Proceedings of the Second European Workshop on Data Mining and Text Mining in Bioinformatics

An Evaluation of Gene Selection Methods for Multi-class Microarray Data Classification

Hong Chai and Carlotta Domeniconi
Information and Software Engineering Department
George Mason University
Fairfax, VA 22030
hchai@gmu.edu carlotta@ise.gmu.edu

Abstract

The fundamental power of microarrays lies in the ability to conduct parallel surveys of gene expression patterns for tens of thousands of genes across a wide range of cellular responses, phenotypes and conditions. Thus microarray data contain an overwhelming number of genes relative to the number of samples, presenting challenges for meaningful pattern discovery. This paper provides a comparative study of gene selection methods for multi-class classification of microarray data. We compare several feature ranking techniques, including new variants of correlation coefficients, and Support Vector Machine (SVM) method based on Recursive Feature Elimination (RFE). The results show that feature selection methods improve SVM classification accuracy in different kernel settings. The performance of feature selection techniques is problem-dependent. SVM-RFE shows an excellent performance in general, but often gives lower accuracy than correlation coefficients in low dimensions.

1 Introduction

Gene expression profiling studies via DNA microarrays offer unprecedented opportunities for advancing fundamental biological research and clinical practice. Microarray technology allows researchers to simultaneously measure the expression level of tens of thousands of genes, creating a comprehensive overview of exactly which genes are being expressed in a specific tissue under various conditions. These studies produce a massive amount of data which poses challenging problems for the discovery of informative patterns.

A microarray can contain up to 20,000 features, each of which recognizes mRNA from a single gene, and a relatively small number of experiments or samples (in the order of hundreds or less). As a consequence, the identification of discriminant genes to classifying tissue types, e.g., presence of cancer, is of fundamental and practical interest. Such genetic markers, in fact, can be found of value in further investigation of the disease and in future therapies. From a machine learning standpoint, the reduction of dimensionality of the data avoids the possibility of overfitting. The classification of a few dozens of points lying in a space with thousands dimensions is a hopeless task due to the curse-of-dimensionality. One can easily find a (linear) decision boundary that perfectly separates the training data. However, such a classifier will perform poorly on previously unseen data. In other words, it won't generalize well.

Regularization techniques can prevent the overfitting of the data, without performing dimensionality reduction [20]. Support Vector Machines (SVMs) are such an example [6]. Nevertheless, our experimental results show that also SVMs can benefit from the reduction of dimensionality due to feature selection. Another solution to high dimensional settings consists in projecting the data onto the principal components, obtained as linear combinations of the input features [9]. A disadvantage is that none of the original features can be discarded. Moreover, the new dimensions can be difficult to intepret, making it hard to understand the groups of data in relation to the original space. SVMs themselves have been used successfully for gene selection (SVM-RFE) [12]. The weights that multiply the inputs in the solution boundary are used as feature ranking coefficients. In general, feature ranking techniques via correlation coefficients can be particularly useful for gene selection. One can consider only top ranked genes (above a certain score threshold, or a fixed number) for further analysis, or to train a classifier.

While binary (two-class) classification has been extensively studied over the past few years [1, 3, 4, 8, 12], the multi-class classification case has received little attention [13, 16, 5]. In this paper, we

focus on multi-class classification, and compare several gene ranking methods, including new variants of correlation coefficients, using different microarray datasets. In analogy with the Structural Risk Minimization principle [20], for each ranking method, we construct nested subsets of features $F_1 \subset F_2 \subset \ldots \subset F$. Given a classification model (SVMs with different kernel functions in our experiments), one can select the subset of features that gives the best cross-validated accuracy. Our main findings are:

1. SVMs classification benefits from gene selection;
2. Gene ranking with correlation coefficients gives higher accuracy than SVM-RFE in low dimensions in most data sets. The best performing correlation score varies from problem to problem;
3. Although SVM-RFE shows an excellent performance in general, there is no clear winner. The performance of feature selection methods seems to be problem-dependent;
4. For a given classification model, different gene selection methods reach the best performance for different feature set sizes;
5. Very high accuracy was achieved on all the data sets studied here. In many cases perfect accuracy (based on leave-one-out error) was achieved;
6. The NCI60 dataset [17] shows lower accuracy values. This dataset has the largest number of classes (eight), and smaller sample sizes per class. SVM-RFE handles this case well, achieving 96.72% accuracy with 100 selected genes and a linear kernel. The gap in accuracy between SVM-RFE and the other gene ranking methods is highest for this dataset (ca. 11.5%).

2 Problem Description and Related Work

In a classification problem, we are given C classes and m training observations. The training obervations consist of n feature measurements $\mathbf{x} = (x_1, \ldots, x_n)^t \in \Re^n$, and the known class labels $y = 1, \ldots, C$. The goal is to predict the class label of a given query $\mathbf{q}$. For the problem we consider here, the features are gene expression coefficients, and the observations correspond to samples or patients. Thus, $n >> m$.

Traditional gene selection methods keep the genes that *individually* best discriminate between training data of different classes. Such methods make use of correlation coefficients and expression ratios, and usually are defined for two-class classification problems. Let us consider class labels $y \in \{-1, +1\}$. The correlation coefficient used in [11] as ranking criterion is defined as: $w_j = \frac{\mu_j^+ - \mu_j^-}{\sigma_j^+ + \sigma_j^-}$, $j = 1, \ldots, n$, where μ_j^+ (μ_j^-) is the mean value of gene j for the + (−) class. Similarly, σ_j^+ and σ_j^- are the respective standard deviations. Large positive w_j values indicate a strong correlation with the positive class. Large negative w_j values indicate a strong correlation with the negative class. The objective in [11] is to select an equal number of genes j with large positive and large negative correlation coefficient. In [10], the authors consider $|w_j|$ as ranking criterion. In [15], the Fisher criterion score [7] is used: $w_j = \frac{(\mu_j^+ - \mu_j^-)^2}{(\sigma_j^+)^2 + (\sigma_j^-)^2}$, $j = 1, \ldots, n$, which gives higher scores to genes whose means differ greatly between the two classes, relative to their variances.

Correlation scores assume that features are independent. Each feature is analyzed in isolation, without taking into consideration the mutual information across features. This fact implies that redundant genes may be selected, and genes individually not important, but complementary to each other, may not be selected.

Another technique for feature ranking uses the concept of Shannon entropy. Given entropy E as a measure of impurity in a set S of training examples, it is possible to define a measure of the effectiveness of a feature (or attribute) A in classifying the training data. The measure, called *Information Gain*, is simply the expected reduction in entropy caused by partitioning the data according to this feature [14]. More precisely, the information gain $I(S, A)$ of a feature A, relative to a set of data S, is defined as $I(S, A) = E(S) - \sum_{v \in V(A)} \frac{|S_v|}{|S|} E(S_v)$, where $V(A)$ is the set of all possible values of feature A, and S_v is the subset of S for which feature A has value v. The first term is the entropy of the entire set S: $E(S) = \sum_{i=1}^{C} -\frac{|C_i|}{|S|} \log_2 \frac{|C_i|}{|S|}$, where $|C_i|$ is the number of training data in class C_i, and $|S|$ is the cardinality of the entire set S. The definition of information gain can be extended to handle continuous

valued features. This is achieved by searching for candidate thresholds sorting the points according to the continuous feature, and identifying adjacent points that differ in their classification label [14].

The *Chi-squared* is another method that evaluates features individually with respect to the classes. The range of continuous valued features needs to be discretized into intervals. A matrix A is then formed, where A_{ij} is the number of samples of the C_i class within the jth interval. Let C_{I_j} be the number of samples in the jth interval. The expected frequency of A_{ij} is $E_{ij} = C_{I_j}|C_i|/m$. The Chi-squared statistic of a feature is then defined as $\chi^2 = \sum_{i=1}^{C}\sum_{j=1}^{I}\frac{(A_{ij}-E_{ij})^2}{E_{ij}}$, where I is the number of intervals. The larger the χ^2 value, the more informative the corresponding feature is.

In another effort, the authors in [5] evaluated the discriminatory power of a gene with test statistics such as the ANOVA F, the Brown-Forsythe, the Cochran, and the Welch test statistics. These are extensions of the t-statistic used in the two-class classification problem.

Linear SVMs have been used successfully for gene selection [12]. The squared weights, w_j^2, that multiply the inputs in the solution boundary $f(\mathbf{x}) = \mathbf{w}\cdot\mathbf{x} + b$ are used as feature ranking coefficients. The underlying idea is that features with largest weights influence most the classification decision. Thus, if the classifier performs well, those features with the largest weights are the most informative. As a result, an iterative procedure (*Recursive Feature Elimination*, or RFE) trains the SVM classifier, computes the ranking w_j^2 for all features, and removes the feature with smallest ranking criterion. The procedure is iterated until a certain number of selected features is obtained. To speed up the process, several features may be removed at each iteration. In contrast to feature ranking using correlation coefficients, the RFE method is a *multivariate* approach which evaluates the relevance of multiple features simultaneously.

3 Feature Correlation Scores for Multiclass Problems

In this section we introduce several correlation scores for feature ranking to handle multi-class classification scores. For each class i and each feature j, we define:

$$\mu_{j,i} = \frac{1}{|C_i|}\sum_{\mathbf{x}\in C_i} x_j. \tag{1}$$

$\mu_{j,i}$ represents the mean value of feature j for class C_i. We also define the *total mean* along feature j:

$$\mu_j = \frac{1}{m}\sum_{\mathbf{x}} x_j. \tag{2}$$

Using equations (1) and (2), we provide a measure of the *between-class scatter* along feature j:

$$B_j = \sum_{i=1}^{C}|C_i|(\mu_{j,i}-\mu_j)^2. \tag{3}$$

This leads to the following score function

$$BScatter_j = \frac{B_j}{\sum_{i=1}^{C}\sigma_{ji}} \qquad j = 1,\ldots n, \tag{4}$$

where σ_{ji} is the standard deviation of class i along feature j. This score is related to Fisher discriminant analysis for multiple classes [7] under feature independence assumption. It credits the largest score to the feature that maximizes the ratio of the between-class scatter to the within-class scatter.

Let us consider: $\mu_{j,max} = \max_l \mu_{j,l}$, $\mu_{j,min} = \min_l \mu_{j,l}$. The second score function we define is

$$MinMax_j = \frac{\mu_{j,max}-\mu_{j,min}}{\sum_{i=1}^{C}\sigma_{ji}} \qquad j = 1,\ldots n. \tag{5}$$

This score function favors features along which the farthest mean-class difference is large, and the within class variance is small.

For each feature j, we sort the C values $\mu_{j,i}$ in non-decreasing order: $\mu_{j,1} \leq \mu_{j,2} \leq \ldots \leq \mu_{j,C}$. Let us define now $b_{j,l} = |\mu_{j,l} - \mu_{j,l+1}|$, $1 \leq l \leq C-1$. $b_{j,l}$ measures the distance between adjacent mean class values along feature j. The third score function rewards the features with large distances between adjacent mean class values:

$$bSum_j = \frac{\sum_{l=1}^{C-1} b_{j,l}}{\sum_{i=1}^{C} \sigma_{ji}} \qquad j = 1, \ldots n. \tag{6}$$

In addition, we consider two variants of the previous function. The following one rewards features j with a large between-neighbor-class mean difference:

$$bMax_j = \frac{\max_l b_{j,l}}{\sum_{i=1}^{C} \sigma_{ji}} \qquad j = 1, \ldots n. \tag{7}$$

Alternatively, we can favor the features with large *smallest* between-neighbor-class mean difference:

$$bMin_j = \frac{\min_l b_{j,l}}{\sum_{i=1}^{C} \sigma_{ji}} \qquad j = 1, \ldots n. \tag{8}$$

Finally, we consider a score function which combines *MinMax* and *bMin*:

$$Comb_j = \frac{\min_l(b_{jl})(\mu_{j,max} - \mu_{j,min})}{\sum_{i=1}^{C} \sigma_{ji}} \qquad j = 1, \ldots n. \tag{9}$$

4 Experimental Analysis

The Datasets. We used the following four datasets. The MLL dataset consists of gene expression profiles of three classes of leukemia and is available at http://research.nhgri.nih.gov/microarray/Supplement. It was first studied by Scott et al. [18] in proposing that a distinct disease type, MLL, can be clearly separated from conventional acute lymphoblastic leukemia (ALL) and acute myelogenous leukemias (AML). The dataset includes 12582 probe sets from the Affymetrix chip, and contains 72 samples. The numbers of samples in the three classes are balanced, 24 in ALL, 20 in MLL, and 28 in MLL. There are no missing values in this dataset.

The Lymphoma dataset covers 9 classes, 96 malignant and normal lymphocyte samples. There are 4026 genes. It was published in [1] and is available at http://llmpp.nih.gov/lymphoma. Classes that contain less than 5 samples were removed in our experiments, and hence 6 classes remained. The numbers of samples in each class are, 46 in DLBCL, 11 in CLL, 9 in FL (malignant classes), 11 in ABB, 6 in RAT, and 6 in TCL (normal samples).

The expression data from budding yeast Saccharomyces cerevisiae consists of 80-sample gene expression vectors for 6221 genes. The samples contain 3 subtypes. The dataset is available at http://rana.lbl.gov/EisenData.htm.

The NCI60 dataset was first studied in [17]. cDNA microarrays were used to examine the variation in gene expression among the 60 cell lines from the National Center Institutes anticancer drug screen. The dataset contains 9 classes and can be downloaded from http://genome-www.stanford.edu/nci60.

There are missing values in the Lymphoma, Yeast and NCI60 datasets. We deleted genes in the Yeast dataset that have above 20 missing expression values, and genes in the NCI60 datasets that have more than 50 missing values. Then, we used the K-nearest neighbor method to impute the remaining missing values in each dataset. For a gene with missing values, the K nearest neighbors are identified from the subset of genes that have complete expression values ($K = 7$ in our experiments). The average of the neighbors' values is used to substitute a missing value [19]. Table 1 summarizes the characteristics of the four datasets used in our experiments.

4.1 Experimental Design

We scaled each dataset so that every gene's expression values are ranged between [-1,1]. Feature selection methods are performed on each dataset to obtain subsets of top-ranked genes. These include our six variants of correlation scores (*BScatter*, *MinMax*, *bSum*, *bMax*, *bMin*, *Comb*), as well

Table 1: The characteristics of the four datasets.

Dataset	#samples	#genes	#classes
MLL	72	12582	3
Lymphoma	88	4026	6
Yeast	80	5775	3
NCI60	61	1155	8

as Chi-squared, Information Gain, and SVM-RFE methods. The six correlation scores are implemented with Python, while the latter three methods are performed using the Weka software available at http://www.cs.waikato.ac.nz/ml/weka. In SVM-RFE, we set the percentage to drop at each iteration to be 10, until 20% of the total number of features is eliminated. Successively, one single feature is dropped at each iteration. Weka handles multi-class problems by ranking attributes for each class separately using a one-vs-all method, and then dealing from the top of each pile to give a final ranking.

We compared the above nine feature selection methods by considering different feature set sizes. Each of the resulting feature set was used to train an SVM classifier with different kernel functions: linear, 2-degree polynomial, and Gaussian. Due to the small number of samples available, leave-one-out cross validation was performed to assess classification performance, using a fixed set of features previously selected with the whole training set. We cross-validated the soft-margin parameter and the width of the Gaussian kernel testing values between [1,100] and [0.001, 2], respectively. The SVM classifications were performed using LIBSVM, which implements the one-vs-one method when classifying multiple categories. The LIBSVM software can be downloaded from http://www.csie.ntu.edu.tw/∼cjlin/libsvm.

4.2 Results

The performance of the nine ranking methods are shown in Tables 2-5. In the tables, the first row represents the number of selected features. Each cell contains the leave-one-out classification accuracy of SVM, achieved using the corresponding number of top genes ranked by one of the feature selection method. It is seen from the experiments that:

1. SVM classification benefits from feature selection methods. In almost all situations, highest prediction rates are achieved for each selection method when using a number of top ranked genes much smaller than the full dimensionality. The accuracy of classification does depend on the use of feature selection methods.

2. The best performing correlation score varies from problem to problem. Among all feature ranking methods, our proposed function *MinMax* gives highest accuracy using 2 through 20 or 100 top-selected genes when classifying the Lymphoma dataset. Overall, gene ranking with correlation coefficients gives higher accuracy than SVM-RFE in low dimensions in most data sets. In higher dimensions, the correlation coefficients may select redundant genes. This is likely the reason why SVM-RFE shows a more robust behavior, in general, in higher dimensions.

3. SVM-RFE yields perfect prediction on the MLL, Lymphoma, and the Yeast datasets. Correlation coefficient scores and information gain, however, also achieves 100% accuracy on these datasets. Chi-squared is the only method that never achieves perfect prediction. Although SVM-RFE shows an excellent performance in general, there is no clear winner. The performance of feature selection methods seems to be problem-dependent.

4. For a given classification model, different gene selection methods reach the best performance at different feature set sizes (as highlighted by the values in boldface in the tables). Very high accuracy was achieved on all the datasets studied here. Perfect accuracy (based on leave-one-out error) was achieved in many cases.

5. The NCI60 dataset shows lower accuracy values. This dataset has the largest number of classes (eight), and smaller sample sizes per class. SVM-RFE handles this case well, achieving 96.72% accuracy with 100 selected genes and a linear kernel. The gap in accuracy between SVM-RFE and the other gene ranking methods is highest for this dataset (ca. 11.5%).

6. For the NCI60 dataset, each of our proposed method gives above 80% prediction accuracy, except in two cases (73.77% and 75.41%). This compared higher to the result in [13], where the best reported performance was 67% from SVM classification using the gene selection program Rankgene.

Table 2: Classification accuracy (%) for MLL data: SVM with linear, polynomial, and Gaussian kernels.

SVM Linear	*2*	*5*	*10*	*20*	*50*	*100*	*200*	*500*	*1000*	*5000*
BScatter	86.11	91.67	95.83	93.05	95.83	**100**	100	97.22	97.22	97.22
MinMax	84.42	91.67	93.05	95.83	95.83	97.22	**98.61**	97.22	97.22	97.22
bSum	86.11	91.67	93.05	94.44	95.83	97.22	**98.61**	97.22	97.22	97.22
bMax	77.78	87.5	95.83	98.61	95.83	**100**	100	100	98.61	97.22
bMin	83.33	90.28	88.89	94.44	91.66	95.83	**97.22**	97.22	95.83	97.22
Comb	83.33	91.67	93.05	93.05	93.05	93.05	95.83	95.83	95.83	**97.22**
Chi-squared	91.67	93.05	90.28	**97.22**	97.22	97.22	97.22	97.22	97.22	97.22
Info Gain	91.67	97.22	91.67	95.83	95.83	97.22	**100**	97.22	97.22	97.22
SVM-RFE	86.11	94.44	**100**	100	100	100	100	100	100	100
SVM Polynomial	*2*	*5*	*10*	*20*	*50*	*100*	*200*	*500*	*1000*	*5000*
BScatter	84.72	93.05	94.44	91.67	95.83	**98.61**	98.61	98.61	95.83	95.83
MinMax	87.5	91.66	93.05	93.06	97.22	95.83	**98.61**	97.22	97.22	97.22
bSum	84.72	91.67	93.05	91.67	97.22	95.83	**98.61**	97.22	95.83	97.22
bMax	79.17	87.5	95.83	98.61	98.61	98.61	**100**	100	98.61	97.22
bMin	83.33	90.28	91.67	94.44	95.83	95.83	95.83	94.44	94.44	**97.22**
Comb	84.72	91.67	93.05	93.05	93.05	94.44	94.44	94.44	95.83	**97.22**
Chi-squared	91.67	91.67	91.67	95.83	95.83	**97.22**	97.22	95.83	95.83	97.22
Info Gain	91.67	97.22	93.05	94.44	93.05	**98.61**	98.61	98.61	95.83	95.83
SVM-RFE	87.5	95.83	98.61	**100**	100	100	100	100	100	100
SVM Gaussian	*2*	*5*	*10*	*20*	*50*	*100*	*200*	*500*	*1000*	*5000*
BScatter	87.5	93.05	94.44	94.44	97.22	98.61	**100**	97.22	97.22	90.28
MinMax	87.5	90.28	93.05	95.83	97.22	97.22	**98.61**	97.22	97.22	93.05
bSum	87.5	90.28	93.05	94.44	**97.22**	95.83	97.22	97.22	97.22	93.05
bMax	76.38	88.89	95.83	98.61	98.61	**100**	100	97.22	97.22	90.28
bMin	84.72	90.28	91.67	95.83	93.05	95.83	95.83	**97.22**	95.83	90.28
Comb	84.72	93.05	93.05	94.44	93.05	93.05	**97.22**	90.28	79.16	70.83
Chi-squared	91.67	91.67	91.67	95.83	95.83	**97.22**	97.22	94.44	80.55	70.83
Info Gain	91.67	97.22	93.05	95.83	93.05	97.22	**100**	93.05	88.89	86.11
SVM-RFE	87.5	95.83	**100**	100	100	100	100	100	100	100

5 Conclusions and Future Work

This paper provides a comparative study on feature selection for multi-class classification of microarray data. The results show that feature selection methods improve SVM classification accuracy in different kernel settings. In many datasets perfect prediction was achieved using one or more subsets of top features. SVM-RFE shows an excellent performance in general, but it gives lower accuracy than correlation coefficients, information gain, and Chi-squared method in low dimensions.

The selection of a fixed set of features over the whole training set induces a bias in the results. Valuable suggestions on how to assess and correct the selection bias are discussed in [2], and will be considered in our future experiments. Our proposed score functions did not take into consideration the fact that the correlation between any pair of selected features should be low. In fact, there may exist redundant genes in a given subset. In future work, our ranking method will be modified so that selected genes have correlation no larger than a certain threshold. In addition, top-ranked genes selected by the proposed score functions will be compared to marker genes identified in other benchmark studies.

Acknowledgements. This research is in part supported by a 2004 R. E. Powe Junior Faculty Award.

References

[1] Alizadeh, A., Eisen, M., et al., Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling, *Nature*, **403**, 2000.

[2] Ambroise, C., McLachlan, G. J., Selection bias in gene extraction on the basis of microarray gene-expression data, *National Academy of Sciences*, **99**(10):6562-6566, 2002.

[3] Ben-Dor, A., Friedman, N., Yakhini, Z., Scoring Genes for Relevance, *Technical Report of the Leibniz Center*, 2000-38, 2000.

Table 3: Classification accuracy (%) for Lymphoma data: SVM with linear, polynomial, and Gaussian kernels.

SVM Linear	*2*	*5*	*10*	*20*	*50*	*100*	*200*	*500*	1000	*4026*
BScatter	60.23	77.27	92.04	95.45	96.59	**98.86**	97.72	97.72	97.72	97.72
MinMax	84.09	90.91	97.72	98.86	98.86	**100**	98.86	98.86	97.72	97.72
bSum	82.95	90.91	96.59	98.86	98.86	**100**	98.86	98.86	97.72	97.72
bMax	59.09	71.59	84.09	94.32	96.59	**97.72**	97.72	97.72	97.72	97.72
bMin	71.59	76.12	94.32	97.72	97.72	**98.86**	98.86	98.86	98.86	97.72
Comb	65.91	81.82	95.45	97.72	98.86	**100**	100	97.72	97.72	97.72
Chi-squared	73.86	90.91	94.32	97.72	96.59	**98.86**	97.72	97.72	97.72	97.72
Info Gain	76.13	88.63	92.04	**97.72**	96.59	96.59	97.72	97.72	97.72	97.72
SVM-RFE	73.50	82.95	93.18	96.59	**100**	100	100	100	100	97.72
SVM Polynomial	*2*	*5*	*10*	*20*	*50*	*100*	*200*	*500*	*1000*	*4026*
BScatter	60.23	78.41	92.04	96.59	96.59	**98.86**	97.72	97.72	97.72	97.72
MinMax	84.09	92.04	97.72	98.86	98.86	**100**	98.86	98.86	98.86	97.72
bSum	84.09	92.04	95.45	97.72	98.86	**100**	98.86	98.86	98.86	97.72
bMax	62.5	72.72	82.89	89.77	95.45	97.72	97.72	**98.86**	97.72	97.72
bMin	71.09	77.27	94.32	95.45	96.59	**98.86**	98.86	98.86	98.86	97.72
Comb	67.04	81.82	94.32	95.45	98.86	**100**	100	97.72	98.86	97.72
Chi-squared	77.27	90.91	94.32	96.59	97.72	**98.86**	97.72	97.72	97.72	97.72
Info Gain	75.0	87.5	90.91	**97.72**	95.45	95.45	96.59	97.72	97.72	97.72
SVM-RFE	77.27	81.82	90.91	94.32	**100**	100	100	100	100	97.72
SVM Gaussian	*2*	*5*	*10*	*20*	*50*	*100*	*200*	*500*	*1000*	*4026*
BScatter	60.23	82.95	90.91	97.72	97.72	97.72	97.72	97.72	**98.86**	86.36
MinMax	84.09	92.04	96.59	98.86	98.86	**100**	98.86	96.59	98.86	86.36
bSum	84.09	90.91	95.45	98.86	98.86	**100**	97.72	98.86	97.72	86.36
bMax	61.36	72.73	82.95	94.32	95.45	97.72	97.72	92.05	**98.86**	86.36
bMin	72.72	78.41	89.77	95.45	93.18	98.86	**100**	88.63	96.59	86.36
Comb	65.91	80.68	95.45	96.59	96.59	**98.86**	98.86	92.04	98.86	86.36
Chi-squared	77.27	90.91	96.59	97.72	96.59	97.72	**98.86**	92.04	98.86	86.36
Info Gain	77.27	86.36	93.18	97.72	96.59	97.72	**98.86**	83.18	98.86	86.36
SVM-RFE	77.27	84.09	92.04	97.72	98.86	**100**	98.86	98.86	98.86	86.36

[4] Brown, M., Grundy, W., Lin, D., Cristianini, N., Sunet, C., Haussler, D., Knowledge-based Analysis of Microarray Gene Expression Data By Using Support Vector Machines, *National Academy of Sciences*, **97**:262-267, 2000.

[5] Chen, D., Hua, D., Reifman, J., Cheng, X., Gene Selection for Multi-class Prediction of Microarray Data, *Computational Systems Bioinformatics*, 2003.

[6] Cristianini, N., Shawe-Taylor, J., *An Introduction to Support Vector Machines and Other Kernel-based Methods*, Cambridge University Press, 2000.

[7] Duda, R. O., Hart, P., *Pattern classification and scene analysis*, Wiley, 1973.

[8] Eisen, M., et al, Cluster analysis and display of genome-wide expression patterns, *National Academy of Sciences*, **95**, 1998.

[9] Fukunaga, K., *Introduction to Statistical Pattern Recognition*, Academic Press, 1990.

[10] Furey, T., Cristianini, N., Duffy, N., Bednarski, D. W., Schummer, M., Haussler, D., Support vector machine classification and validation of cancer tissue samples using microarray expression data, *Bioinformatics*, **16**, 2000.

[11] Golub, T.R., Slonim, D.K., Tamayo, P., Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring, *Science*, **286**, 1999.

[12] Guyon, I., Weston, J., Barnill, S., Gene Selection for Cancer Classification Using Support Vector Machines, *Machine Learning*, **46**:389-422, 2002.

[13] Li, T., Zhang, C., Ogihara, M., A Comparative Study of Feature Selection and Multiclassification Methods for Tissue Classification Based on Gene Expression, *Bioinformatics*, 2004 (In Press).

[14] Mitchell, T. M., *Machine Learning*, McGraw-Hill, 1997.

[15] Pavlidis, P., Weston, J., Cai, J., Grundy, W. N., Gene Functional Analysis from Heterogeneous Data, *Research in Computational Molecular Biology* (RECOMB), 2001.

Table 4: Classification accuracy (%) for Yeast data: SVM with linear, polynomial, and Gaussian kernels.

SVM Linear	*2*	*5*	*10*	*20*	*50*	*100*	*200*	*500*	*1000*	*5775*
BScatter	87.5	86.25	90.0	**96.25**	92.5	93.75	95.0	95.0	95.0	96.25
MinMax	90.0	91.25	95.0	96.25	96.25	96.25	96.25	**97.5**	97.5	96.25
bSum	88.75	91.25	95.0	95.0	95.0	95.0	96.25	96.25	**97.5**	96.25
bMax	86.25	86.25	87.5	86.25	88.75	**96.25**	96.25	96.25	96.25	96.25
bMin	81.25	85.0	92.5	95.0	95.0	96.25	96.25	**97.5**	97.5	96.25
Comb	82.5	88.75	91.25	93.75	95.0	**96.25**	96.25	96.25	96.25	96.25
Chi-squared	80.0	88.75	92.5	95.0	92.5	95.0	95.0	**96.25**	95.0	96.25
Info Gain	87.5	91.25	93.75	95.0	**96.25**	96.25	95.0	96.25	96.25	96.25
SVM-RFE	78.75	95.0	96.25	98.25	**100**	100	100	100	100	96.25
SVM Polynomial	*2*	*5*	*10*	*20*	*50*	*100*	*200*	*500*	*1000*	*5775*
BScatter	87.5	86.25	90.0	90.0	95.0	95.0	**96.25**	96.25	96.25	96.25
MinMax	90.0	92.5	95.0	95.0	**96.25**	96.25	96.25	96.25	96.25	96.25
bSum	92.5	91.25	95.0	95.0	**96.25**	96.25	96.25	9.25	96.25	96.25
bMax	86.25	**96.25**	86.25	86.25	87.5	96.25	95.0	96.25	96.25	96.25
bMin	83.75	85.0	88.75	93.75	95.0	**96.25**	96.25	96.25	96.25	96.25
Comb	82.5	87.5	90.0	93.75	93.75	96.25	96.25	**97.5**	96.25	96.25
Chi-squared	91.25	90.0	93.75	95.0	**96.25**	95.0	96.25	96.25	96.25	96.25
Info Gain	91.25	92.5	96.25	93.75	96.25	96.25	96.25	96.25	**97.5**	96.25
SVM-RFE	81.25	96.25	96.25	98.75	98.25	**100**	100	97.5	96.25	96.25
SVM Gaussian	*2*	*5*	*10*	*20*	*50*	*100*	*200*	*500*	*1000*	*5775*
BScatter	87.5	86.25	90.0	88.75	95.0	**96.25**	96.25	96.25	93.75	85.0
MinMax	91.25	92.5	**96.25**	95.0	96.25	96.25	96.25	96.25	93.75	85.0
bSum	88.75	92.5	95.0	95.0	**96.25**	96.25	96.25	96.25	93.75	85.0
bMax	86.25	86.25	86.25	88.75	95.0	**96.25**	96.25	96.25	93.75	85.0
bMin	82.5	83.75	87.5	93.75	**96.25**	93.75	96.25	95.0	93.75	85.0
Comb	82.5	90.0	91.25	93.75	95.0	95.0	**96.25**	95.0	93.75	85.0
Chi-squared	90.0	92.5	93.75	95.0	95.0	**96.25**	92.5	86.25	85.0	85.0
Info Gain	91.25	93.75	95.0	95.0	**96.25**	96.25	93.75	86.25	86.25	85.0
SVM-RFE	80.0	95.0	95.0	98.75	**100**	95.0	76.25	87.5	85.0	85.0

Table 5: Classification accuracy (%) for NCI60 data: SVM with linear, polynomial, and Gaussian kernels.

SVM Linear	*2*	*5*	*10*	*20*	*50*	*100*	*200*	*500*	*800*	*1155*
BScatter	29.51	29.51	52.46	77.04	78.68	**81.97**	81.97	81.97	80.33	75.41
MinMax	24.59	22.95	44.26	70.49	**81.97**	78.68	81.97	81.96	78.69	75.41
bSum	24.59	19.67	45.90	65.57	**81.97**	78.68	81.97	81.97	78.68	75.41
bMax	24.59	19.67	52.46	57.38	73.77	**81.97**	81.97	80.33	78.69	75.41
bMin	32.78	34.42	62.29	65.57	68.85	**80.33**	78.69	78.69	75.41	75.41
Comb	29.51	65.57	65.57	72.13	77.05	81.97	**83.60**	77.05	75.41	75.41
Chi-squared	42.62	62.29	68.85	81.97	81.97	**85.24**	83.61	78.68	80.33	75.41
Info Gain	37.70	63.93	78.69	70.49	81.97	**83.61**	83.61	78.69	80.33	75.41
SVM-RFE	34.43	59.01	68.85	78.69	91.80	**96.72**	96.72	93.44	83.60	75.41
SVM Polynomial	*2*	*5*	*10*	*20*	*50*	*100*	*200*	*500*	*800*	*1155*
BScatter	29.50	34.46	57.37	77.05	**81.96**	78.68	81.96	81.96	81.96	73.77
MinMax	21.31	22.95	42.62	68.85	**85.25**	80.33	80.33	83.16	81.97	73.77
bSum	22.95	26.23	44.26	67.21	**83.16**	80.33	80.33	83.61	81.96	73.77
bMax	21.31	19.62	49.18	68.85	68.85	**81.97**	81.97	81.97	77.05	73.77
bMin	31.14	42.62	55.74	60.65	70.49	**80.33**	78.69	80.33	78.69	73.77
Comb	39.34	75.41	67.21	75.41	**80.33**	78.69	78.69	80.33	77.05	73.77
Chi-squared	49.18	67.21	67.21	78.58	81.96	**85.25**	85.25	83.61	77.05	73.77
Info Gain	45.90	65.57	78.69	75.41	**85.25**	85.25	83.61	80.33	77.05	73.77
SVM-RFE	34.43	52.46	65.57	83.60	93.44	**95.08**	95.08	91.80	86.88	73.77
SVM Gaussian	*2*	*5*	*10*	*20*	*50*	*100*	*200*	*500*	*800*	*1155*
BScatter	27.89	29.50	54.09	77.05	80.33	77.05	80.33	**81.97**	68.85	60.65
MinMax	26.23	22.95	37.70	72.13	80.32	**83.60**	77.05	78.69	70.49	60.65
bSum	24.59	22.95	44.26	72.13	**83.16**	83.16	83.16	78.68	70.49	60.65
bMax	24.59	21.31	52.46	59.01	52.46	67.21	**73.77**	70.49	67.21	60.65
bMin	36.06	37.71	50.82	59.02	70.49	**75.41**	73.77	70.49	63.93	60.65
Comb	31.14	47.54	50.82	68.85	73.77	**81.96**	77.05	68.85	60.65	60.65
Chi-squared	49.18	62.29	73.77	73.77	80.33	**83.16**	81.96	70.49	63.93	60.65
Info Gain	47.54	70.49	73.77	77.04	80.33	**83.60**	78.68	62.29	65.57	60.65
SVM-RFE	34.42	57.37	72.13	80.33	93.44	**95.08**	77.05	77.05	72.13	60.65

[16] Ramaswamy, S., et al., Multiclass cancer diagnosis using tumor gene expression signatures, *National Academy of Sciences*, **98**:26, 2001.

[17] Ross, D. T., et al., Systematic variation in gene expression patterns in human cancer cell lines, *Nature Genetics*, **24**(3):227-235, 2000.

[18] Scott, A., Armstrong, et al., MLL Translocations Specify A Distinct Gene Expression Profile that Distinguishes A Unique Leukemia. *Nature Genetics*, **30**:41-47, January 2002.

[19] Szymkowiak, A. et al., Imputing Missing Values in Diary Records in Sun-Exposure Study, *IEEE Workshop on Neural Networks for Signal Processing*, 489-498, 2001.

[20] Vapnik, V. N., *Statistical Learning Theory*, Wiley Interscience, 1998.