SELECTING GENES WITH DISSIMILAR DISCRIMINATION STRENGTH FOR SAMPLE CLASS PREDICTION

ZHIPENG CAI, RANDY GOEBEL, MOHAMMAD R. SALAVATIPOUR, YI SHI, LIZHE XU, AND GUOHUI LIN*

Department of Computing Science, University of Alberta
Edmonton, Alberta T6G 2E8, Canada
Email: zhipeng, goebel, mreza, ys3, ghlin@cs.ualberta.ca, lizhe.xu@dhs.gov

One of the main applications of microarray technology is to determine the gene expression profiles of diseases and disease treatments. This is typically done by selecting a small number of genes from amongst thousands to tens of thousands, whose expression values are collectively used as classification profiles. This gene selection process is notoriously challenging because microarray data normally contains only a very small number of samples, but range over thousands to tens of thousands of genes. Most existing gene selection methods carefully define a function to score the differential levels of gene expression under a variety of conditions, in order to identify top-ranked genes. Such single gene scoring methods suffer because some selected genes have very similar expression patterns so using them all in classification is largely redundant. Furthermore, these selected genes can prevent the consideration of other individually-less but collectively-more differentially expressed genes. We propose to cluster genes in terms of their class discrimination strength and to limit the number of selected genes per cluster. By combining this idea with several existing single gene scoring methods, we show by experiments on two cancer microarray datasets that our methods identify gene subsets which collectively have significantly higher classification accuracies.

1. Introduction

DNA microarrays provide the opportunity to measure the expression levels of thousands of genes simultaneously. This novel technology supplies us with a large volume of data to systematically understand various gene regulations under different conditions. As one of the main applications, it is very important to determine the gene expression profiles of diseases and disease treatments. Among the thousands of genes in the arrays, many of them do not have their expression values distinguishably changed across different condition, e.g., so-called “house keeping” genes. These genes certainly would not be very useful in profiling since they do not contribute much to disease or treatment class recognition. In practice, a small number, typically in the tens, of genes that are highly differentially expressed across different conditions are to be selected to compose profiles for the purpose of class prediction. This process is known as *gene selection*; there are many existing methods, which typically define a function to score the level of how differentially expressed a gene is, under different conditions, and identify those top ranked genes. [1,2,3,4,5] Such single gene scoring

[*] To whom correspondence should be addressed.

methods typically suffer the problem that some selected genes have very similar expression patterns, therefore using them all in classification is largely redundant, and those selected genes prevent other individually-less but collectively-more differentially expressed genes from being selected.

Several other gene selection methods have recognized the problem with the redundancy of some highly expressed genes, and look for a subset of genes that collectively maximize the classification accuracy. For example, Xiong *et al.* define a function to measure the classification accuracy for individual genes and select a subset of genes through *Sequential Forward* [*Floating*] *Selection* (SF[F]S),[6] which was developed decades ago for general feature selection. Guyon *et al.* propose another method that uses *Support Vector Machines* (SVMs) and *Recursive Feature Elimination* (RFE).[7] In terms of effectiveness, these gene selection methods perform much better than those single gene scoring methods, since they measure the classification strength of the whole set of selected genes. Computationally, they are essentially heuristics which replace exhaustive enumeration of an optimal subset of genes which typically takes much longer to return a solution. The inefficiency of these methods actually prevent them from being used in practice. Nevertheless, there are alternative implementations of the key idea, which is to exclude a gene when there is already a similar gene selected.

We propose another implementation which first clusters genes according to their class discrimination strength, namely, two genes that have very close class discrimination strength are placed in a common cluster; we then limit the number of genes per cluster to be selected. This provides a more efficient clustering process which, when combined with a single gene scoring method, leads to an efficient and effective gene selection algorithm. We call our method an *EEGS-based gene selection method*. In the next section, we present the details of a novel measure of class discrimination strength difference between two genes, using their expression values. With this distance measure, we briefly explain how to adopt the k-*means* algorithm[8] to cluster genes. We also briefly introduce three single gene scoring methods, namely F-test[3], Cho[4] and GS,[5] and two classifiers, namely, a linear kernel SVM classifier[7] and a k Nearest Neighbor (KNN) classifier.[3] Finally, we outline a complete high level description of the EEGS-based gene selection methods. In Section 3, we briefly introduce our performance measurements, followed by the dataset descriptions, and our experimental results. Section 4 discusses parameter selection, the effects of variety within data sets, classifiers and the performance measurements, and finally, the overall results compared to single gene scoring methods. Section 5 summarizes our main contributions, our conclusions on the suitable datasets for the EEGS-based methods, and some plans for future work.

2. The EEGS-Based Gene Selection Methods

There are two challenges in microarray data classification. One is class discovery to define previously unrecognized classes. The other is to assign individual samples to already-defined classes, which is the focus here.

2.1. *The Performance Measurements*

The genes selected by a method are evaluated by their class discrimination strength, measured by the classification accuracy, defined as follows. For gene selection purposes, a number of microarray samples with known class labels are provided, which form a *training dataset*. The selected genes are then used for building a classifier, which can take a new microarray sample and assign it a class label. The set of such samples for testing purpose is referred to as the *testing dataset*, and the percentage of the correctly labeled samples is defined as the *classification accuracy* of the method (on this particular testing dataset). Note that we have to have the class labels for the samples in the testing dataset in order to calculate the classification accuracy. For computational convenience, given a microarray dataset whose samples all have known class labels, only a portion of it is used to form the training dataset; the rest of the samples have their class labels removed and are used to form the testing dataset. There are two popular cross validation schemes adopted in the literature to evaluate a method, which are *ℓ-Fold* and *Leave One Out* (LOO). We adopt the ℓ-Fold cross validation in this work, in which the whole dataset is (randomly) partitioned into ℓ equal parts and, at one time, one part is used as testing dataset and the other $\ell - 1$ parts are used as training dataset. The process is repeated for each part and the average classification accuracy over these ℓ ones is taken as the final classification accuracy. Here set $\ell = 5$ and repeat the process for 20 iterations. Therefore, the final classification accuracy is the average over 100 values. We report the 5-Fold classification accuracies for all the six tested gene selection methods in Section 3.

2.2. *The Classifiers*

We adopt two classifiers in our study. One is a linear kernel SVM classifier that has been used in Guyon *et al.*[7] and the other is a KNN classifier that has been used in Dudoit *et al.*[3] Essentially, with a given set of selected genes determined by some gene selection method, the SVM classifier, which contains multiple SVMs, finds decision planes to best separate the labeled samples based on the expression values of these selected genes. Subsequently, it uses this set of decision planes to predict the class label of a test sample. For a more detailed explanation of how the decision planes are constructed, the readers are referred to Guyon *et al.*[7] The KNN classifier predicts the label of a testing sample in a different way. Using the expression values of (only) the selected genes, the classifier identifies the k most similar samples in the training dataset. It then uses the class labels of these k similar samples through a majority vote. In our experiments, we set the value of k to be 5 as default, after testing for several values in the range 4 to 10.

2.3. *The Single Gene Scoring Methods*

Many of the existing gene selection methods are single gene scoring methods that define a function to approximate the class discrimination strength of a gene.[1,2,3,4,5] Typically, an F-test gene selection method[2,9] is presented, which basically captures the variance of the class variances of the gene expression values in the dataset. A bigger variance indicates that a gene is more differentially expressed and thus ranked higher. Because class sizes might differ a lot, Cho *et al.*[4] proposed a weighted variant, which was further refined by Yang *et al.*[5] We denote these three single gene scoring methods as F-test, Cho and GS, respectively, and combine the EEGS idea with them to have the EEGS-based methods, denoted as EEGS-F-test, EEGS-Cho and EEGS-GS, respectively.

2.4. *Gene Clustering*

Gene clustering in microarray data analysis is an independent research subject, in which genes having a similar expression pattern are clustered for certain applications. In our work here, we are particularly interested in the class discrimination strength of the genes, since we do not want to select too many genes that have similar class discrimination strength. Note that genes having a similar expression pattern would certainly have similar class discrimination strength, but the other way around is not necessarily true. Therefore, we define a new measure trying to better capture the difference in the class discrimination strength between two genes.

Assume there are p genes and n samples in the microarray training dataset, and these n samples belong to L distinct classes. Let a_{ij} denote the expression value of gene i in sample j. This way, the training dataset can be represented as a matrix $A_{p\times n} = (a_{ij})_{p\times n}$. Let $C_1, C_2, \ldots, C_L$ denote the L classes, and $n_q = |C_q|$, for $q = 1, 2, \ldots, L$. Let $\overline{a}_{iq}$ be the mean expression value of gene i in class C_q: $\overline{a}_{iq} = \frac{1}{n_q}\sum_{j\in C_q} a_{ij}$, for $q = 1, 2, \ldots, L$. The centroid matrix is thus $\overline{A}_{p\times L} = (\overline{a}_{iq})_{p\times L}$. The *discrimination strength vector* of gene i is defined as $v_i = \langle |\overline{a}_{iq_1} - \overline{a}_{iq_2}| \mid 1 \leq q_1 < q_2 \leq L\rangle$, where the order of $\frac{1}{2}L(L-1)$ vector entries is fixed the same for all genes, for example the lexicographical order. After all the discrimination strength vectors have been calculated, the k-means algorithm[8] is applied to cluster these p genes into k clusters using their discrimination strength vectors. Essentially, k-means is a centroid-based clustering algorithm that partitions the genes based on their pairwise distances. We adopt both the Euclidean distance and the Pearson correlation coefficient in our experiments. Again, we have tested several values of k in the k-means algorithm (cf. Section 4.1) and we have set it to 100 as default.

2.5. *The Complete EEGS-Based Gene Selection Methods*

Given a microarray training dataset containing p genes and n samples in L classes, an EEGS-based gene selection method first calls the k-means algorithm (with $k = 100$) to cluster genes. Next, depending on the detailed single gene scoring method integrated in the method, which is one of F-test, Cho and GS, it calls the single gene scoring method to score all the genes and sort them into non-increasing order. Using this gene order and the gene cluster information, the EEGS-based method selects a pre-specified number, x, of top ranked genes with the constraint that there are at most T genes per cluster can be selected. In more details, it scans through the gene order and picks up a gene only if there are less than T genes from the same cluster selected. These x selected genes are then fed to classifier construction, either the SVM classifier or the KNN classifier. In our experiments, we have tested x ranging from 1 to 80 and several values for T (cf. Section 3.2). We have set $T = 1$ as default (cf. Section 4.1).

Depending on the single gene scoring method integrated into the EEGS-based gene selection method, which is one of F-test, Cho and GS, the method is referred to as EEGS-F-test, EEGS-Cho and EEGS-GS, respectively.

3. Experimental Results

We compare the three EEGS-based gene selection methods with the three ordinary gene selection methods, measured by the 5-Fold cross validation classification accuracy. Note that we have adopted two distance measures in the k-means clustering algorithm. We have a broader collection of experimental results, but here report only those based on the Euclidean distance, as there is essentially no difference between the results based on the Pearson correlation coefficient (cf. Section 4.1). Note also that we have adopted two classifiers, a linear kernel SVM classifier and a KNN classifier. We choose to plot their classification accuracies together labeled by different notations, for instance, EEGS-Cho-KNN labels the accuracies of the KNN classifier. The experiments are done on two real cancer microarray datasets, CAR[10] and LUNG,[9] whose details are described in the following subsection.

3.1. *Dataset Descriptions*

The CAR dataset contains 174 samples in eleven classes: *prostate*, *bladder/ureter*, *breast*, *colorectal*, *gastroesophagus*, *kidney*, *liver*, *ovary*, *pancreas*, *lung adenocarcinomas*, and *lung squamous cell carcinoma*, which have 26, 8, 26, 23, 12, 11, 7, 27, 6, 14, and 14 samples, respectively.[10] Each sample originally contained 12,533 genes. We preprocessed the dataset as described in Su *et al.*[10] to include only those probe sets whose maximum hybridization intensity in at least one sample is ≥ 200; Subsequently, all hybridization intensity values ≤ 20 were raised to 20, and the values were log transformed. After preprocessing, we obtained a dataset of 9,182 genes.

The LUNG dataset[9] contains in total 203 samples in five classes: *adenocarcinomas*, *squamous cell lung carcinomas*, *pulmonary carcinoids*, *small-cell lung carcinomas* and *normal lung*, which have $139, 21, 20, 6, 17$ samples, respectively. Each sample originally had 12,600 genes. A preprocessing step which removed genes with standard deviations smaller than 50 expression units, produced a dataset with 3,312 genes. [9]

3.2. *Cross Validation Classification Accuracies*

The classification accuracies reported here were obtained under the default setting which uses Euclidean distance, $k = 100$ in the k-means clustering algorithm, and at most $T = 1$ gene per cluster could be selected. On each of the two datasets, all six gene selection methods, F-test, Cho, GS, EEGS-F-test, EEGS-Cho, and EEGS-GS, were run and the 5-Fold cross validation classification accuracies were collected and plotted in Figure 1.

Obviously, these plots show that regardless of which cross validation scheme and which classifier were used, the classification accuracies of the EEGS-based gene selection methods were significantly higher than that of their non-EEGS-based counterparts. Typically, on the CAR dataset, the classification accuracies of the EEGS-based methods were significantly higher — though the difference between the classification accuracies became smaller with the increasing number of selected genes, it remained to be more than 10%. From Figure 1, among the single gene scoring methods, another observation is that the GS method performed better than the Cho method and the F-test method. The EEGS-based methods had the same performance tendency on the CAR dataset. On the LUNG dataset, similar results were obtained, although the performance differences between the EEGS-based methods and the non-EEGS-based methods were smaller than those on the CAR dataset.

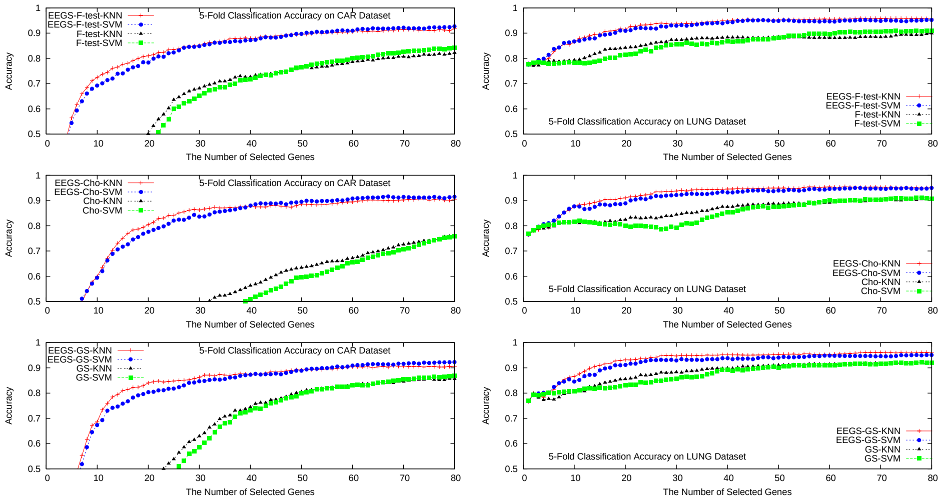

Figure 1. The 5-Fold classification accuracies of the six gene selection methods on the CAR and LUNG datasets.

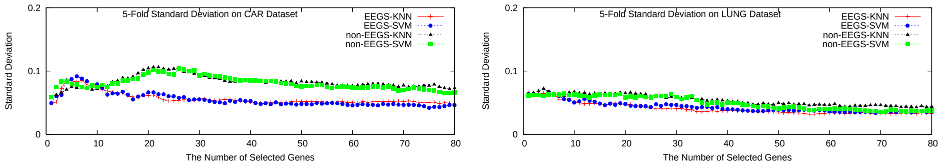

Figure 2. Plots of average standard deviations of the 5-Fold classification accuracies of the EEGS-based and the non-EEGS-based gene selection methods, combined with the KNN and the SVM classifiers, on the CAR dataset (left) and LUNG dataset (right), respectively.

We have also calculated the standard deviations of the 5-Fold cross validation classification accuracies. Note that the accuracies plotted in Figure 1 were averages over 100. Figure 2 plots the average standard deviations of the EEGS-based methods and the non-EEGS-based methods, on the CAR and the LUNG datasets, respectively. Namely, the EEGS-KNN plot records the average standard deviations of the three EEGS-based methods (EEGS-F-test, EEGS-Cho and EEGS-GS) combined with the KNN classifier, and the non-EEGS-SVM plot records the average standard deviations of the three non-EEGS-based methods (F-test, Cho and GS) combined with the SVM classifier, and so on. These results show that the standard deviations of the classification accuracies of the EEGS-based methods were even smaller than those of the non-EEGS-based methods, indicating that the EEGS-based methods performed more consistently. The statistical significance of the outperformance of the EEGS-based methods over the non-EEGS-based methods was done and the p values in the Analysis of variance (ANOVA) was always less than 0.001. (For the complete results, the readers might refer to supplementary materials at http://www.cs.ualberta.ca/~ghlin/src/WebTools/cgs.php.)

4. Discussion

4.1. *Gene Clustering*

We adopted the k-means algorithm for gene clustering, in which k, the number of expected clusters, has to be set beforehand. Obviously, the value of k will affect the sizes of resultant clusters, and therefore will affect T ultimately, which is the maximum number of genes per cluster to be selected. We chose to empirically determine these two values. To this end, we experimented with 15 values for k: from 10 to 150 (in the tens), and five values for T: 1, 2, 3, 4 and 5. All three EEGS-based methods combined with two classifiers were tested on the CAR dataset, under the 5-Fold cross validation, for each combination of k and T. Associated with each value of T, a classification accuracy is defined as the mean value of $100 \times 3 \times 2 \times 15 = 9,000$ values, where there are 100 runs in the 5-Fold cross validation, 3 EEGS-based methods, 2 classifiers, and 15 values of k in the test. These classification accuracies, with respect to the number of selected genes, are plotted in Figure 3 (left), where $T = 1$ clearly performed the best. Similarly, associated with each value of k, a classification accuracy is defined as the mean value of $100 \times 3 \times 2 \times 5 = 3,000$ values, where there are 5 values of T in the test. Again, these classification accuracies, with respect to the number of selected genes, are plotted in Figure 3 (right), where we can see that the value of k didn't really affect the performance. Since we decided to set the maximum number of selected genes to be 80, we determined to set $k = 100$ and $T = 1$ as default.

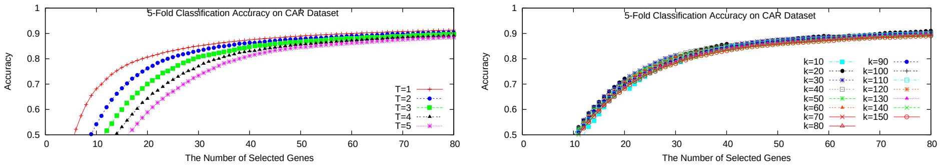

Figure 3. The effects of the number of clusters in gene clustering and the maximum number of genes per cluster can be selected.

Another important factor in gene clustering that might affect its performance is the

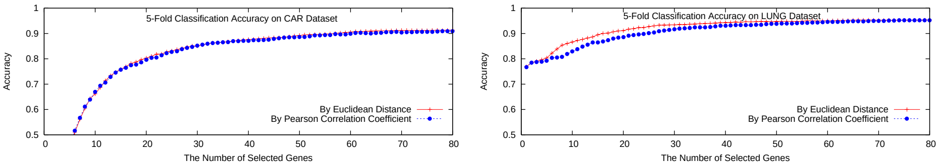

Figure 4. The effects of the Euclidean distance and the Pearson correlation coefficient in gene clustering.

distance measure, for which the Euclidean distance and the Pearson correlation coefficient are two most commonly adopted ones. We have experimented with both of them in the k-means clustering algorithm on the CAR and the LUNG datasets. With the default setting for k and T, we collected the 5-Fold classification accuracies which are the mean values of $100 \times 3 \times 2 = 600$ values and plotted them in Figure 4. It can be clearly seen that, the detailed distance measure did not seem to affect the overall performance of the EEGS-based methods, in terms of their classification accuracy. Therefore, we chose the Euclidean distance as our default setting.

4.2. *Datasets*

Note that in the EEGS-based gene selection methods, a discrimination strength vector is computed for every gene, and genes are clustered using the Euclidean distance defined on their discrimination strength vectors. The main intention for such clustering is to limit the number of genes that have very similar class discrimination strength to be selected, and thus to provide space for other individually-less but collectively-more differentially expressed genes to participate in the class prediction. This goal would not be achieved when there are only two classes in the dataset (binary classification), which would mean that the discrimination strength vectors have only one entry and the EEGS-based method reduces to its component basic gene selection method. For similar reasons, we suspect that the EEGS-based gene selection methods would work well when the number of classes in the dataset is three. The CAR and the LUNG datasets contain eleven and five classes, respectively, and therefore the discrimination strength vectors have 55 and 10 entries, respectively. The EEGS-based gene selection methods all performed excellent on them.

For various reasons, microarray datasets are often imbalanced, that is, the sizes of the classes are highly variable. For example, in the LUNG dataset, the maximum class size is 139 while the minimum class size is only 6. Since it is possible that during the 5-Fold cross validation the random partition produces a training dataset containing only a few samples, or maybe even none, from a small class, the testing would make mistakes on the samples from the same class. To verify how much the dataset unbalance would affect the performance of a gene selection method, we removed the classes of sizes smaller than 10 from the CAR and the LUNG datasets to produce two reduced but more balanced datasets, denoted as CAR^r and LUNG^r, respectively. Consequently, the CAR^r and the LUNG^r datasets contain 153 samples in 8 classes and 197 samples in 4 classes, respectively. We then ran all six methods combined with both the KNN classifier and the SVM classifier on the full and the reduced datasets, and plotted the average classification accuracies (each over three methods with two classifiers, i.e., six values) in Figure 5. In the figure, one can see that the performance of the EEGS-based methods did not change a lot on the reduced CAR^r and the LUNG^r datasets, compared with their performance on the full datasets. Interestingly, for the non-EEGS-based methods, their performance increased significantly on the CAR^r dataset, but not on the LUNG^r dataset. Nevertheless, these results show that the EEGS-based methods performed more stable (and better) than the non-EEGS-based methods on imbalanced datasets. One of the possible reasons is that the EEGS-based methods might be able to select some genes that are signatures of the samples in the small classes, for which further studies are needed to understand better.

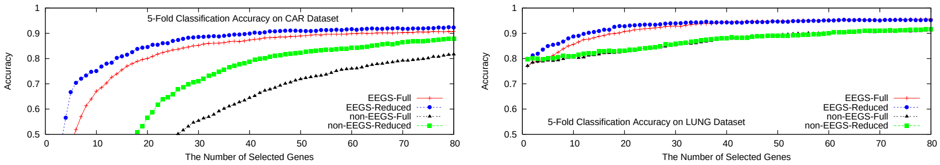

Figure 5. Classification accuracies of the EEGS-based and the non-EEGS-based methods on the full and reduced datasets, where EEGS-Full plots the average classification accuracies of the EEGS-based methods on the full dataset.

Our further statistics on the classification *precision* and *recall* for each class in the 5-Fold cross validation shows that they seemed independent of the class size (data not shown but can be accessed through http://www.cs.ualberta.ca/˜ghlin/src/WebTools/cgs.php).

4.3. *A Case Study on the CAR Dataset*

The EEGS-based gene selection methods are designed to not select too many genes having similar class discrimination strength, so as to consider individually-less but collectively-more differentially expressed genes. In this sense, some selected genes might be lower in the gene order, but have strength to discriminate some classes on which its preceding genes might not do well. To examine if this indeed happened, we collected more detailed results on the first 10 genes selected by F-test and EEGS-F-test, respectively. The 5-Fold cross validation classification accuracies of the SVM classifier built on the first x genes were also collected, for $x = 1, 2, \ldots, 10$. We summarized the results in Table 1, in which column 'Probe Set' records the probe set (gene) id in the CAR dataset, column 'R' records the rank of the gene in the gene order by F-test, and column 'Accuracy' records the classification accuracy of the gene subset up to the gene at the row.

Note that the third gene (probe set) selected by EEGS-F-test, *765_s_at*, has a rank 17, which was thus not selected by F-test. The classification accuracy of the top 10 genes selected by F-test was only 30.63%, while adding the third gene *765_s_at* in EEGS-F-test lifted the classification accuracy to 42.18%, already significantly higher than 30.63%. On average, the contribution of each gene, except the first, selected by EEGS-F-test was 6.10% in terms of classification accuracy; the contribution of each gene, except the first, selected by F-test was only 1.23%. These figures suggested that when the number of selected genes was fixed, the genes selected by EEGS-F-test had much higher class discrimination strength compared to the genes selected by F-test.

Table 1. The first 10 genes selected by the EEGS-F-test and the F-test methods on the CAR dataset, respectively, and the respective 5-Fold cross validation classification accuracies of the SVM classifiers built on the genes. Column 'R' records the rank of the gene in the gene order by F-test, and column 'Accuracy' records the classification accuracy of the gene subset up to the gene at the row.

EEGS-F-test-SVM			F-test-SVM		
Probe Set	R	Accuracy	Probe Set	R	Accuracy
40794_at	1	19.54%	40794_at	1	19.54%
41238_s_at	4	28.91%	660_at	2	25.46%
765_s_at	17	**42.18**%	32200_at	3	25.40%
1500_at	21	60.52%	41238_s_at	4	30.00%
35220_at	24	64.54%	34941_at	5	30.69%
32771_at	27	68.62%	41468_at	6	30.52%
34797_at	38	70.75%	36141_at	7	30.12%
35194_at	50	73.56%	617_at	8	30.35%
36806_at	52	73.16%	37812_at	9	30.29%
40511_at	63	74.48%	217_at	10	30.63%

Acknowledgments. This research is supported in part by AHFMR (to LX), AICML (to RG), CFI (to GL), iCore (to YS and RG), NSERC (to RG, MS and GL) and the University of Alberta (to MS).

References

1. T. R. Golub *et al.* Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537, 1999.
2. P. Baldi and A. D. Long. A Bayesian framework for the analysis of microarray expression data: Regularized t-test and statistical inferences of gene changes. *Bioinformatics*, 17:509–519, 2001.
3. S. Dudoit, J. Fridlyand, and T. P. Speed. Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, 97:77–87, 2002.
4. J. Cho, D. Lee, J. H. Park, and I. B. Lee. New gene selection for classification of cancer subtype considering within-class variation. *FEBS Letters*, 551:3–7, 2003.
5. K. Yang, Z. Cai, J. Li, and G.-H. Lin. A stable gene selection in microarray data analysis. *BMC Bioinformatics*, 7:228, 2006.
6. M. Xiong, X. Fang, and J. Zhao. Biomarker identification by feature wrappers. *Genome Research*, 11:1878–1887, 2001.
7. I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46:389–422, 2002.
8. F. Blanchot-Jossic *et al*. Up-regulated expression of ADAM17 in human colon carcinoma: co-expression with EGFR in neoplastic and endothelial cells. *Oncogene*, 207:156–163, 2005.
9. A. Bhattacharjee *et al*. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proceedings of National Academy of Sciences of the United States of America*, 98:13790–13795, 2001.
10. A. I. Su *et al*. Molecular classification of human carcinomas by use of gene expression signatures. *Cancer Research*, 61:7388–7393, 2001.