with support vector machines as the method of classification and gene selection. The obtained position p-values help to determine the number of genes to be selected and enable scientists to analyze selection results by sophisticated multivariate methods under the same statistical inference paradigm as for simple hypothesis testing methods.

Index Terms—Significance of gene ranking, gene selection, classification, microarray data analysis.

1 Introduction

An important application of DNA microarray technologies in functional genomics is to classify samples according to their gene expression profiles, e.g., to classify cancer versus normal samples or to classify different types or subtypes of cancer. Selecting genes that are informative for the classification is one key issue for understanding the biology behind the classification and an important step toward discovering those genes responsible for the distinction. For this purpose, researchers have applied a number of test statistics or discriminant criteria to find genes that are differentially expressed between the investigated classes [1], [2], [3], [4], [5], [6], [7]. This category of gene selection methods is usually referred to as the filtering method since the gene selection step usually plays the role of filtering the genes before doing classification with some other methods. Another category of methods is the so-called wrapper methods, which use the classification performance itself as the criterion for selecting the genes and genes are usually selected in a recursive fashion [8], [9], [10], [11], [12]. A representative method of this category is SVM-RFE based on support vector machines (SVM), which uses linear SVM to classify the samples and ranks the contribution of the genes in the classifier by their squared weights [10].

All these selection methods produce rankings of the genes. When a test statistic, such as the t-test, F-test, or bootstrap test, is used as the criterion, the ranking is attached by p-values derived from the null distribution of the test statistic, which reflects the probability of a gene showing the observed difference between the classes simply due to chance. Such p-values give biologists a clear understanding of the information that the genes probably contain. The availability of the p-value makes it possible to investigate the microarray data under the solid framework of statistical inference and many theoretical works have been built based on the extension of the concept of p-value, such as the false discovery rate (FDR) study [13].

Existing gene selection methods that come with p-values are of the filtering category and are all univariate methods. To consider possible combinatorial effects of genes, most wrapper methods adopt more sophisticated multivariate machine learning strategies such as SVMs and neural networks. These have been shown in many experiments to be more powerful in terms of classification accuracy. However, for gene selection, the gene rankings produced with these methods do not come with a measure of statistical significance. The ranking is only a relative order of genes according to their relevance to the classifier. There is no clear evaluation of a gene's contribution to the classification. For example, if a gene is ranked 50th according to its weight in the SVM classifier, it is only safe to say that this gene is perhaps more informative than the gene ranked at 51st. However, there is no way to describe how significant it is and there is no ground to compare the information it contains with a gene also ranked as 50th by the same method in another experiment. This nature of relative ranking makes it hard to interpret and further explore the gene selection results achieved with such advanced machine learning methods. For example, it is usually difficult to decide on the proper number of genes to be selected in a specific study with such machine learning methods. Most existing works usually select a subgroup of genes with some heuristically decided numbers or thresholds [6], [8], [10]. The advanced estimation techniques, such as FDR, based on significance measures do not apply for such methods.

- *C. Zhang is with the Cold Spring Harbor Laboratory and the Department of Biomedical Engineering, State University of New York at Stony Brook, NY 11794. E-mail: zhangc@cshl.edu.*
- *X. Lu and X. Zhang are with the MOE Key Laboratory of Bioinformatics/Department of Automation, Tsinghua University, Beijing 100084, China. E-mail: lxs97@mails.tsinghua.edu.cn, zhangxg@tsinghua.edu.cn.*

Manuscript received 28 Sept. 2004; revised 14 May 2005; accepted 5 Oct. 2005; published online 31 July 2006.
For information on obtaining reprints of this article, please send e-mail to: tcbb@computer.org, and reference IEEECS Log Number TCBB-0159-0904.

Evaluating the statistical significance of the detected signal is the central idea in the paradigm of statistical inference from experimental data. There should be an equivalent study on those machine-learning-based multivariate gene selection methods which produce ranks according to their own criteria. Strategies such as permutation can be utilized to assess the significance of the classification accuracy, but they do not measure the significance of the selected genes directly. Surprisingly, this question has not been addressed by the statistics or bioinformatics community in existing literature. We therefore propose that the question be asked in this way: For an observed ranking of genes by a certain method, what is the probability that a gene is ranked at or above the observed position due to chance (by the same method) if the gene is, in fact, not informative to the classification? ("Being informative" is in the sense of the criteria defined or implied by the classification and ranking method. It may have different meanings for different methods.) We call this problem the significance of gene ranking or feature ranking. We raise this problem in this paper and describe our strategy toward a solution. The problem is discussed in the context of microarray classification of cancer samples, but the philosophy and methodology is not restricted to this scenario.

2 The Significance-of-Ranking Problem

Suppose a microarray data set contains m cases $X = \{\mathbf{x}_i, i = 1, \cdots, m\}$. Each case is characterized by a vector of the expression values of n genes $\mathbf{x}_i = [x_{i1}, x_{i2}, \cdots, x_{in}]^T \in R^n, i = 1, \cdots, m$. Each gene is a vector of their expression values across the cases $\mathbf{g}_j = [x_{1j}, x_{2j}, \cdots, x_{mj}]^T$ and we denote the set of all genes as $G = \{\mathbf{g}_j,\ j = 1, \cdots, n\}$. Each case has a label $y_i = \{-1, 1\}$, $i = 1, \cdots, m$ indicating the class it belongs to among the studied two classes, e.g., normal versus cancer, or two subtypes of a cancer, etc. Among the n genes, usually some are informative to the classification and some are not, but we do not know which genes are informative and which are not. For the convenience of description, we denote the set of informative genes as I_G and that of the uninformative genes as U_G. To simplify the problem, we assume that

$$I_G \cap U_G = \Phi \text{ and } I_G \cup U_G = G. \tag{1}$$

The goal is to build a classifier that can predict the classes $\hat{y}_i$ of the cases from $\mathbf{x}_i$ and, at the same time, to identify the genes that most likely belong to I_G. The former task is called classification and the latter one is called gene selection. In the current study, we assume that there has already been a ranking method *RM* which produces a ranking position for each gene according to some criterion assessing the gene's relevance with the classification:

$$r_j = rank\big(g_j | \{(\mathbf{x}_i, y_i), i = 1, \cdots, m\}\big), j = 1, \cdots, n \tag{2}$$

and we do not distinguish the specific types of the *RM*. The ranking is obtained based on the samples, thus r_j is a random variable. The significance-of-ranking problem is to calculate the following probability:

$$p(r_j) \triangleq P\big(rank(g_j) \le r_j | g_j \in U_G\big), \tag{3}$$

i.e., given a gene is uninformative to the classification (according to *RM*'s criterion), what is the probability that it is ranked at or above the observed ranking position by the ranking method? We call this probability the p-value of a gene's ranking position or, simply, position p-value.

This significance-of-ranking problem is distinct from existing statistics for testing differentially expressed genes in several aspects. It applies to more complicated multivariate classification and gene selection methods. Even when it is applied on gene ranking methods based on univariate hypothesis tests like t-test, the position p-value is different with the t-test p-value by definition. The t-test p-value of a gene is calculated from the expression values of this gene in the two sample sets by comparing with the assumed null distribution model when the gene is not differentially expressed in the two classes. The position p-value of a single gene, however, is defined on its context, in the sense that its value depends not only on the expression of this gene in the samples, but also on other genes in the same data set. A gene with the same expression values may have different position p-values in different data sets. The null distributions of ranks of uninformative genes are different in different data sets and, therefore, the foremost challenge for solving the problem is that the null distribution has to be estimated from the specific data set under investigation.

3 The r-Test Scheme

The significance-of-ranking problem is formulated as a hypothesis testing problem. The null hypothesis is that the gene is not informative or $g_j \in U_G$, the alternative hypothesis is $g_j \in I_G$ (the gene is informative) since we have assumption (1) and the statistic to be used to test the hypothesis is the ranking position. As in standard hypothesis testing, the key to solving the problem is to obtain the distribution of the statistic under the null hypothesis, i.e., the distribution of the ranks of uninformative genes:

$$P(r | g \in U_G). \tag{4}$$

For the extreme case when $I_G = \Phi$ (all the genes are uninformative) and the ranking method is not biased, it is obvious that the null distribution is uniform. In a real microarray data set, however, usually some genes are informative and some are not, thus the uniform null distribution is not applicable. The null rank distribution in a practical investigation depends on many factors, including the separability of the two classes, the underlying number of informative genes, the power of the ranking method, the sample size, etc. The characteristics of these factors are not well understood in either statistics or biology and, therefore, we have to estimate an empirical null distribution from the data set itself.

We propose to tackle this problem in two steps. First, we identify a set of putative uninformative genes (PUGs) which are a subset of U_G. This is possible in practice because, although we do not know U_G, discovering a number of genes that are irrelevant to the classification is usually not hard in most microarray data sets. We denote the identified subset as

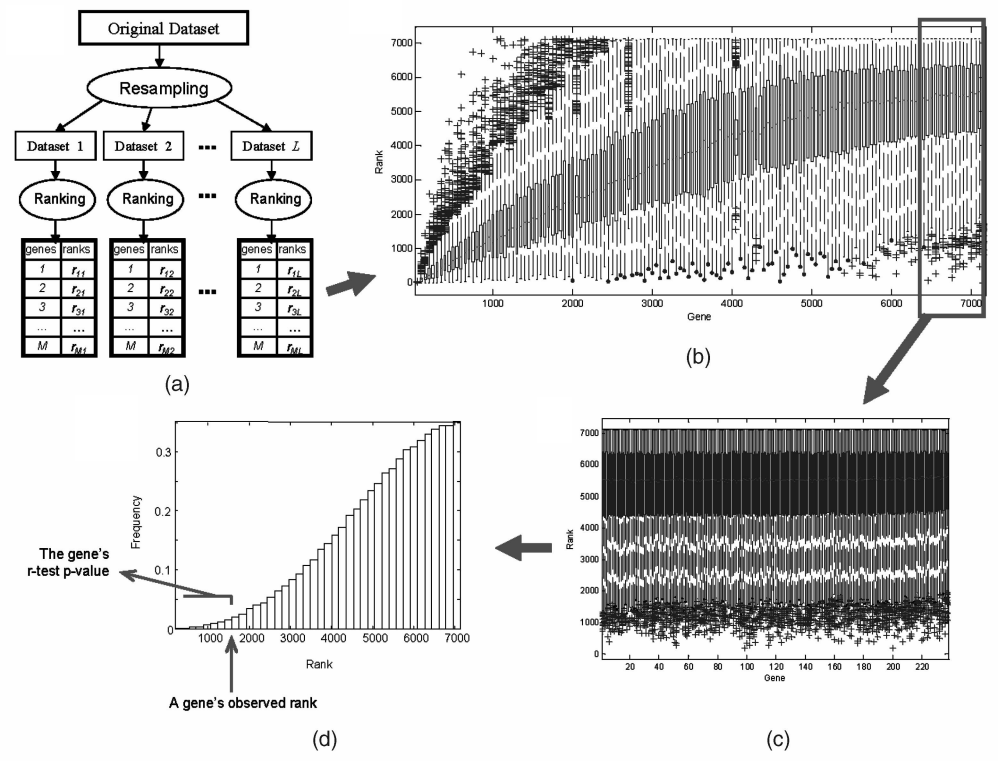

Fig. 1. The diagram showing the principle of r-test (pr-test). (a) A number (L) of new data sets is resampled from the original data set. A ranking is generated for each new data set with the ranking method, resulting in a total of L rankings. (b) The genes are ordered by their average positions of the L rankings. The horizontal axis is genes by this order and the vertical axis is ranking position in the resample experiments. For each gene, its ranking positions in the L experiments are drawn in a box plot, with a short dash in the middle showing the median. (c) From the bottom (rightmost) of the ordered gene list, k genes are selected as putative uninformative genes or PUGs. The box-plots of the ranks of the k PUGs are illustrated in this enlarged image. The null distribution of ranks of uninformative genes is to be estimated from these ranks. (d) An example null distribution estimated from PUGs. For each gene on the microarray, its actual ranking is compared with the null distribution to calculate the position p-value of the gene being noninformative. For mr-test and tr-test, the average position in the L rankings is used in the calculation of the p*-value. In tr-test, the PUGs are not selected from the bottom of the ranking, but rather from the genes with the largest t-test p-values.

U'_G. The next step is to estimate the null distribution of ranks with the ranking positions of these PUGs.

From the original data set, we resample L new data sets and apply the ranking method on each of them, producing, for each gene L, ranks $r^l_j, l = 1, \cdots, L$. In our implementation, we randomly resample half of the cases in the original data set each time. Other resampling schemes such as bootstrapping can also be used to obtain similar results according to our experiments (data not shown). Since, usually, the size of U_G is much larger than that of I_G (i.e., most genes are uninformative), if a gene tends to always be ranked at the bottom in the L rankings, it is very likely that the gene is an uninformative one. Thus, we define $\bar{r}_j$ as the average position of gene j in the L rankings,

$$\bar{r}_j = \frac{1}{L}\sum_{l=1}^{L} r^l_j, \quad j = 1, \cdots, n \tag{5}$$

and select the bottom k genes with the largest $\bar{r}_j$ as the PUGs to form U'_G, where k is a preset number. We rewrite U'_G as U'^k_G when we need to emphasize the role of k in this procedure. We assume that U'^k_G is a random sample of U_G and r^l_j, $l = 1, \cdots, L$ for $g_j \in U'^k_G$ is a random sample from the underlying null distribution of the ranks of uninformative genes. Thus, we have $k \cdot L$ observations of the null distribution of ranks from which we estimate the null distribution using a histogram. More sophisticated nonparametric methods can be adopted to fit the distribution if necessary. We denote the estimated null distribution as

$$\hat{P}(r|g \in U_G) = P_{histogram}(r^l_j | g_j \in U'^k_G). \tag{6}$$

With this estimated null distribution, the calculation of the position p-value is straightforward: For gene i with ranking position r_i,

$$\hat{p}(r_i) = \hat{P}(r \le r_i | g \in U_G) = P_{histogram}(r^l_j \le r_i | g_j \in U'^k_G). \tag{7}$$

Applying this on all the genes, we convert the ranking list to a list of position p-values reflecting the significance of the genes' being informative to the classification.

This whole procedure for estimating the p-value of a ranking is illustrated in Fig. 1. We name this scheme the r-test and call the position p-value thus calculated the r-test p-value.

4 Compensation for Bias in the Estimated PUGs

One important problem with the r-test scheme is the selection of the PUGs. Ideally, the ranks used to select PUGs and the ranks used to estimate null distribution should be independent. However, this is impractical in that

there is actually only one data set available. In our strategy, the same ranks are used to estimate both PUGs and the distribution of their ranks. This is an unplanned test in the sense that the PUGs are defined after the ranks are observed [14]. The PUGs in $U_G^{\prime k}$ are not an unbiased estimate of U_G. In the extreme case when k is small, uninformative genes that are ranked higher are underrepresented and the ranks of $U_G^{\prime k}$ might represent only a tail of the ranks of U_G on the right. If this happens, it will cause an overoptimistic estimation of the r-test p-values and result in more genes being claimed significant. Therefore, we propose two modified strategies to compensate for the possible bias.

4.1 Modified r-Test with Average Ranks

In (7), the position p-value is calculated by comparing the rank r_i of gene g_i obtained from the whole data set with the estimated null distribution. Intuitively, when the sample size is small, one single ranking based on a small sample set can have a large variance, especially when all or most of the genes are uninformative. We propose replacing the rank r_i by the $\bar{r}_i$ defined in (5), i.e., to use the average position of gene g_i in the L resampling experiments as the estimate of the true rank, and to calculate the position p-value with this estimated rank rather than the single observation of the rank:

$$p^*(\bar{r}_i) = \hat{P}^*(r \le \bar{r}_i | g \in U_G) = P_{histogram}(r_j^l \le \bar{r}_i | g_j \in U_G^{\prime k}). \quad (8)$$

The estimated null distribution is the distribution of single ranks of putative uninformative genes, but the $\bar{r}_i$ to be compared to it is the averaged rank and (8) is no longer a p-value in the strict sense. Therefore, we name it p*-value instead and call this modified r-test the mr-test for convenience. Ideally, if a gene is informative to the classification and the ranking method can consistently rank the gene according this information on both the whole data set and on the resampled subsets, we'll have

$$\bar{r}_i = r_i, \text{ for } g_i \in I_G, \quad (9)$$

in which case the p*-value will be equivalent to the original r-test p-value for these genes. In practice, when the sample size is small and the signal in some informative genes are not so strong, we always have $\bar{r}_i \ge r_i$ when r_i is small; therefore, the estimated ranks move toward the right on the rank distribution comparing with the single-run ranks, which, as an effect, can be a compensation to the bias in the estimated null distribution. (For the genes ranked in the lower half of the list, the averaged rank will move leftward, but these genes are not of interest to us in this study since we assume only a minority of the genes can be informative.)

4.2 Independent Selection of PUGs

The ultimate reason that may cause biased estimation of the null distribution is that the PUGs in the above r-test scheme are estimated from the same ranking information as that being used for the calculation of the test statistics. A solution is to select a group of PUGs that are an unbiased sample from the U_G. This is a big challenge because estimating the rank position distribution of U_G is the question itself.

When the ranking method *RM* is a multivariate one such as SVM-based methods, the ranking of the genes will not directly depend on the differences of single genes between the classes. We therefore can use a univariate statistic such as the t-test to select a group of nondifferentially expressed genes as the PUGs since these genes will have a high probability of not being informative as they are basically the same in the two classes. This selection will be less correlated with the ranking by *RM*. Applying a threshold η on the t-test p-value p^t, we select the PUG set $U_G^{\prime\eta}$ as:

$$U_G^{\prime\eta} \triangleq \{g_j | g_j \in G, p^t(g_j) \ge \eta\} \quad (10)$$

and estimate the null position distribution according the ranking of the $U_G^{\prime\eta}$ genes by *RM* in the resampled data:

$$\hat{P}^t(r | g \in U_G) = P_{histogram}(r_j^l | g_j \in U_G^{\prime\eta}). \quad (11)$$

The position p*-value of a gene ranked on average at $\bar{r}_i$ is calculated as:

$$\hat{p}^*(\bar{r}_i) = \hat{P}^t(r \le \bar{r}_i | g \in U_G) = P_{histogram}(r_j^l \le \bar{r}_i | g_j \in U_G^{\prime\eta}). \quad (12)$$

For the convenience of discussion, we call this strategy the tr-test and call the primary r-test defined by (7) the pr-test. We view the pr-test, mr-test, and tr-test as three specific methods under the general r-test scheme.

It should be noted that if the ranking produced by *RM* is highly correlated with t-test ranking, the result of tr-test will be close to that of the original pr-test. On the other hand, since insignificant genes evaluated individually may not necessarily be uninformative when combined with certain other genes, the PUGs selected by (10) may include informative genes for *RM*. Therefore, the estimated null distribution may bias toward the left end in some situations, making the results overconservative. However, in the experiments described below, it is observed that the tr-test results are not sensitive to changes in the p-value cut-offs used for selecting PUGs (10), which is an implication that the method is not very biased.

5 Experiments with SVM on Real and Simulated Data

5.1 r-Test with SVM Gene Ranking

Due to the good generalization ability of support vector machines (SVM) [15], they are regarded as one of the best multivariate algorithms for classifying microarray data [9], [10], [16]. In the experiments for r-test in this work, we adopted linear SVM as the ranking machine *RM*. The linear SVM is trained with all genes in the data set, producing the discriminate function

$$f(\mathbf{x}) = \mathbf{w}^* \cdot \mathbf{x} + b^*, \quad (13)$$

where $\mathbf{w}^* = \sum_{i=1}^n \alpha_i^* y_i \mathbf{x}_i$ and α_i^* are the solutions of the following quadratic programming problem:

$$L_p = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^n \alpha_i y_i (\mathbf{x}_i \cdot \mathbf{w} + b) + \sum_{i=1}^n \alpha_i. \quad (14)$$

Following [10], the contribution of each gene in the classifier can be evaluated by

$$DL_p = (1/2)\frac{\partial^2 L_p}{\partial w_i^2}(Dw_i)^2 = (w_i)^2 \quad (15)$$

and, thus, the genes are ranked by $(w_i)^2$. There are other ways of assessing the relative contribution of the genes in a SVM classifier [17], but, since the scope of this paper is not to discuss the ranking method, we adopt the ranking criterion given in (15) here. The ranking only reflects the relative importance of the genes in the classifier, but cannot reveal how important each gene is. The r-test converts the ranking to position p-values (or p*-values) to evaluate the significance.

5.2 Data Sets

Experiments were done on six microarray data sets: three real data sets and three simulated data sets. The leukemia data set [1] contains the expression of 7,129 genes (probe sets) of 72 cases, 47 of them are of the ALL class and 25 are of the AML class. The colon cancer data set [18] contains 2,000 genes of 62 cases, among which 40 are from colon cancers and 22 from normal tissues. These two data sets have been widely used as benchmark sets in many methodology studies. Another data set used in this study is a breast cancer data set [19] containing 12,625 genes (probe sets) of 85 cases. The data set is used to study the classification of two subclasses of breast cancer. Forty-two of the cases are of class 1 and 43 are of class 2.

Simulated data sets were generated to investigate the properties of the methods in different situations. The first case is for an extreme situation where none of the genes are informative. The simulated data set contains 1,000 genes and 100 cases. The expression values of the genes are independently generated from normal distributions with randomized means and variations in a given range. The 100 cases are generated with the same model, but are assigned arbitrarily to two fake classes (50 cases in each class). So, the two classes are, in fact, not separable and all the genes are uninformative. We refer to this data set as the "fake-class" data set in the following description.

Each of the other two simulated data sets also contains 1,000 genes and 100 cases of two classes (50 cases in each class). In one data set (we call it "simu-1"), 700 of the genes follow N(0, 1) for both classes and the 300 genes follow N(0.25, 1) for class 1 and N(-0.25, 1) for class 2. In the other data set (we call it "simu-2"), 700 of the genes follow N(0, 1) for both classes and the 300 genes follow N(0.5, 1) for class 1 and N(-0.5, 1) for class 2. With these two simulated data sets, we hope to mimic situations where there are weak and strong classification signals in the data.

All the data sets except simu-1 and simu-2 were standardized to 0-mean and standard deviation 1 first across the cases and then across the genes. This is to prevent possible bias in the ranking affected by the scaling. In practical investigations, this step might not be needed or might need to be done in some other way according to the specific situation of the data and the specific ranking methods to be adopted.

The six data sets used in our experiments represent different levels of separability of the investigated classes. For the leukemia data set, almost perfect classification accuracy has been achieved [1], [9], [10], so it represents a relatively easy classification task. For the colon cancer data set, the samples can still be well separated, but with some errors [10], [18]. The two subclasses studied in the breast cancer data set are hardly separable as observed in this data set, but it is believed that there could be some degree of separability [20], [21]. The fake-class simulation represents a situation where the two classes are completely nonseparable and the simu-1 and simu-2 simulation represents an ideal situation where separation is defined on a subset of the genes and the uninformative genes are i.i.d. To check the classification accuracy that can be achieved on these data sets, we randomly split them into independent training and test sets and applied linear SVM on them. These experiments were done 200 times for each data set and the classification accuracy obtained at different gene selection levels is summarized in Table 1. It can be seen that the accuracies are consistent with the reports in the literature and with the design of the simulations. (Note that the error rates reported here are independent test results based on only half of the samples for training, so they are larger than the cross-validation errors reported elsewhere. The scope of this paper is not to improve or discuss classification accuracy.)

TABLE 1
Separability of the Classes of the Six Data Sets

Datasets	Number of genes[1]	SVM error rate[2]
Leukemia	1000	0.050 ± 0.034
	100	0.065 ± 0.034
	20	0.110 ± 0.042
Colon cancer	1000	0.185 ± 0.037
	100	0.185 ± 0.040
	20	0.241 ± 0.055
Breast cancer	1000	0.468 ± 0.017
	100	0.463 ± 0.033
	20	0.490 ± 0.017
Fake-class data	1000	0.440 ± 0.035
	100	0.460 ± 0.042
	20	0.483 ± 0.014
Simu-1 data	1000	0.0089 ± 0.0091
	100	0.0471 ± 0.0220
	20	0.2376 ± 0.0372
Simu-2 data	1000	0.0 ± 0.0
	100	0.0013 ± 0.0037
	20	0.1772 ± 0.0341

1. Number of genes selected with SVM-RFE (Guyon et al, 2002).
2. The error rates are calculated in this way: In each run, half of the cases were used as training set for SVM-RFE, and the other half as independent test set. The error rates here are the summary of 200 runs.

TABLE 2
The Number of Genes Selected at Various r-Test p*-Value Levels with SVM

Dataset	k: number of PUGs	p*-value level				
		0.001	0.005	0.01	0.05	0.1
Leukemia	700	396	652	853	1609	2341
	800	374	618	813	1556	2241
Colon cancer	50	33	76	110	366	657
	70	29	70	101	331	600
Breast Cancer	900	18	130	242	1057	2324
	1100	17	104	209	963	2113
Fake-class Data	30	2	2	4	32	77
	40	2	2	4	32	73
	200	0	2	2	17	47
Simu-1 Data	30	39	96	126	227	310
	40	36	91	124	227	306
	200	20	58	91	190	257
Simu-2 Data	30	300	301	306	321	353
	40	300	301	306	320	349
	200	297	300	301	315	328
	500	288	297	300	307	315

5.3 Number of Significant Genes According to the mr-Test and tr-Test

We systematically experimented with the SVM-based pr-test, mr-test, and tr-test methods on the six data sets and studied the number of genes claimed as significantly informative with each method at various significant levels. The results of the pr-test are affected by different choices of the number k of selected PUGs (data not shown), indicating that the pr-test can be very biased unless we know the accurate number of informative genes. Therefore, we focus on the mr-test and tr-test in the following discussion.

Table 2 shows the number of significant genes according to the mr-test at different p*-value levels, with different choices of ks on the six data sets. Comparing with the pr-test results, the mr-test results are less sensitive to changes in the number k. This is especially true when there are ideal classification signals, as in the simu-2 data, where we can see a more than 10-fold change of k causes only little variance in the estimated gene numbers. With p*-value levels from 0.001 to 0.1, the estimated significant gene numbers are all around the correct number (300). The claimed significant genes are all those true informative genes in the model when the estimated genes are less than 300. For the situations where the number of estimated informative genes is larger than 300, all the true informative genes are discovered. When the data are less ideal, we see that the results are stable within a smaller variation of k. More experiment results with larger variations in the choice of k are provided in the supplemental material, which can be found on the Computer Society Digital Library at http://computer.org/tcbb/archives.htm. From Table 2, it can also be observed that the number of significant genes is not directly correlated with the classification accuracy. For example, the breast cancer data and fake-class data both look nonseparable according to the classification errors (Table 1), However, for the breast cancer data, more than 200 genes are identified as significantly informative among the 12,625 genes ($\sim 1.6\%$) at the p*-value = 0.01 level, but, for the fake-class data, this number is only about 0.4 percent of the 1,000 genes.

Results of the tr-test with different t-test p-value cut-offs are shown in Table 3. It can be seen that different cut-offs result in different numbers of PUGs, but the variation in estimated position p*-values due to PUG number difference is even smaller than in the mr-test. This implies that the tr-test results are not biased by the selection of PUGs since, if the PUG selection was biased, different numbers of PUGs at t-test p-value cut-offs would have caused different degrees of bias and the results would have varied greatly. Comparing between Table 2 and Table 3, as well as the results in the supplemental material, which can be found on the Computer Society Digital Library at http://computer.org/tcbb/archives.htm, we observe that, for the mr-test, although there is a range of k for each data set in which the results are not very sensitive to variations of k, this range can be different with different data sets. On the other hand, for the tr-test, within the same ranges of cut-off t-test p-values, results on all the data sets show good consistency with regard to variations in the cut-off value. This makes the tr-test more applicable since users do not need to tune the parameter specifically to each data set.

Comparing the number of genes selected by the tr-test and mr-test (Table 2 and Table 3), it is obvious that the tr-test is more stringent and selects much fewer genes than

TABLE 3
The Number of Genes Selected at Various Position p-Value Levels by tr-Test with SVM

Data Set	η[1]	k: number of PUGs	p*-value level				
			0.001	0.005	0.01	0.05	0.1
Leukemia	0.7	1414	15	65	112	416	760
	0.9	466	13	56	104	406	756
Colon Cancer	0.7	361	0	0	0	13	54
	0.9	127	0	0	0	13	54
Breast Cancer	0.7	4208	0	1	3	44	191
	0.9	1371	0	1	4	58	217
Fake-class Data	0.7	325	0	2	2	8	34
	0.9	102	0	2	2	11	37
Simu-1 Data	0.7	202	12	36	59	156	221
	0.9	67	13	39	67	175	237
Simu-2 Data	0.7	195	294	298	300	307	317
	0.9	59	295	299	300	307	318

[1] The t-test p-value cut-off for selecting PUGs

the mr-test on the real data sets. The differences are smaller on the simulated data. Similarly to the results of the mr-test, almost all the informative genes in simu-2 data can be recovered at p*-value levels from 0.001 to 0.05 and there are only a very few false-positive genes (e.g., the 307 genes selected at p*-value = 0.05 contains all the 300 true-positive genes and seven false-positive genes). This shows that the SVM method is good in both sensitivity and specificity in selecting the true informative genes for such ideal case, and both of the two r-test methods can detect the correct number of informative genes at a wide range of significance levels. For simu-1 data, not all the informative genes can be recovered in the experiments. This reflects the fact of the large overlap of the two distributions in this weak model. Many of original 300 "informative" genes are actually not statistically significant in the contexts of both univariate methods and multivariate methods.

For the real data sets, there are no known answers for the "true" number of informative genes. The mr-test uses the tail in the ranking list to estimate the null distribution for assessing the significance of the genes on the top of the list, therefore there is a higher possibility of the p*-values being underestimated, although this has been partially compensated by using the average rank positions. Thus, the number of genes being claimed significant by mr-test might be overestimated. In this sense, the tr-test scheme provides a more unbiased estimation of the null distribution, which is supported by the decreased sensitivity to PUG numbers. With the tr-test, at the 0.05 p*-value level, we get about 410 significant genes from the 7,129 genes (5.75 percent) in the leukemia data. On the other two real data sets, the results tend to be too conservative: about 13 out of 2,000 genes (0.65 percent) in the colon cancer data and 50 out of 12,625 genes (0.4 percent) in the breast cancer data are claimed as significant at this level.

It should be noted that the PUGs selected according to the t-test may contain informative genes for SVM, which considers the combined effects of genes. This will cause the number of genes called significant by the tr-test to be underestimated. This is especially true for data sets in which the major classification signal exists in the combinatorial effects of genes instead of differences in single genes. The correct answer may be somewhere between the two estimations of the tr-test and mr-test. When the signal is strong, the two estimations will be close as we see in the simu-2 data. In practice, one can choose which one to use according to whether the purpose is to discover more possibly informative genes or to discover a more manageable set of significant genes for follow-up investigations.

6 Discussion

Statistical hypothesis testing is a fundamental framework for scientific inference from observations. Unfortunately, existing hypothesis testing methods are not sufficient to handle high-dimensional multivariate analysis problems arising from current high-throughput genomic and proteomic studies. Many new data mining techniques have been developed both in statistics and in the machine learning field. These methods are powerful in analyzing complicated high-dimensional data and helped greatly in functional genomics and proteomics studies. However, the analysis of the statistical significance of data mining results has not been paid enough attention. One reason might be that many methods are rooted in techniques aimed at solving problems in engineering and technological applications rather than in scientific discoveries. As an example, many machine-learning-based gene selection and classification methods may achieve very good performance in solving the specific classification problems, but the results are usually of a

"black-box" type and judging the significance of the features being used for the classification was usually not deemed important. This fact compromises their further contribution in helping biologists to understand the mechanisms underlying the investigated disease classification.

This paper raises the problem of the significance of gene rankings in microarray classification study and proposes a solution strategy called the r-test that converts the ranking of genes obtained with any method to position p-values (p*-values) that reflect the significance of the genes being informative. The concept of this question is important and the formulation and solution are challenging for several reasons as addressed in the paper. First of all, the definition of a gene being informative to the classification may not yet be completely clear for many classification methods. Even under the same criterion, there may not be a clear boundary between informative and uninformative genes. A biological status may be affected by several genes with different levels of contribution and it may affect the expression of many other genes. Differences between individuals and instrumental noises may make the genes that have no relation with the studied biological process show some relevancy in the limited samples. All these (and other) complexities make it hard to mathematically model microarray data. We propose the r-test methods based on intuitive reasoning under certain assumptions about the nature of the data. As shown in the experiments, the methods provide reasonable solutions, but the decision by the mr-test and tr-test method can be very different for some situations. Theoretically, rigorous methods are still to be developed.

Under the proposed r-test framework, the key issue is the choice of putative uninformative genes or PUGs. Since the null distribution has to be estimated from the data themselves, avoiding bias in the estimation is the most challenging task. Besides the methods used by the mr-test and tr-test, we have also tried several other ways to tackle the problem, including selecting the PUGs according to the distribution of the ranks of all genes in the resample experiments, deciding the number of PUGs recursively according to the rank with an EM-like strategy, selection of an independent set of PUGs by fold-change, etc. Different resampling strategy has also been experimented with. Among these efforts, the reported mr-test and tr-test give the most satisfactory results. They both perform perfectly on ideal simulations. For practical cases, the mr-test has a tendency to be overoptimistic by claiming more significant genes and the tr-test has a tendency to be conservative by approving only a small number of significant genes. Note that both r-test schemes do not change the ranking itself; therefore, it is the role of the classification and gene selection method (the *RM*) to guarantee that the ranking itself is reasonable for the biological investigation. The r-test only helps to decide on the number of genes to be selected from the list at given significance levels. Since there is currently no theoretical solution to completely avoid estimation bias, one can make a choice between mr-test and tr-test results by balancing between the two opposite trends of possible biases according to the particular biological problem at hand.

ACKNOWLEDGMENTS

The authors would like to thank Drs. J.D. Iglehart and A.L. Richardson for providing them with their microarray data for the experiments. They would also like to thank the editor and reviewers for their valuable suggestions that contributed a lot to the work. They thank Dustin Schones for helping to improve their writing. This work is supported in part by NSFC projects 60275007, 60234020, and the National Basic Research Program (2004CB518605) of China. This work was performed while Chaolin Zhang was with the MOE Key Laboratory of Bioinformatics/Department of Automation, Tsinghua University, Beijing. Chaolin Zhang and Xuesong Lu contributed equally in this work and should be regarded as joint authors. The corresponding author is Xuegong Zhang.

REFERENCES

[1] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, and E.S. Lander, "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring," *Science,* vol. 286, no. 5439, pp. 531-537, 1999.

[2] C.-A. Tsai, Y.-J. Chen, and J.J. Chen, "Testing for Differentially Expressed Genes with Microarray Data," *Nuclear Acids Research,* vol. 31, no. 9, p. e52, 2003.

[3] M.S. Pepe, G. Longton, G.L. Anderson, and M. Schummer, "Selecting Differentially Expressed Genes from Microarray Experiments," *Biometrics,* vol. 59, no. 1, pp. 133-142, 2003.

[4] P. Broberg, "Statistical Methods for Ranking Differentially Expressed Genes," *Genome Biology,* vol. 4, no. 6, p. R41, 2003.

[5] W. Pan, "A Comparative Review of Statistical Methods for Discovering Differentially Expressed Genes in Replicated Microarray Experiments," *Bioinformatics,* vol. 18, no. 4, pp. 546-554, 2002.

[6] S. Ramaswamy, K.N. Ross, E.S. Lander, and T.R. Golub, "A Molecular Signature of Metastasis in Primary Solid Tumors," *Nature Genetics,* vol. 33, no. 1, pp. 49-54, 2003.

[7] L.J. van 't Veer, H. Dai, M.J. van de Vijver, Y.D. He, A.A.M. Hart, M. Mao, H.L. Peterse, K. van der Kooy, M.J. Marton, A.T. Witteveen, G.J. Schreiber, R.M. Kerkhoven, C. Roberts, P.S. Linsley, R. Bernards, and S.H. Friend, "Gene Expression Profiling Predicts Clinical Outcome of Breast Cancer," *Nature,* vol. 415, no. 6871, pp. 530-536, 2002.

[8] M. Xiong, X. Fang, and J. Zhao, "Biomarker Identification by Feature Wrappers," *Genome Research,* vol. 11, no. 11, pp. 1878-1887, 2001.

[9] X. Zhang and H. Ke, "ALL/AML Cancer Classification by Gene Expression Data Using SVM and CSVM Approach," *Proc. Conf. Genome Informatics,* pp. 237-239, 2000.

[10] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene Selection for Cancer Classification Using Support Vector Machines," *Machine Learning,* vol. 46, pp. 389-422, 2002.

[11] C. Furlanello, M. Serafini, S. Merler, and G. Jurman, "Entropy-Based Gene Ranking without Selection Bias for the Predictive Classification of Microarray Data," *BMC Bioinformatics,* vol. 4, no. 1, p. 54, 2003.

[12] H. Yu, J. Yang, W. Wang, and J. Han, "Discovering Compact and Highly Discriminative Features or Feature Combinations of Drug Activities Using Support Vector Machines," *Proc. 2003 IEEE Bioinformatics Conf. (CSB '03),* 2003.

[13] J.D. Storey and R. Tibshirani, "Statistical Significance for Genome Wide Studies," *Proc. Nat'l Academy of Science USA,* vol. 100, no. 16, pp. 9440-9445, 2003.

[14] R.R. Sokal and F.J. Rohlf, *Biometry.* San Francisco: Freeman, 1995.

[15] V. Vapnik, *The Nature of Statistical Learning Theory.* Springer-Verlag, 1995.

[16] S. Ramaswamy, P. Tamayo, R. Rifkin, S. Mukherjee, C.-H. Yeang, M. Angelo, C. Ladd, M. Reich, E. Latulippe, J.P. Mesirov, T. Poggio, W. Gerald, M. Loda, E.S. Lander, and T.R. Golub, "Multiclass Cancer Diagnosis Using Tumor Gene Expression Signatures," *Proc. Nat'l Academy of Sciences,* vol. 98, no. 26, pp. 15149-15154, 2001.

[17] X. Zhang and W.H. Wong, "Recursive Sample Classification and Gene Selection Based on SVM: Method and Software Description," technical report, Dept. of Biostatistics, Harvard School of Public Health, 2001.
[18] U. Alon, N. Barkai, D.A. Notterman, K. Gish, S. Ybarra, D. Mack, and A.J. Levine, "Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Arrays," *Proc. Nat'l Academy of Sciences,* vol. 96, no. 12, pp. 6745-6750, 1999.
[19] Z.C. Wang, M. Lin, L.-J. Wei, C. Li, A. Miron, G. Lodeiro, L. Harris, S. Ramaswamy, D.M. Tanenbaum, M. Meyerson, J.D. Iglehart, and A. Richardson, "Loss of Heterozygosity and Its Correlation with Expression Profiles in Subclasses of Invasive Breast Cancers," *Cancer Research,* vol. 64, no. 1, pp. 64-71, 2004.
[20] E. Huang, S.H. Cheng, H. Dressman, J. Pittman, M.H. Tsou, C.F. Horng, A. Bild, E.S. Iversen, M. Liao, and C.M. Chen, "Gene Expression Predictors of Breast Cancer Outcomes," *The Lancet,* vol. 361, no. 9369, pp. 1590-1596, 2003.
[21] M. West, C. Blanchette, H. Dressman, E. Huang, S. Ishida, R. Spang, H. Zuzan, J.A. Olson Jr., J.R. Marks, and J.R. Nevins, "Predicting the Clinical Status of Human Breast Cancer by Using Gene Expression Profiles," *Proc. Nat'l Academy of Sciences,* vol. 98, no. 20, pp. 11462-11467, 2001.

Chaolin Zhang received the BE degree from the Department of Automation at Tsinghua University, Beijing, China, in 2002. From 2002 to 2004, he worked as a graduate student on machine learning applications in microarray data analysis and literature mining at the MOE Key Laboratory of Bioinformatics, Tsinghua University. He is now a PhD student at Cold Spring Harbor Laboratory and the Department of Biomedical Engineering, the State University of New York at Stony Brook.

Xuesong Lu received the BE degree from the Department of Automation, Tsinghua University, Beijing, China, in 2001. He is currently a PhD candidate in the Department of Automation and the MOE Key Laboratory of Bioinformatics at Tsinghua University, Beijing, China. His research interests include microarray data mining, gene network modeling, and literature mining.

Xuegong Zhang received the PhD degree in pattern recognition and intelligent systems from Tsinghua University, Beijing, China, in 1994. He is currently a professor in the Department of Automation and the MOE Key Laboratory of Bioinformatics at Tsinghua University. His research interests include machine learning and pattern recognition, bioinformatics, computational genomics, and systems biology.