DNA AND CELL BIOLOGY
Volume 23, Number 10, 2004

Pp. 607–614

Gene Selection for Sample Classifications in Microarray Experiments

CHEN-AN TSAI,[1] CHUN-HOUH CHEN,[2] TE-CHANG LEE,[3,4] I-CHING HO,[4] UENG-CHENG YANG,[5] and JAMES J. CHEN[1]

ABSTRACT

DNA microarray technology provides useful tools for profiling global gene expression patterns in different cell/tissue samples. One major challenge is the large number of genes relative to the number of samples. The use of all genes can suppress or reduce the performance of a classification rule due to the noise of nondiscriminatory genes. Selection of an optimal subset from the original gene set becomes an important prestep in sample classification. In this study, we propose a family-wise error (FWE) rate approach to selection of discriminatory genes for two-sample or multiple-sample classification. The FWE approach controls the probability of the number of one or more false positives at a prespecified level. A public colon cancer data set is used to evaluate the performance of the proposed approach for the two classification methods: *k* nearest neighbors (*k*-NN) and support vector machine (SVM). The selected gene sets from the proposed procedure appears to perform better than or comparable to several results reported in the literature using the univariate analysis without performing multivariate search. In addition, we apply the FWE approach to a toxicogenomic data set with nine treatments (a control and eight metals, As, Cd, Ni, Cr, Sb, Pb, Cu, and AsV) for a total of 55 samples for a multisample classification. Two gene sets are considered: the gene set Ω_F formed by the ANOVA *F*-test, and a gene set Ω_T formed by the union of one-versus-all *t*-tests. The predicted accuracies are evaluated using the internal and external crossvalidation. Using the SVM classification, the overall accuracies to predict 55 samples into one of the nine treatments are above 80% for internal crossvalidation. Ω_F has slightly higher accuracy rates than Ω_T. The overall predicted accuracies are above 70% for the external crossvalidation; the two gene sets Ω_T and Ω_F performed equally well.

INTRODUCTION

DNA MICROARRAY TECHNOLOGY provides tools to simultaneously study the expression profiles of thousands of distinct genes in a single experiment. Application of this technology ranges from the study of gene expression in yeast under a variety of experimental conditions (e.g., Eisen *et al*., 1998) to the study of differences between normal and tumor tissues (Alon *et al*., 1999) and differences between different tumor subtypes (Golub *et al*., 1999). DNA array technology can also be applied to toxicology testing for toxicity screening of unknown compounds and as a tool for mechanistic studies (Burczynski *et al*., 2000). Clustering analysis and classification are two commonly used analyses for determining relationships between genes or gene clusters to identify biological functions or to predict specific biological sample outcomes.

One major challenge in the analysis of gene expression data is the large number of genes in the data set. Many of those genes are not relevant to clustering or classification. Prior to clustering/classifying the data, there are often questions about adjusting the data in some way to enhance relationships between genes and samples. Data can be removed if they do not provide significant incremental information, or more importantly, they may confuse the analysis and make it unnecessar-

[1]Division of Biometry and Risk Assessment, National Center for Toxicological Research, Food and Drug Administration, Jefferson, Arkansas.
[2]Institute of Statistical Science, and [4]Institute of Biomedical Sciences, Academia Sinica, Taipei, 115, Taiwan.
[3]Institute of Biopharmaceutical Sciences, and [5]Institute of Biochemistry, National Yang-Ming University, Taipei, 112, Taiwan.

ily complex. If the expression of a particular gene is the same in all samples, it will not be useful for distinguishing these samples. If the expressions for a gene are very different over all samples, it may contain useful information to distinguish them (Raychaudhuri *et al*., 2001). Therefore, selection of an "optimal" subset (gene identification) from the original data set is an important prestep in clustering and classification. A common approach is to select a fixed number of the highest ranked genes based on *t*-test-like statistics or some discrimination measures (e.g., Liu *et al*., 2002). A problem with this approach is that the selected differentially expressed genes do not meet statistical criterion of controlling false positive error rates. Also, the number of genes selected for a follow-up analysis is arbitrary.

Selection of an appropriate gene set can be obtained using statistical significance testing. A test statistic with its corresponding *p*-value is calculated for each gene to determine differential expressions (either overexpressed or underexpressed) among experimental samples. A small *P*-value indicates an evidence of differential expressions. Typically, an investigator either selects those genes with *P*-values below the prespecified cutoff "significance" level (discussed below) or selects a fixed number of the genes with the smallest *P*-values (e.g., Alon *et al*., 1999; Nguyen and Rocke, 2002). A subset of "significantly" differentially expressed genes is then identified.

The *P*-value is ordinarily defined under a single hypothesis. The *P*-value is the probability for the experimental outcome if there is no difference among samples for an individual gene. A test procedure is said to control the Type I error probability at the significance level α if the observed *P*-value $\leq \alpha$. The level α is a *marginal* Type I error; the probability inference refers only to the particular gene, *irrespective of the results for the other genes*. In a microarray experiment, hundreds or thousands of tests (genes) are conducted, simple use of *P*-values (comparing the observed *P*-values with the α to determine significant genes) without adjustment for multiple testings could lead to a large chance of false positive findings. For example, if m tests are made with each at a significance level of α, then the probability of one or more false positives can be as large as $1 - (1 - \alpha)^m$. For $m = 1000$ and $\alpha = 0{:}05$, the probability of one or more false positives is almost $1 - (1 - 0{:}05)^{1000} \approx 1$. That is, some genes will have *P*-values less than 0.05, even when the samples are not different.

This article proposes using the family-wise error rate (FWE) controlled approach to gene selection for sample classification (prediction). The goal of sample prediction is to develop a decision rule that accurately predicts the class membership of a new sample based on the expression profiles of the selected genes. Several classification algorithms have been adopted for classification of cancer subtypes or gene functions. Three commonly used classification algorithms are the Fisher's linear discriminant function, nearest-neighbor classifiers, and support vector machines (e.g., Ramaswamy *et al*., 2001, Dudoit *et al*., 2002). The Fisher's discriminant function method has not performed as well on most of the data sets evaluated. In this paper, we evaluate the proposed FWE approach using the nearest-neighbor classifiers and support vector machine algorithms. We use the public colon cancer data set (Alon *et al*., 1999) and a toxicogenomic data set to illustrate and evaluate the predictive accuracy of the proposed approach.

MATERIALS AND METHODS

Sample classification can be decomposed into three steps: (1) selection of discriminatory genes, (2) selection of prediction methods, and (3) crossvalidation to estimate accuracy of prediction.

Selection of a discriminatory gene set

Let $x_{i,c_1} \ldots, x_{i,cn_1}$ denote the intensity from the control (normal) group with n_1 samples for gene i and $x_{i,t_1} \ldots, x_{i,tn_2}$ denote the intensity from the treated (diseased) group with n_2 samples, $i = 1, \ldots, m$. To determine a differentially expressed gene, say gene i, the procedure can be formulated as the hypothesis:

$$H_0 : \mu_{ic} = \mu_{it} \text{ versus } H_1 : \mu_{ic} \neq \mu_{it},$$

where μ_{ic} and μ_{it} denote the mean of gene i for a normal group and a diseased group, respectively. Common statistical significance testing approach is to compute the *t*-statistic

$$t_i = \frac{|\bar{x}_{ic} - \bar{x}_{it}|}{\sqrt{s_{ic}^2/n_1 + s_{it}^2/n_2}}$$

where $\bar{x}_{ic}$ and $\bar{x}_{it}$ are the means of gene i for the control and treatment groups, and s_{ic}^2 and s_{it}^2 are the sample variances for the control and treated groups, respectively. Because the gene expression data are generally not normally distributed, the random permutation test is often recommended to compute the (unadjusted) *P*-values of the *t*-statistic.

An approach to account for multiple testings is to control the family-wise error rate (FWE). The family-wise error rate (experiment-wise error rate) is the error probability associated with one or more false rejections for all tests included in the experiment. That is, the FWE approach ensures that the probability of making one or more false positives (among the genes tested) is less than a predetermined level α (e.g., 0.05). The simplest FWE method is the *Bonferroni correction* by dividing α by the number of genes m as a significance level for each individual test. Because the test for each gene has an α/m probability of making Type I error, the family-wise error rate is at most α. The *Bonferroni correction* procedure equivalently can be performed by multiplying all *P*-values by m; the *P*s so obtained are then compared with α. These *P*-values are referred to as the adjusted *P*-values. Thus, the adjusted *P*-values take into account the multiplicity. The adjusted *P*-values can be compared directly with the FWE-based significance level. In the FWE approach, the genes with adjusted *P*-values less than or equal to α are selected. The Bonferroni procedure is known to be conservative in that the actual family-wise error is less than α. Several modified Bonferroni procedures and resampling methods (Hochberg and Tamhane, 1987; Westfall and Young, 1993) have been developed to reduce the conservatism.

Mathematical theory to compute the adjusted *P*-values is explained briefly below. The probability of one or more false pos-

itives is equal to the probability that the smallest P-value is less than or equal to α; alternatively, the largest of the observed t-statistics $t_1, \ldots, t_m$ is significant at the α level. Thus, the FWE can be defined according to the probability distribution of the maximum t-statistic, denoted by $t(m)$, under the null hypothesis. The FWE corrected (adjusted) P-value for gene i is

$$p[t]_i = P\{t_{(m)} \geq t_i | \text{ under the null hypothesis}\}.$$

The exact distribution of $t_{(m)}$ cannot be computed mathematically. The adjusted P-values are computed using the permutation resampling method (Westfall and Young, 1993). The algorithm for computing adjusted P-values is given as follow.

Resampling algorithm for computing adjusted P-values

0. Set $k_i = 0$ for $i = 1, \ldots, m$.
1. Compute raw statistic t_i from the sample data set $x_{i,c_1}, \ldots, x_{i,c_{n_1}}$, and $x_{i,t_1}, \ldots, x_{i,t_{n_2}}$ for $i = 1, \ldots, m$
2. Generate one bootstrap sample by sampling without replacement from the pooled sample of the original data set with the same sample size for each group.
3. Compute the statistic t_i^* (using the same method in step 1) from the bootstrap sample for $i = 1, \ldots, m$.
4. Find the maximum of $t_1^*, \ldots, t_m^*$ based on the bootstrap sample, $t_{(m)}^* = \max\{t_1^*, \ldots, t_m^*\}$.
5. Compare each t_i to $t_{(m)}^*$, and set $k_i = k_i + 1$, if $t_i \leq t_{(m)}^*$.
6. Repeated step 2–4 enough times B^*, say, 50,000.
7. Compute the proportion of bootstraps samples for which $t_{(m)}^* \geq t_i$ to obtain the adjusted P-value as $\tilde{p}_i = k_i/B^*$.

The adjusted P-values for more than two sample comparisons can be obtained using the analysis of variance (ANOVA) F-test statistic. The F-test assumes a constant variance across all experimental samples. Finally, because the FWE approach is a very stringent criterion, it is possible that no gene is selected. In this case, we assign each sample with an equal probability to each class.

Classification algorithms

A classification rule is derived from a training data set and then is used to classify (predict) new observation data. Two classification methods, based on a preselected set of genes to form discrimination rule, are considered: the k-nearest neighbor classifiers (k-NN), and support vector machines (SVM).

The k-NN classifiers are a typical memory-based prediction method (Fix and Hodges, 1951). Given a set of training data X^μ with class labels C^μ, the class label of a new testing sample $\mathbf{x}_0$ is determined in the following steps: (1) calculating the dissimilarity of $\mathbf{x}_0$ to each sample of the training data set X^μ, (2) finding the k closest points, $\{\mathbf{x}_1^{\mu^*}, \ldots, \mathbf{x}_k^{\mu^*}\} \in X^\mu$, in the training data set, and (3) assigning the class label by using the majority vote among the k closest neighbors. Apparently, the choice of k will influence the performance of the k-NN algorithm. The optimal setting of k can be determined using crossvalidation techniques. Based on a preliminary analysis, k was set to be 1.

The SVM classifiers are a machine-learning prediction (Vapnik, 1998). In a two-class classification, an SVM classifier tries to draw a hyperplane in the m-dimensional gene expression space between the two classes. If no separating hyperplane exits, the samples are mapped into a higher dimensional space where such a separator does exist. In this paper, we use the Gaussian kernel. We also extend the procedure to the multiple class classification.

Crossvalidation

Microarray gene expression data are characterized by the number of genes (variables) far exceeding the number of samples. This presents challenges for classification algorithms, which are generally designed with a large number of samples over few variables. A common problem is overfitting the data. That is, the predicted model can fit the original data well but may predict poorly for new data. Classification algorithms, typically, involve a training phase run on samples whose classes are already known, and a testing phase generalizes the algorithm developed from the training data to predict classes of the test data and to estimate the predicted error rate. Training data and testing data ideally should be independent and identically distributed data sets. However, in the context of microarray data, because the number of arrays is usually small, the crossvalidation approach often uses a large fraction of data in the training phase and the remaining fraction in the testing phase to estimate the error rate. For example, in the "leave-one-out" crossvalidation, one sample is excluded each time from the whole sample, and then is classified in the testing phase based on the predictive model developed from the training set. This process is iterated until all samples are classified.

In the leave-one-out crossvalidation procedure, the total number of predictions for estimating classification error rates is n, one for each array. A more general crossvalidation algorithm is the V-fold crossvalidation. In the V-fold crossvalidation, the entire data set is divided into V subsets of roughly equal size and the classification method is repeated V times. Each time, the prediction rule is trained on (V-1) subsets together and then the classification rule is applied to the remaining subset as the test data set. This process is iterated for all V subsets (i.e., until all n samples are classified). The error (or accuracy) rate across all V subsets is computed for each class. This process can be repeated b times with different partitions of V subsets. The average accuracy rate over the b trials is calculated.

One important goal in crossvalidation is to provide an unbiased estimate of error rates. Crossvalidation can be performed prior to the gene selection (external crossvalidation) or after gene selection (internal crossvalidation). In the internal crossvalidation the test data set is a subset of the original data set used in gene selection. Ambroise and McLachlan, (2002) argued that the external crossvalidation should be used to avoid gene selection bias in estimating the error rate of a classification algorithm. One problem with extending the gene selection to the external crossvalidation is that only the V-1 subsets of observations are used in the testing; this results in the loss of power of identifying differentially expressed genes. In the internal crossvalidation, the same selected gene set is used in each of training samples. The internal crossvalidation can be regarded as an evaluation of the performance of the selected gene

set for a classification algorithm. On the other hand, in the external crossvalidation, a new gene set is selected for each training sample set. The external crossvalidation is an evaluation of the selection procedure for a classification algorithm. The external crossvalidation is more consistent with the notion of crossvalidation using independent samples. The internal and external crossvalidations are computed to evaluate the performance of the proposed FWE approach to gene selection for sample classification.

RESULTS

Colon data

The colon tumor data set (Alon *et al*., 1999) consists of 40 tumor and 22 normal colon tissue samples on 2000 human genes with highest minimal intensity across the 62 samples. The normal and colon samples were compared using the *t*-statistic given in Materials and Methods section. The number of permutations to compute the adjusted *P*-value was 50,000. The 2000 genes were ranked according the adjusted *P*-values for gene selection.

Several researchers have applied different gene selection and/or classification procedures to this data set. In the evaluation of the performance of a procedure, they all used internal crossvalidation to estimate the prediction error. Alon *et al*. (1999) clustered the 62 samples into two clusters. Three normal tissues (n8, n12, n34) and five tumor tissues were misclustered (T2, T30, T33, T36, T37). Using the top 1000 *t*-test–based ranked genes and an SVM approach with the leave-one-out crossvalidation, Furey *et al*. (2000) misclassified six tissues (T30, T33, T36) and (n8, n34, n36). Using top 35 entropy-based ranked genes and a (prediction by collective likelihood) classification algorithm, Li and Wong (2002) misclassified the five tissues (T28, T33) and (n1, n2, n39). Other researchers proposed multivariate methods to select an optimal subset of genes that have the smallest error with respect to a particular classification algorithm (Li *et al*., 2001; Xiong *et al*., 2001; Hellem Bø and Jonassen, 2002). Except for two and three genes, these methods typically use Monte Carlo simulation to select a "near optimal' gene set. The smallest error rate reported was 6.5% with only three genes. However, most reported error rates were about 10% or above.

We considered the FWE = 0.01, 0.05, and 0.10 for a cutoff to select a discriminatory gene set. Four different partitions were considered for V-fold crossvalidation, V = 2, 5, 10, and 62. For the 2-, 5- and 10-fold crossvalidations, we repeated $b = 100$ times. Note that the 62-fold (leave-one-out) crossvalidation consists of only one partition; therefore, one iteration ($b = 1$) is sufficient. For example, in the 10-fold crossvalidation, we randomly divided the 62 samples into 10 groups with the sizes (6, 6, 6, 6, 6, 6, 6, 6, 7, 7). The classification started with the first six samples as the test set and the remaining 56 samples as the training set. The process was iterated until the final seven samples were a test set with the first 55 samples as a training set. The whole process was repeated 100 times. The average error rate was estimated. Tables 1 and 2 show the error rate estimates of the the 1 − NN and SVM algorithms for the internal and external crossvalidations, respectively. It appears that the SVM outperforms the 1 − NN in all cases. Our discussion mainly focuses on the results from the SVM method.

The internal crossvalidation estimates the error rates for the selected gene set. The number of genes selected, denoted by *r*, were 12, 27, and 44 for $\alpha = 0.01$, 0.05, and 0.10, respectively. On average, $r = 12$ gives the smallest error rates; the error rates for $r = 27$ and 44 are similar. Except for V = 2, the number of partitions (folds) does not appear to have much effect on the error estimates. Different fold partitions represent different numbers of samples to establish the discrimination rule. In general, a highly discriminatory gene set (e.g., $r = 12$) is expected to have the predictive accuracy be positively correlated with the number of fold (V) for the crossvalidation. On average, these estimates are consistent with the results of about 10% reported by many works discussed above. We have the following findings with regard to which samples were misclassified using the leave-one-out crossvalidation. Five samples were misclassified (T30, T33, T36, n34, n36) for $r = 12$, six samples were misclassified (T2, T30, T33, T36, n34, n36) for $r = 27$, and seven samples were misclassified (T2, T30, T33, T36, n1 2, n34, n36) for $r = 44$. Note that six of seven misclassified samples (T2, T30, T33, T36) and (n12, n34) were among the eight samples misclassified by Alon *et al*. (1999). The normal sample n36 was also misclassified by Furey *et al*. (2000). In summary, the selected gene sets from the proposed procedure appears to perform better than or comparable to several results reported in the literature using univariate analysis without performing multivariate search.

The external crossvalidation estimates the error rates for a selection procedure. Gene selection and classification rule are carried out in each training set. Because of the loss of power the number of genes selected will decrease as the number of partitions decreases. For example, for $\alpha = 0{:}01$, the average number of genes selected are 12.2, 8.5, and 5.5 for V = 62, 10, and 5, respectively. We do not report the results for V = 2,

TABLE 1. ERROR RATE ESTIMATES OF THE INTERNAL CROSSVALIDATION FOR THE COLON TUMOR DATA USING THE 1-NN AND SVM CLASSIFICATION ALGORITHMS

		1-NN classification				*SVM classification*			
α	n	*Two-fold*	*Five-fold*	*10-fold*	*62-fold*	*Two-fold*	*Five-fold*	*10-fold*	*62-fold*
0.01	12	0.1552	0.1618	0.1597	0.1613	0.1077	0.0981	0.0832	0.0806
0.05	27	0.1618	0.1837	0.1937	0.1935	0.1179	0.1019	0.1271	0.0968
0.10	44	0.1594	0.1806	0.1897	0.1935	0.1242	0.1065	0.1011	0.1129
1.00	2000	0.2753	0.2745	0.2795	0.2742	0.1744	0.1511	0.1411	0.1613

TABLE 2. ERROR RATE ESTIMATES OF THE EXTERNAL CROSSVALIDATION FOR THE COLON TUMOR DATA USING THE 1-NN AND SVM CLASSIFICATION ALGORITHMS

	Average number of genes			*1-NN classification*			*SVM classification*		
α	*Five-fold*	*10-fold*	*62-fold*	*Five-fold*	*10-fold*	*62-fold*	*Five-fold*	*10-fold*	*62-fold*
0.01	5.5	8.5	12.2	0.2724	0.2216	0.1774	0.2365	0.1815	0.1290
0.05	16.3	22.4	26.8	0.1966	0.1731	0.1613	0.1811	0.1565	0.1452
0.10	26.1	33.9	42.3	0.1856	0.1671	0.2097	0.1692	0.1576	0.1290

since it performs poorly. The error rates estimated from the external crossvalidation are considerably higher than those from the internal crossvalidation. The two factors that contribute to the higher error estimates in the external crossvalidation are: (1) loss of power due to a reduced sample size in the V-fold crossvalidation, and (2) selection variation due to failing to identify some critical genes or identifying some irrelevant genes that mask the performance. In the leave-one-out partition, the numbers of genes selected by the external and the internal crossvalidations are similar; the differences between the internal and external crossvalidations can be attributable to the selection variation. But in the fivefold and 10-fold crossvalidation, the differences are attributable to both the loss of power and selection variation.

Ambroise and McLachlan (2002) compared the difference in error estimates between the internal and external crossvalidations using an SVM method with linear kernel and a backward elimination gene selection procedure. They reported error estimates from the internal and external 10-fold crossvalidation for the selected number of genes $r = 2^k$, $k = 1, \ldots, 11$. The estimated internal error rates were below 5% for $k = 2, \ldots, 9$. For external crossvalidation, all estimated error rates were well above 15%; the lowest estimate, 17.5%, occurred for a subset of $64 = 2^6$ genes. Our smallest error estimate was 15.65% for $\alpha = 0{:}05$. However, it should be noted that Ambroise and McLachlan (2002) used only 31 samples in the analysis.

With regard to the k-NN method, except for larger prediction errors compared to the SVM method, the results from k-NN are consistent with the results from SVM. However, an interesting difference is that for $\alpha = 0{:}05$, the error estimates between internal and external crossvalidation are comparable. Finally, for $r = 2000$, without the gene selection process, it gives substantially large error predictions.

Internal and external crossvalidations with respect to the number of the selected set of genes (the approach of selecting a fixed number of genes) were further evaluated. Selection of a fixed number of genes can be based on the ranking of unadjusted P-values, since the adjustment of P-values does not affect their ranking. The selection of a fixed number of genes is asymptotically equivalent to using the unadjusted P-values to determine a cutoff. For example, the selection of 100 genes out of 2000 genes corresponds to the nominal level of $0.05 = 100/2000$. We evaluated the following sizes: $r = 10$; $20 \sim 100$, 500, and 2000. Five hundred genes of the smallest P-values were considered by Alon *et al*. (1999).

Figures 1 and 2 are the plots of the averaged error estimates from the internal and external crossvalidations for 1-NN and SVM, respectively. The results are consistent with the results of Tables 1 and 2. The error rates from the internal crossvalidation are much larger without gene selection ($r = 2000$). For the 1-NN method, Figure 1 shows that except for $r = 10$, it does not appear that the external crossvalidation will give larger error predictions than the internal crossvalidation. For the SVM method, Figure 2 shows that the error estimates from the external validation do not vary much with r, between 13 to 18%. It is worthwhile to note that the objective of gene selection is to select the smallest best subset. The smallest error estimate from the external crossvalidation occurs at $r = 20$ using the SVM methods. However, the Type I error probability of the 20 selected gene is not known without computing the adjusted P-values (Tables 1 and 2).

Toxicogenomic data

This study was conducted at the Academia Sinica, Taiwan, to examine gene expression patterns with respect to metal exposures in human skin fibroblast cells using a colorimetric cDNA microarray technology (Chen *et al*., 1998). The data set consisted of control and eight different metal treatment samples for a total of 55 arrays. The entire experiment was conducted in 14 sepa-

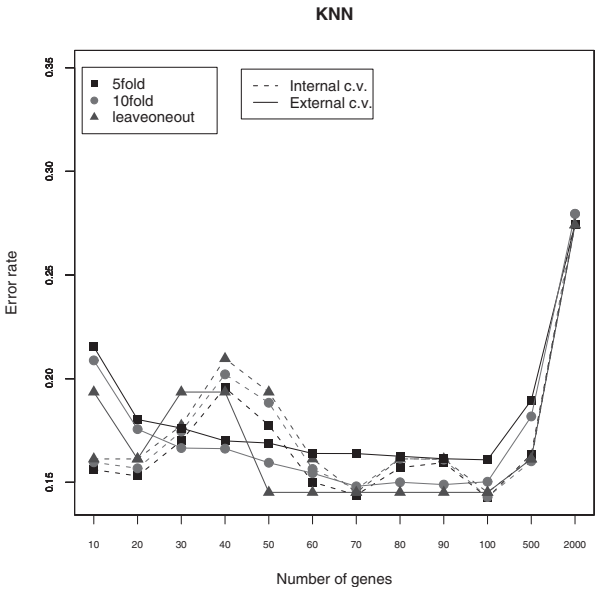

FIG. 1. Internal and external error rates of the 1-NN classification, based on 100 replications for the fivefold and 10-fold crossvalidation.

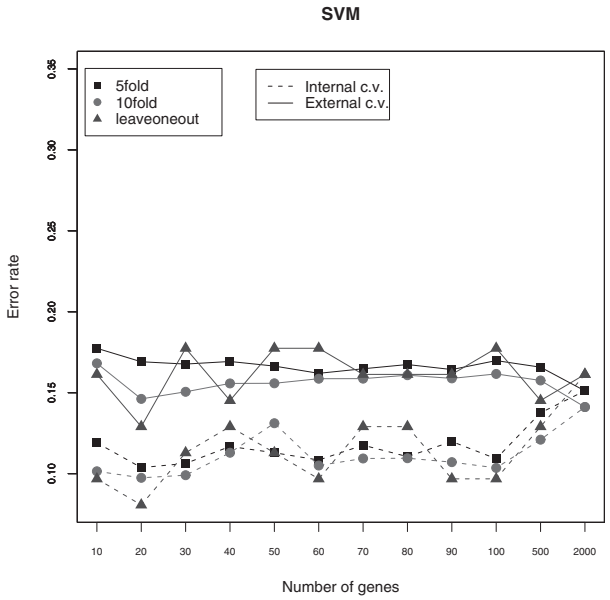

FIG. 2. Internal and external error rates of the SVM classification, based on 100 replications for the fivefold and 10-fold crossvalidation.

rate days. In each day, two to six arrays, in which one was a control array, were conducted. Thus, the control group consisted of 14 replicates. Treatments were eight different metal exposures (As, AsV, Cd, Cr, Cu, Ni, Pb, and Sb). The number of arrays for individual metals ranged from four to seven. The microarray was As-chip-TCL01 array. There were 708 genes (some genes contained duplicate or triplicate spots) on each array. Sixteen housekeeping genes and eight plant (control) genes were spotted to monitor for nonspecific background binding. These 24 genes were excluded from the analysis. The final data set contained 684 genes (spots). The intensity data were first log-transformed (base 2) to stabilize the variance and adjust the extremes. The range of the log-transformed intensities was between 0 and 16. All analyses were performed in the log base 2 scale. We applied the generalized additive model (Tsai *et al*., 2004) to adjust for the effects due to different hybridization dates.

The normalized intensity data were analyzed using the analysis of variance (ANOVA) F-test to determine an overall difference in expression levels among the nine (control and eight metals) treatment groups. The adjusted *P*-values were computed based on the method described earlier. An adjusted *P*-value less than or equal to 0.05 was used as a cutoff to select the significant gene set, denoted by Ω_F. The number of genes selected was 48. The ANOVA *F*-test is a global test; it does not address the questions of individual treatment effects. Alternatively, the one-versus-all (OVA) test was used to compare each group with the remaining eight groups for all nine groups. In each comparison, the OVA test identified a set of genes that were significantly different between the tested group and the average of the remaining groups. Using a cutoff of 0.05, nine sets of significance genes were obtained, denoted as T_i for $i = 1, \ldots, 9$. Each gene set contained treatment-specific marker genes that are different from the average of the remaining groups. The union of the nine sets T_is, $\Omega_T = \cup_{i=1}^{9} T_i$, consisted of all genes that distinguish each single group from the remaining groups. However, the gene set Ω_T may not control the FWE at 0.05, since nine comparisons were performed. The total number of genes was 32. The numbers of genes in T_is were 21, 2, 0, 4, 1, 0, 3, 0, and 4 for Control, As, AsV, Cd, Cr, Cu, Ni, Pb, and Sb, respectively. Note that there were genes that were significant in two or more groups. The gene Hs.72984 was significant in Control, Cd, and Sb groups simultaneously, and the gene Hs.80464 was significant in both Control and Cd groups. There are no significant genes for the, AsV, Cu, and Pb groups. This can affect the performance of Ω_T.

The performance of the gene sets Ω_F, Ω_T, and Ω_A were evaluated, where Ω_A consists of all 684 genes. Prediction accuracy for each gene set was computed using the 1-NN and SVM clas-

TABLE 3. ESTIMATED ACCURACY RATES (%) OF THE INTERNAL AND EXTERNAL USING THE 11-FOLD (CV11) AND LEAVE-ONE-OUT (CV55) CROSSVALIDATIONS FOR THE TOXICOGENOMIC DATA USING THE SVM MULTICLASS CLASSIFICATION ALGORITHM

		Internal crossvalidation				*External crossvalidation*				*All genes*	
		CV11		*CV55*		*CV11*		*CV55*		*CV11*	*CV55*
Metal	n	Ω_T	Ω_F	Ω_T	Ω_F	Ω_T	Ω_F	Ω_T	Ω_F	Ω_A	Ω_A
Ctrl	14	99.9	100	100	100	96.5	98.6	92.9	100	82.6	100.0
As	7	84.7	96.9	85.7	100	75.0	81.5	57.1	71.4	35.6	0.
AsV	5	79.6	83.6	80.0	80.0	71.6	63.9	80.0	80.0	36.8	0.
Cd	6	97.3	77.4	100	66.7	79.0	63.2	83.3	66.7	13.6	0.
Cr	5	59.4	88.9	60.0	100	47.4	40.3	60.0	60.0	0.3	0.
Cu	5	77.0	76.8	80.0	80.0	75.2	56.9	80.0	60.0	20.3	0.
Ni	4	68.5	81.2	75.0	100	49.8	65.5	75.0	75.0	12.4	0.
Pb	4	36.3	39.9	50.0	50.0	42.3	44.0	50.0	50.0	12.7	0.
Sb	5	97.0	78.0	100	80.0	54.4	69.6	80.0	80.0	40.9	0.
Total	55	82.9	84.8	85.5	87.3	72.0	71.3	76.4	76.4	37.8	25.5
No. of genes		32	48	32	48	24.4[a]	30.4[a]	29.6[a]	43.0[a]	684	684

The crossvalidation 11-fold accuracy was obtained by the average over 100 random split into training and test sets.
[a]Average number of gene selected over all partitions.

sification algorithms with 11-fold and leave-one-out crossvalidation. The 11-fold crossvalidation was repeated $b = 100$ times. The total number of misclassified samples from all trials was calculated for every class in each iteration. The average accuracy rate was calculated for each class. Table 3 shows the prediction results from the three gene sets using the SVM classification. The results from the 1-NN were similar.

Except for the set Ω_A, the internal crossvalidation has the overall accuracy rates above 80% and the external crossvalidation has the overall accuracy rates above 70%. The number of genes selected from the internal crossvalidation are 32 and 48 for Ω_T and Ω_F, respectively. Ω_F has slightly higher accuracy rates than Ω_T. The average numbers of genes selected from the external 11-fold crossvalidation are 24.4 and 30.4 for Ω_T and Ω_F, respectively, while the average numbers from the leave-one-out crossvalidation are 29.6 and 43.0. The lower accuracy rates for the 10-fold crossvalidation, compared to the leave-one-out crossvalidation, can be due to failure to identify critical discriminatory genes. In the external crossvalidation, the two sets Ω_T and Ω_F perform equally. The accuracies of individual classifications for the Pb samples are low. This may not be a surprise, since there is no discriminatory gene for the Pb class.

DISCUSSION

DNA array technology has been applied to sample clustering and sample prediction by many researchers. Because of a large number of genes measured, selection of an appropriate number of discriminatory genes from the original gene set is critical to the accuracy of the clustering and prediction. Many investigators have proposed different gene selection methods. The gene selection methods can be grouped into two approaches: univariate and multivariate analysis. The univariate analysis is the most commonly used approach (including the present approach). This approach examines one gene at a time; it ranks all genes according to discriminatory measures and selects a subset of top-ranked genes to be used for classification. This approach does not take multigene correlation into account. The multivariate analysis approach examines the joint discriminative ability of several genes based on a specific scoring criterion. The objective is to find the best subset among all possible subsets. Typically, hundreds of thousands of subsets of discriminative genes are evaluated with a classification algorithm. After that, an optimal subset is determined. This approach is limited to a very small number of genes in a subset, since exhaustive search is computationally prohibitive. Related approaches include the stepwise (forward or backward) selection method to obtain the "optimal" subset for a fixed r. The major problem of the multivariate approach is that the number of genes in the subset considered needs to be known *a priori*.

Dimension reduction is a closely related problem to the gene selection. Dimension reduction techniques define a smaller number of hybrid genes that are a composite of the original genes. These hybrid genes are chosen to provide independent information about different samples. Principal component analysis and multidimensional scaling (MDS) are the two most recognized dimension reduction methods for microarray data analysis. Recently, Nguyen and Rocke (2002) proposed a two-step procedure: dimension reduction using partial least squares and classification using logistic discrimination for tumor classification. They used partial least squares to select three gene components. One problem with the dimension reduction approach is that the interpretation of the gene components is often difficult.

The performance of a sample classification procedure depends on the gene selection method, the number of selected genes, and the classification method. Regardless of the selection method, different numbers of genes selected will give different classification results. There is no theoretical estimation of the optimal number of selected genes even for a given specific classification algorithm on a particular application. The optimal gene set may depend on a classification algorithm, and can vary from data to data. It is not feasible to come up with a general procedure to determine the optimal gene set combined with a classification algorithm that gets the best accuracy.

Selection of discriminatory genes can be independent of the classification algorithms. The performance of different classification methods can be evaluated under the same selection method. Selection of genes often attempts to identify a minimum number of genes that are useful (Raychaudhri *et al*., 2001). In this paper we propose using the FWE approach to determine the number of genes for sample classification. The FWE approach ensures the selection of the minimum number of differentially expressed genes such that each selected gene is truly positive with a confidence of $(1-\alpha)$ probability. However, because of the stringent criterion imposed in the selection, it may not select any gene. The classification rule then becomes a random assignment. On the other hand, in a study to develop genetic profiles, many genes that might be involved in complex functional relationships with other genes might have moderate differen-tial expressions between experimental samples. These genes would have larger P-values. This application requires a procedure to select a large number of potentially differentially expressed genes involved in gene regulation. For this application, the proposed FWE approach cannot completely reveal this information.

REFERENCES

ALON, U., BARKAI, N., NOTTERMAN, D.A., GISH, K., YBARRA, S., MACK, D., and LEVINE, A.J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc. Natl. Acad. Sci. USA **96,** 6745–6750.

AMBROISE, C., and MCLACHLAN, G.J. (2002). Selection bias in gene extraction on the basis of microarray gene-expression data. Proc. Natl. Acad. Sci. USA **99,** 6562–6566.

BURCZYNSKI, M.E., MCMILLIAN, M., CIERVO, J., LI, L., PARKER, J.B., DUNN II, R.T., HICKEN, S., FARR, S., and JOHNSON, M.D. (2000). Toxicogenomics-based discrimination of toxic mechanism in HepG2 human hepatoma cells. Toxicol. Sci. **bf 58,** 399–415.

CHEN, J.J., WU, R., YANG, P.C., HUANG, J.Y., SHER, Y.P., HAN, M.H., KAO, W.C., LEE, P.J., CHIU, T.F., CHANG, F., et al. (1998). Profiling expression patterns and isolating differentially expressed genes by cDNA microarray system with colorimetry detection. Genomics **51,** 313–324.

DUDOIT, S., FRIDLYAND, J., and SPEED, T.P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. J. Am. Stat. Assoc. **97,** 77–87.

EISEN, M.B., SPELLMAN, P.T., BROWN, P.O., and BOTSTEIN, D. (1998). Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. USA **95,** 14863–14868.

FIX, E., and HODGES, J.L. (1951). Discriminatory analysis: Nonparametric discrimination: Consistency properties. Technical Report Project 21-49-004, Report Number 4, USAF School of Aviation Medicine, Randolf Field, TX.

FUREY, T.S., CHRISTIANINI, N., DUFFY, N., BEDNARSKI, D.W., SCHUMMER, M., and HAUSSLER, D. (2000). Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics **16,** 906–914.

GOLUB, T., SLONIM, D., TAMAYO, P., HUARD, C., GASSENBEEK, M., MESIROV, J., COLLER, H., LOH, M., DOWNING, J., CALIGIURI, M., et al. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science **286,** 531–537.

HELLEM BØ, T. and JONASSEN, I. (2002). New feature subset selection procedures for classification of expression profiles. Genome Biol. **3,** research0017.1–0017.11.

HOCHBERG, Y., and TAMHANE, A.C. (1987). *Multiple Comparison Procedures* (John Wiley & Sons, New York).

LI, J., and WONG, L. (2002). Identifying good diagnostic gene groups from gene expression profiles using the concept of emerging patterns. Bioinformatics **18,** 725–734.

LI, L., WEINBERG, C.R., DARDEN, T.A., and PEDERSEN, L.G. (2001). Gene selection for sample classification based on gene expression data: Study of sensitivity to choice of parameters of the GA/KNN method. Bioinformatics **17,** 1131–1142.

LIU, H., LI, J., and WONG, L. (2002). A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns. Genome Informat. **13,** 51–60.

NGUYEN, D.V., and ROCKE, D.M. (2002). Multi-class cancer classification via partial least squares with gene expression profiles. Bioinformatics **18,** 1216–1226.

RAMASWAMY, S., TAMAYO, P., RIFKIN, R., MUKHERJEE, S., YEANG, C.-H., ANGELO, M., LADD, C., REICH, M., LATULIPPE, E., MESIROV, J., *et al*. (2001). Multiclass cancer diagnosis using tumor gene expression signatures. Proc. Natl. Acad. Sci. USA **98,** 15149–15154.

RAYCHAUDHURI, S., SUTPHIN, P.D., CHANG, J.T., and ALTMAN, R.B. (2001). Basic microarray analysis: Grouping and feature reduction. Trends Biotechnol. **19,** 189–193.

TSAI, C.A., HSUEH, H., and CHEN, J.J. (2004). A generalized additive model for microarray gene expression data analysis. J. Biopharmaceut. Stat., **14,** 553–573.

VAPNIK, V. (1998). *Statistical Learning Theory* (Wiley, New York).

WESTFALL, P.H., and YOUNG, S.S. (1993). *Resampling-Based Multiple Testing*. (John Wiley & Sons, New York).

XIONG, M., LI, W., ZHAO, J., JIN, L., and BOERWINKLE, E. (2001). Feature (Gene) selection in gene expression-based tumor classification. Mol. Genet. Metab. **73,** 239–247.

Address reprint requests to:
James J. Chen, Ph.D.
NCTR/FDA/HFT-20
Jefferson, AR 72079

E-mail: jchen@nctr.fda.gov