2006 IEEE Conference on
Systems, Man, and Cybernetics
October 8-11, 2006, Taipei, Taiwan

A Novel Approach Searching for Discriminative Gene Sets

Feng Chu and Lipo Wang

***Abstract*— We propose an algorithm of searching for good discriminative gene sets (DGSs) in microarray cancer data, which we call active mining discriminative gene sets (AM-DGS). Tests in the leukemia data set and the prostate data set indicate that our method is able to achieve better accuracy with much smaller DGSs compared to 3 widely used methods, i.e., TS, FS, and SVM-RFE.**

I. Introduction

Accurate classification of homogenous cancers is a key problem for disease prognosis, treatment selection, pathology research, and drug discovery. In recent years, gene expression profiles have been extensively applied to classifying cancers at the molecular level [4], [5], [6]. A typical gene expression data set can be described as a high dimensional $n \times m$ matrix B. In B, each column stands for a cancer sample (i.e., an observation) and each row stands for a gene. Here m usually ranges from several tens to over one hundred and n usually ranges from several thousands to tens of thousands. Since n is much larger than m, it is of great importance to select a group genes for classification because of the following two points. First, among all the genes, only a part of them have discriminating power. Furthermore, some genes even act as "noise" and undermine the classification accuracy. Second, some genes are highly correlated and their expression profiles behave very similarly in classification. Excluding some of such correlated genes will reduce redundancy in the discriminative gene sets (DGS).

Since mid-1990s, a number of gene selection approaches [8], [9] have been proposed. The majority of these methods can be regarded as filter schemes [15], which rank genes first according to their discriminative ability and then select a certain number of, e.g., 20, 50, or 100, top-ranked genes for classification. Although these top-ranked genes can lead to highly accurate classification results, they may still contain great redundancy. Some other methods use wrapper scheme [15]. In [1], a support vector machine [12], [13] based recursive feature elimination method (SVM-RFE) was proposed. Although this method takes advantages of a classifier, i.e., the SVM, it is a greedy backward search scheme that requires a large amount of computing time. In [11], a method called *Markov blanket* was used to reduce redundancy in DGSs. In [7], Y. Wang *et al.* used clustering methods to identify the redundancy in DGSs and then reduce the redundancy by "collapsing dense clusters".

Feng Chu and Lipo Wang are with the College of Information Engineering, Xiangtan University, Xiangtan, Hunan, China and also with the School of Electrical and Electronic Engineering, Nanyang Technological University, Block S1, Nanyang Avenue, 639798, Singapore.

Here we propose a simple yet very effective and efficient method of searching for DGSs that lead to high classification accuracy. Our method is a top-down forward wrapper search scheme, which is much computationally efficient than the SVM-RFE scheme [1] and is able to greatly reduce the redundancy of DGSs.

The rest of this paper is organized as follows. In Section II, we introduce our SVM-based method for searching DGSs, i.e., active mining discriminative gene sets (AM-DGS), and its related techniques. In Section III, we apply our SVM-based AM-DGS algorithm to a well-known benchmark gene expression data set, i.e. the leukemia data set [4]. In Section IV, we discuss our results and conclude the paper.

II. Active Mining Discriminative Gene Sets

Recently, *active learning* has attracted much attention in machine learning [2]. An active learner, AL, has three components $\{X, F, Q\}$. Here X is the input matrix. F is the mapping function from input space to output space that describes the objective (or function) of the AL. Q is a query function that is used to determine the sequence of unlabelled samples to be learned by the AL.

In almost all the active learning approaches proposed to date, the function Q is used to search for the unlabelled *samples*, i.e., *observations*, to be learned by the AL. In the following parts of this section, we will propose a learning scheme with a query function $\tilde{Q}$ that is used to search for *features* (i.e., *genes* in this application). Hence we call our algorithm *active mining* as opposed to *active learning*.

A. T-statistic

In the first step of our scheme, we rank all the features (genes) according to their t-statistics (TSs). The TS of gene i is defined as follows [10].

$$TS_i = |\frac{\overline{x}_{c1} - \overline{x}_{c2}}{s_{pi}\sqrt{1/n_1 + 1/n_2}}| \tag{1}$$

where

$$\overline{x}_{c1} = \sum_{j \in C_1} \overline{x}_{ij}/n_1 \tag{2}$$

$$\overline{x}_{c2} = \sum_{k \in C_2} \overline{x}_{ik}/n_2 \tag{3}$$

$$s_{pi}^2 = \frac{\sum_{j \in C_1}(x_{ij} - \overline{x}_{c1})^2 + \sum_{k \in C_2}(x_{ik} - \overline{x}_{c2})^2}{n_1 + n_2 - 2} \tag{4}$$

There are 2 classes, i.e., C_1 and C_2, which include n_1 and n_2 samples, respectively. x_{ij} and x_{ik} are the expression values

of gene i in C_1 and C_2, respectively. $\overline{x}_{c1}$ and $\overline{x}_{c2}$ are the mean expression values of C_1 and C_2. s_{pi} is the pooled standard deviation of gene i.

B. Seeds

After ranking all the genes with TSs, the gene with the largest TS is selected as the first feature in the discriminative gene set (DGS). We call this first feature a *seed*.

C. Support Vector Machines

We use support vector machines (SVMs) [12] [13] as our classifier, i.e., we input our DGS into an SVM to carry out training and classification. A standard SVM classifier aims to solve the following problem. Given l training vectors $\{\mathbf{x}_i \in R^n, i = 1, ..., l\}$ that belong to two classes, with desired output $y_i \in \{-1, 1\}$, find a decision boundary:

$$\mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}_i) + b = 0, \tag{5}$$

where $\mathbf{w}$ is the weight vector and b is the bias. $\boldsymbol{\phi}(\mathbf{x}_i)$ is the function that maps $\mathbf{x}_i$ to a potentially much higher dimensional feature space. This decision boundary is determined by minimizing the cost function:

$$\psi = \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{l}\xi_i, \tag{6}$$

subject to:

$$y_i(\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_i) + b) \geq 1 - \xi_i, \tag{7}$$

$$\xi_i \geq 0. \tag{8}$$

where $\{\xi_i, i = 1, 2, ..., l\}$ are slack variables and C is a constant that determines the tradeoff between the training error and the generalization capability of the SVM. This optimization problem has a quadratic programming (QP) dual problem:

$$\text{maximize: } Q(\alpha) = \sum_{i=1}^{l}\alpha_i - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j y_i y_j \boldsymbol{\phi}(\mathbf{x}_i)^T\boldsymbol{\phi}(\mathbf{x}_j), \tag{9}$$

subject to:

$$\sum_{i=1}^{l}\alpha_i y_i = 0, \tag{10}$$

$$C \geq \alpha_i \geq 0, \tag{11}$$

where $\{\alpha_i, i = 1, 2, ..., l\}$ are Lagrange multipliers. For this problem, we use the sequential minimum optimization [3] as the QP-solver.

D. Correction Score

We define a ranking scheme, which we call correction score (CS), to measure a feature's ability to separate the samples that are misclassified by the DGS obtained in the previous round of training. (Here we define a process of picking out a feature and adding it into a DGS as a *round* of training.) The CS of gene i is defined as:

$$CS_i = S_{bi}/S_{wi} \tag{12}$$

where

$$S_{bi} = \sum_{j\in C_1}(e_{ij} - \overline{x}_{c1})^2 + \sum_{k\in C_2}(e_{ik} - \overline{x}_{c2})^2 \tag{13}$$

$$S_{wi} = \sum_{j\in C_1}(e_{ij} - \overline{x}_{c2})^2 + \sum_{k\in C_2}(e_{ik} - \overline{x}_{c1})^2 \tag{14}$$

where e_{ij} and e_{ik} are the expression values of *misclassified* samples in $C1$ and $C2$, respectively. $\overline{x}_{c1}$ and $\overline{x}_{c2}$ are defined in Eq.2 and Eq.3. S_{bi} is the sum of squares of the inter-class distances [16] (the distances between samples of different classes) among the misclassified samples. S_{wi} is the sum of squares of the intra-class distances (the distances of samples within the same class) among the misclassified samples.

E. SVM-based AM-DGS

Inputs:

Training samples: $\mathbf{X}_{tr} = [\mathbf{x}_{tr1}, \mathbf{x}_{tr2}, ..., \mathbf{x}_{trl}]^T$, validation samples: $\mathbf{X}_v$, testing samples $\mathbf{X}_{test}$

Class labels for training, validation, and testing samples: $\mathbf{Y}_{tr} = [y_{tr1}, y_{tr2}, ..., y_{trl}]^T$, $\mathbf{Y}_v$, $\mathbf{Y}_{test}$

The number of top-ranked genes to search for DGSs: M.

Initialize:

Initialize DGS to an empty matrix: DGS=[ ].

Initialize the training error to 1: $E_{tr} = 1$.

Initialize the validation error to 0: $E_v = 0$.

Initialize the repeat counter to 0: $Rpt = 0$.

Choose a seed:

Calculate the TS for each feature in $\mathbf{X}_{tr}$.

for($m = 1$;until $m < M$; $m++$)

{

Select a feature with the m-th largest TS as the seed (**S**).

S →DGS.

Repeat until: $E_{tr} = 0$ or $E_v < E_{vpre}$ or $Rpt > 2$:

{

$E_{trpre} = E_{tr}$;

$E_{vpre} = E_v$;

Train an SVM with DGS then obtain E_{tr}.

Pick out the misclassified samples $\mathbf{X}_e = [\mathbf{x}_{e1}, \mathbf{x}_{e2}, ..., \mathbf{x}_{et}]^T$.

Validate the SVM using $\mathbf{X}_v$ and obtain E_v.

If $E_{vpre} = E_v$, $Rpt = Rpt + 1$.

Calculate CS for each feature in $\mathbf{X}_e$.

Pick out the feature with the largest CS and put it into the DGS.

}

}
Output
DGS

III. Experimental Results

We tested our method of searching for discriminating gene sets in two well-known gene expression data sets, i.e., the leukemia data set [4]and the prostate data set [14].

A. Leukemia Data Set

The leukemia data set [4] (http://www-genome.wi.mit.edu/cancer/) contains two types of leukemia samples, i.e., acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL). Golub *et al.* divided the data into 38 samples for training and the other 34 independent samples for testing. We normalized this data set by subtracting the mean and dividing the standard deviation across each sample.

We processed the leukemia data set with our SVM-based AM-DGS algorithm and showed the results in Table I. Here we list the 8 DGSs whose seeds are the top 8 genes according to their TSs. For each DGS, the first gene (i.e., the first line in the DGS) is its seed. The second, third (and so on) genes are the genes included in the DGS in the corresponding rounds, respectively.

From these results, we found that our SVM-based AM-DGS is very effective and efficient in finding good DGSs. Let us use DGS 1 to illustrate this. Since DGS 1 contained only 2 genes, we plotted the gene expression values of the two genes in DGS 1 in Fig.1. In the first round training, only the seed, i.e., gene $U50136\_rna1\_at$, was input to the SVM. Because gene $U50136\_rna1\_at$ has a high TS, the SVM misclassified only three samples that were indicated with arrows. In the second round training, the algorithm selected the gene that had the best capability to separate the three misclassified samples, i.e., gene $X17042\_at$. We found in Fig.1(a) that gene $X17042\_at$ "dragged" the misclassified samples away from the classes to which these 3 samples were mistakenly assigned in the previous round of training. Therefore, with the help of the second gene $X17042\_at$, DGS 1 increased its training accuracy from 92.11% to 100%: the 38 training samples were perfectly separated by DGS 1.

The best testing accuracy was obtained by DGS 8, which included 4 genes. The SVM obtained 100% training accuracy and 97.1% testing accuracy (i.e., 1 errors in the 34 testing samples) using DGS 8. In this data set, we used the 8 genes with the largest TSs as the seeds (M=8 in our algorithm summarized in the previous section). If more seeds were used, more DGSs could be found.

B. Prostate Data Set

The prostate data set [14] (http://www-genome.wi.mit.edu/MPR/prostate.) contains 50 nontumor samples and 52 prostate tumor samples. In this data set, there are expression values of 12600 genes. Considering the huge amount of genes included in the data set, we firstly ranked all the genes according to their TSs and then picked out the 1000 genes with the largest TSs for further processing. We followed the exactly same procedures as what we did in the colon data set to process this data set, i.e., we carried out 10-fold CV to the data using the SVM based AM-DGS. Tables II, III, IV show the testing, validation, and training results, respectively. The SVM-based AM-DGS achieved 87.3% 10-fold CV accuracy (i.e., 13 errors in the 102 samples) using the DGSs with 3.8 genes on average. In addition, we also compared out methods with the t-statistic, f-statistic, and SVM-RFE based methods in Table VI. From this comparsion, we found that the TS-based method achieved 87.3% with at least 15 genes. It also achieved a slightly better accuracy, i.e., 88.2% (12 errors in the 102 samples), with 200 genes. The FS-baesed method achieved 87.3% accuracy with at least 20 genes and 88.2% accuracy with 200 and 500 genes, respectively. The best accuracy for this data, i.e., 89.2% (11 errors in the 102 samples), was achieved by the SVM-RFE-based method with 100 genes. It also achieved 87.3% accuracy with 30 genes. Although the SVM-based AM-DGS did not achieved the best accuracy in this data set, it obtained the result very close to the best one with much smaller DGSs.

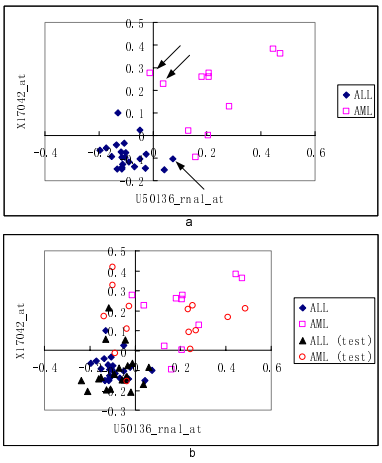

Fig. 1. Gene expression values for the two genes in DGS 1 in the leukemia data set. (a) a plot includes only the training samples; (b) a plot includes all the training and testing samples.

IV. Discussion

The results of leukemia data set visually indicate the effectiveness of our SVM-based AM-DGS algorithm. Except the seeds, all the genes in a DGS are selected according to their capability for correcting misclassification. Therefore, the SVM-based AM-DGS can optimize the cooperation among genes and hence leads to good accuracy and smaller DGSs. Compared with the filter approaches, e.g., TS and FS, the SVM-based AM-DGS can greatly reduce the redundancy in a DGS.

In conclusion, the SVM-AMDGS proposed here is effective and computationally efficient in searching for good DGSs, simulations wih the leukemia, and the prostate data sets show that our algorithm leads to highly accurate classifications with the smallest gene sets found in the literature.

REFERENCES

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

[2] P. Mitra, C. A. Murthy, S. K. Pal, "A Probabilistic Active Support Vector Learning Algorithm," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 26, pp. 413-418, 2004.

[3] J. C. Platt, "Sequential Minimum Optimization: A Fast Algorithm for Training Support Vector Machines," Microsoft Research, Cambridge, U.K., Technical Report, 1998.

[4] 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, et al. "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science,* vol. 286, pp. 531-537, 1999.

[5] A. A. Alizadeh, M.B. Eisen, R.E. Davis, C. Ma, I.S. Lossos, A. Rosenwald, J.C. Boldrick, H. Sabet, T. Tran, X. Yu, et al., "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling," *Nature,* Vol. 403, pp. 503-511, 2000.

[6] J. M. Khan, J.S. Wei, M. Ringner, L.H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C.R. Antonescu, C. Peterson, et al. "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks," *Nature Medicine,* Vol. 7, pp. 673-679, 2001.

[7] Y. Wang, F. Makedon, J. Ford, J. Pearlman, "Hykgene: a hybrid approach for selecting marker genes for phenotype classification using microarray gene expression data," *Bioinformatics* Advance Access Published.

[8] Y. Lai, B. Wu, L. Chen, H. Zhao, "Statistical method for identifying differential gene-gene coexpression patterns," *Bioinformatics* vol. 21, pp. 1565-1571, 2005.

[9] P. Broet, A. Lewin, S. Richardson, C. Dalmasso, H. Magdelenat, "A mixture model-based strategy for selecting sets of genes in multiclass response microarray experiments," *Bioinformatics* vol. 20, pp. 2562-2571, 2004.

[10] J. Devore, and R. Peck, *Statistics: the Exploration and Analysis of Data*, 3rd ed. Pacific Grove, CA.: Duxbury Press, 1997.

[11] E. P. Xing, M. I. Jordan, R. M. Karp, "Feature selection for high-dimensional genomic microarray data," *Proc. of the 18th international conference on machine learning,* pp. 601-608, Morgan Kaufmann Publishers Inc.

[12] V. Vapnik, *Statistical Learning Theory,* New York: Wiley, 1998.

[13] L. P. Wang (Ed.), *Support Vector Machines: Theory and Applications*, Springer, Berlin, 2005.

[14] D. Singh, P. G. Febbo, K. Ross, D. G. Jackson, J. Manola, "Gene expression correlates of clinical prostate cancer behavior" *Cancer Cell,* Vol. 1, pp. 203-209, 2002.

[15] P. Devijver, J. Kittler, *Pattern Recognition: a statistical approach.* London: Prentice Hall, 1982.

[16] X. Fu, L. P. Wang, "Data dimensionality reduction with application to simplifying RBF network structure and improving classification performance" *IEEE Trans. on systems, man, and cybernetics-part b: cybernetics,* Vol. 33, pp. 399-409, 2003.

Set No.	Gene Sets	Training Accuracy (%)	Testing Accuracy (%)
1	U50136_rna1_at	92.11	79.41
	X17042_at	100	82.35
2	X95735_at	97.37	94.12
	M23197_at	100	94.12
3	M55150_at	97.37	82.35
	M84526_at	92.11	82.35
	M23197_at	97.37	91.18
4	M16038_at	92.11	79.41
	U22376_cds2_s_at	94.74	82.35
5	Y12670_at	94.74	64.71
	U22376_cds2_s_at	100	82.35
6	M23197_at	92.11	85.29
	U22376_cds2_s_at	97.37	88.24
	M63138_at	97.37	94.12
7	D49950_at	97.37	94.12
	U22376_cds2_s_at	86.84	67.65
	X04085_rna1_at	94.74	76.47
	U50136_rna1_at	100	82.35
8	X17042_at	89.47	79.41
	U22376_cds2_s_at	89.47	85.29
	M86406_at	94.74	64.71
	X95735_at	100	97.06

TABLE I

TRAINING AND TESTING ACCURACIES FOR VARIOUS DGSS OBTAINED BY OUR SVM-BASED AM-DGS ALGORITHM TO THE LEUKEMIA DATA SET.

Fold No.	1	2	3	4	5	6	7	8	9	10
Fold Size	11	11	10	10	10	10	10	10	10	10
Seed 1 (%)	72.7	***63.6***	80	90	100	90	90	90	90	80
Seed 2 (%)	72.7	63.6	90	***100***	100	***90***	80	90	100	90
Seed 3 (%)	90.9	63.6	90	90	***100***	90	70	100	100	80
Seed 4 (%)	90.9	72.7	90	80	90	80	90	90	90	90
Seed 5 (%)	90.9	63.6	70	80	80	70	80	90	90	90
Seed 6 (%)	72.7	81.8	80	80	80	90	***80***	90	100	80
Seed 7 (%)	***90.9***	81.8	100	90	100	70	90	80	90	90
Seed 8 (%)	90.9	81.8	80	100	100	90	90	***90***	100	***90***
Seed 9 (%)	81.8	100	***80***	50	50	70	80	100	80	70
Seed 10 (%)	90.9	63.6	70	80	80	80	90	80	***90***	80

TABLE II

THE TESTING RESULTS FOR THE 10-FOLD CROSS VALIDATION CONDUCTED IN THE PROSTATE DATA WITH OUR SVM-BASED AM-DGS. EACH RESULT CONTAINS AN ACCURACY AND AN ERROR NUMBERS. FOR EXAMPLE, 72.7 MEANS THAT THE CORRESPONDING DGS OBTAINED 72.7% ACCURACY. THE ITALIC RESULTS ARE OBTAINED BY THE DGSS CHOSEN TO CLASSIFY THE TESTING DATA.

Fold No.	1	2	3	4	5	6	7	8	9	10
Seed 1 (%)	81.0	90.5	77.3	86.4	81.8	81.8	86.4	77.3	72.7	72.7
Seed 2 (%)	81.0	90.5	77.3	95.5	77.3	95.5	86.4	81.8	81.8	90.9
Seed 3 (%)	85.7	90.5	86.4	77.3	95.5	86.4	86.4	86.4	86.4	86.4
Seed 4 (%)	81.0	90.5	81.8	77.3	86.4	81.8	81.8	86.4	95.5	77.3
Seed 5 (%)	76.2	90.5	81.8	77.3	81.8	90.9	77.3	77.3	72.7	77.3
Seed 6 (%)	81.0	85.7	86.4	81.8	77.3	90.9	95.5	72.7	81.8	86.4
Seed 7 (%)	90.5	95.2	86.4	81.8	72.7	72.7	86.4	86.4	77.3	86.4
Seed 8 (%)	81.0	85.7	86.4	77.3	86.4	95.5	86.4	95.5	86.4	95.5
Seed 9 (%)	90.5	81.0	90.9	81.8	81.8	90.9	81.8	77.3	90.9	86.4
Seed 10 (%)	71.4	90.5	81.8	81.8	81.8	77.3	86.4	90.9	95.5	90.9

TABLE III

THE VALIDATION ACCURACY FOR THE 10-FOLD CROSS VALIDATION CONDUCTED IN THE PROSTATE DATA WITH OUR SVM-BASED AM-DGS.

Fold No.	1	2	3	4	5	6	7	8	9	10
Seed 1 (%)	88.6	95.7	94.3	90	94.3	91.4	88.6	88.6	90	92.9
Seed 2 (%)	87.1	94.3	94.3	94.3	90	94.3	95.7	85.7	84.3	95.7
Seed 3 (%)	87.1	94.3	87.1	92.9	90	97.1	92.9	95.7	94.3	88.6
Seed 4 (%)	91.4	91.4	91.4	94.3	92.9	80	94.3	84.3	85.7	94.3
Seed 5 (%)	90	90	87.1	94.3	87.1	87.1	87.1	97.1	90	91.4
Seed 6 (%)	82.9	85.7	85.7	80	80	94.3	94.3	95.7	84.3	88.6
Seed 7 (%)	92.9	92.9	92.9	84.3	88.6	91.4	84.3	87.1	78.6	84.3
Seed 8 (%)	94.3	94.3	94.3	91.4	82.9	88.6	91.4	91.4	84.3	91.4
Seed 9 (%)	91.4	91.4	91.4	78.6	77.1	84.3	97.1	92.9	82.9	88.6
Seed 10 (%)	91.4	94.3	80	74.3	74.3	95.7	94.3	82.9	91.4	87.1

TABLE IV

THE TRAINING ACCURACY FOR THE 10-FOLD CROSS VALIDATION CONDUCTED IN THE PROSTATE DATA WITH OUR SVM-BASED AM-DGS.

Fold No.	1	2	3	4	5	6	7	8	9	10
Number of Genes in the Set with Seed 1	3	***5***	4	2	4	4	3	2	3	4
Number of Genes in the Set with Seed 2	1	4	3	***3***	4	***4***	4	2	1	4
Number of Genes in the Set with Seed 3	1	3	1	4	***4***	3	3	6	3	1
Number of Genes in the Set with Seed 4	4	4	4	2	3	1	5	1	3	4
Number of Genes in the Set with Seed 5	3	4	1	4	2	4	2	4	3	3
Number of Genes in the Set with Seed 6	2	5	5	1	1	5	***3***	4	1	2
Number of Genes in the Set with Seed 7	***4***	5	5	1	3	3	1	2	1	1
Number of Genes in the Set with Seed 8	4	4	4	3	1	4	4	***3***	2	***4***
Number of Genes in the Set with Seed 9	4	4	***4***	1	1	1	4	5	3	2
Number of Genes in the Set with Seed 10	3	5	1	1	1	3	4	3	***4***	4

TABLE V

THE NUMBERS OF GENES USED BY THE DGSS IN THE PROSTATE DATA SET FOR THE 10-FOLD CROSS VALIDATION. THE ITALIC NUMBERS ARE FOR THE DGSS CHOSEN TO CLASSIFY TESTING SAMPLES.

Number of Genes	1	2	5	10	15	20	30	50	100	200	500
t-statistics	24	18	15	14	13	13	17	15	14	12	13
f-statistics	24	19	15	14	14	13	16	14	14	12	12
SVM-RFE	30	24	14	17	15	14	13	14	11	14	15
SVM-based AM-DGS	13 errors using DGSs with 3.8 genes on average.										

TABLE VI

THE COMPARISON OF THE ERRORS IN THE PROSTATE DATA SET WITH DIFFERENT APPROACHES.