Taylor & Francis
Taylor & Francis Group

Communications in Statistics - Simulation and Computation

ISSN: 0361-0918 (Print) 1532-4141 (Online) Journal homepage: http://www.tandfonline.com/loi/lssp20

Comparison of Support Vector Machines to Other Classifiers Using Gene Expression Data

Grace S. Shieh , Y.C. Jiang & Yu-shan Shih

To cite this article: Grace S. Shieh , Y.C. Jiang & Yu-shan Shih (2006) Comparison of Support Vector Machines to Other Classifiers Using Gene Expression Data, Communications in Statistics - Simulation and Computation, 35:1, 241-256, DOI: 10.1080/03610910500416215

To link to this article: http://dx.doi.org/10.1080/03610910500416215

Published online: 15 Feb 2007.

Communications in Statistics—Simulation and Computation®, 35: 241–256, 2006
Copyright © Taylor & Francis Group, LLC
ISSN: 0361-0918 print/1532-4141 online
DOI: 10.1080/03610910500416215

Statistics in Genetics

Comparison of Support Vector Machines to Other Classifiers Using Gene Expression Data

GRACE S. SHIEH[1], Y.C. JIANG[1], AND YU-SHAN SHIH[2]

[1]Institute of Statistical Science, Academia Sinica, Taipei, Taiwan
[2]Institute of Statistical Science, National Chung Cheng University, Chia-Yi, Taiwan

Support vector machines (SVMs) was shown to outperform Fisher's linear discriminant analysis and two classification trees (C4.5 and MOC1) in binary classification of microarray gene expression data (MGED) (Brown et al., 2000; Furey et al., 2000). However, multiclass classification is more commonly encountered in identifying tumor subtypes using MGED. Using MGED, Dudoit et al. (2002) showed that diagonal linear discriminant analysis (DLDA) outperformed other linear and quadratic discriminants, nearest neighbor, and classification trees with univariate splits. It is of interest, therefore, to compare performance of SVMs to DLDA and the latest two classification trees with linear splits, which performered better than trees with univariate splits, in multiclass classification of MGED.

Furthermore, the performance of SVMs with different types of kernels were studied by three types of multiclass MGED. Finally, we investigate how irrelevant and correlated variables (features) influence the performance of the three classifiers. Some suggestions are made for multiclass classification of MGED.

Keywords Classification; Machine learning; Microarray gene expression data; Support vector machines; Tumor class.

Mathematics Subject Classification 62P10; 62H30.

1. Introduction

DNA microarray experiments can simultaneously generate several thousands, or even tens of thousands, of gene expression data during one experiment. A challenging problem relating to these data is that the number of features (or variables, in statistics) is much larger than the number of experiments (samples). Some of these experiments may be conducted to monitor each gene under the same

Received November 29, 2004; Accepted August 18, 2005

Address correspondence to Grace S. Shieh, Institute of Statistical Science, Academia Sinica, Nan-Gang dist., Taipei 115, Taiwan; E-mail: gshieh@stat.sinica.edu.tw

environment, but in different tissues, for example, in cancerous tissues versus normal ones (DeRisi et al., 1996; Golub et al., 1999). DNA microarray experiments have been shown to provide a new method in the diagnosis of cancer via gene expression patterns (Golub et al., 1999).

Recently, SVMs have become very popular in classification problems in the area of Bioinformatics (Brown et al., 2000; Furey et al., 2000; Golub et al., 1999). These latter two studies applied 2-class SVMs to classify protein and tumor subclasses, respectively, and SVMs were shown to perform well with binary classification of MGED. However, multiclass classification is more commonly encountered in identifying tumor subtypes using MGED. It is known that classification trees with univariate splits did not perform as well as those with linear splits (Kim and Loh, 2001; Lim et al., 2000). Thus it is of interest to study how the latest two classification trees with both splits, QUEST (Loh and Shih, 1997) and CRUISE (Kim and Loh, 2001), will perform when compared to SVMs for multiclass classification of MGED.

We compared SVMs, QUEST, and CRUISE using three types of MGED sets, and these datasets were leukemia (Golub et al., 1999), breast cancer (Sorlie et al., 2001), and NCI 60 cell lines (Scherf et al., 2000). The data sets were characterized according to their large, medium, and small sub-class sample sizes, respectively. We then compare these three classifiers to DLDA, the best in Dudoit et al. (2002) where classification trees with univariate splits and nearest neighbor were also studied using three sets of cancer gene expression data.

Next, we conducted two types of experiments: (a) adding housekeeping genes, and (b) adding correlated genes to a few top ranked genes selected by the criterion in Sec. 3.3. Results of Experiment (a) can elucidate the effect of irrelevant features (housekeeping genes) on the performance of SVMs, QUEST, CRUISE, and DLDA. Experiment (b) can clarify how correlated features affect the four classifiers. We note that interactions (correlations) do exist among genes. The findings from Experiments (a) and (b) can provide practitioners with guidelines when applying the four classifiers to MGED.

In Sec. 2, we introduce three types of gene expression datasets: leukemia, breast cancer, and NCI 60 cell lines dataset. Data-preprocessing are presented in Secs. 2.2 and 2.3. We then introduce the three classifiers (SVMs, QUEST, and CRUISE) in Sec. 3. Variable (feature) selection criteria and study design are laid out in Secs. 3.3 and 3.4. In Sec. 4, the performance of the three classifiers are compared with the three datasets; they are further compared to DLDA as studied in Dudoit et al. (2002). Since there is no theoretical basis on types of kernels being used, we also compare the performance of SVMs with three different types of kernels. In Sec. 5, we show how irrelevant variables (genes) and correlated variables affect the performance of the four classifiers. Suggestions for multiclass classification of MGED are given in Sec. 6.

2. Data and Order Constraints

We compared the performance of SVMs, QUEST, and CRUISE with three publicly available MGED sets: the leukemia dataset (Golub et al., 1999), the breast cancer (Sorlie et al., 2001), and the NCI 60 dataset (Scherf et al., 2000). To further compare the three classifiers to DLDA and to supplement the study by Dudoit et al. (2002), we used the same data imputation, standardization schemes, and variable (feature) selection method used in Dudoit et al. (2002).

2.1. *Datasets*

2.1.1. Leukemia dataset. The leukemia dataset was first employed to identify two types of acute leukemias: acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML) (Golub et al., 1999). This data set can be down-loaded from http://www-genome.wi.mit.edu/MPR. Affymetrix high-density oligonucleotide microarrays were hybridized with RNA samples from bone marrow mono-nuclear cells. Gene expression levels were measured from these arrays. There were 47 cases of ALL (38 B-cell and 9 T-cell ALL) and 25 cases of AML.

An R program, available at the above website, comprises the following pre-processing steps: (1) excluding genes whose intensity is below 100 or above 16,000, and (2) excluding genes whose max/min < 5 or $(\max - \min) \leq 500$, where min and max denote the minimum and maximum of expression levels of a gene across 72 samples. After the pre-processing steps, 3,571 genes were kept and a base 10 logarithm was taken before further analysis.

2.1.2. Breast cancer dataset. Eighty-five samples, originally derived from Stanford Microarray Database, were down-loaded from the website http://genome-www4.stanford.edu/MicroArray/SMD. All experiments, as well as the production of microarrays, were performed as described in Perou et al. (2000). There were 78 breast carcinomas (71 ductal, 5 lobular, and 2 ductal carcinomas in situ), 3 fibroadenomas, and 4 normal breast samples. The list of all 85 samples (microarray profiles) with at least 9,216 genes and their clinical data has been published as supporting information on the PNAS web site www.pnas.org. Sorlie et al. (2001) incorporated clinical outcomes (survival, survival time, relapse, tumor category, node status, tumor grade, and metastasis) and then applied average-linkage hierarchical clustering to cluster the 85 samples into six subclasses. These subclasses are basal-like (14 samples), ERBB2+ (11 samples), normal basal-like (13 samples), luminal subtype A (32 samples), subtype B (5 samples), and subtype C (10 samples).

The downloaded data were filtered and normalized by ScanAlyze version 2.5 and Genepix Pro 5.0. Although each array consisted of at least 9,216 genes, we used 6,228 genes which were present in all 85 samples. For those genes which had replications in one array, we used the average to represent that gene's expression level.

2.1.3. NCI 60 dataset. This data came from the National Cancer Institute's drug screen study (Scherf et al., 2000) and was downloaded from http://discover.nci.nih.gov/nature2000/. The cell lines were derived from the following tumors: 7 breast, 6 central nervous system, 7 colon, 6 leukemia, 8 melanoma, 9 non small-all-lung-carcinoma, 6 ovarian, 2 prostate, 8 renal, and 1 unknown. The data were measured from cDNA microarrays which were hybridized with mRNA samples from one of the 60 cell lines (red-fluorescent dye cy5) and mixtures of mRNA from 12 of the cell lines (green-fluorescent dye cy3). A base 2 logarithm of cy5/cy3 fluorescence was used in the analysis. We excluded the 2 prostate and 1 unknown cell lines due to their small class sizes.

2.2. *Imputation of Missing Data*

Following Dudoit et al. (2002), we applied k nearest neighbor method with $k = 5$ to impute missing entries. For a given gene, we computed its correlation with other

$p - 1$ genes, and we imputed its missing entry by the average of k nearest genes that have complete data for this entry.

2.3. *Data Standardization*

Data (arrays) were standardized to mean 0 and variance 1 across variables (genes). With data having been standardized this way, the correlation between the gene expression profiles of two samples can be measured by their Euclidean distance.

3. Classification Methods

3.1. *SVMs*

SVMs were originated from the statistical learning theory of Vapnik and co-workers in the 1970s (Vapnik, 1998). SVMs have been popular since the 1990s due to the advancement of modern computing. When used in classification, SVMs separate binary labeled training data by constructing a hyperplane, which separates class members from non members. The hyperplane, called a maximum margin hyperplane, is maximally distant from both members and non members. When the data are not linear-separable, SVMs map the data by a function ϕ into a higher dimensional space (called a feature space), and define a separating hyperplane there. The kernels of the SVMs, $k(\mathbf{x_i}, \mathbf{x_j}) = \phi(\mathbf{x_i})^T\phi(\mathbf{x_j})$, realize a nonlinear mapping to a feature space. The hyperplane found by an SVM in the feature space corresponds to a decision boundary in the input space (Cristianini and Shawe-Taylor, 2000).

Let $(\mathbf{x_i}, y_i)$, $i = 1, \ldots, n$, be a training set of gene expression profile-label pairs where $\mathbf{x_i} \in \mathbf{R}^q$ and $y_i \in \{1, -1\}^n$, SVMs is equivalent to solving the optimization problem of $\min(w^T w/2 + C\sum_{i=1}^{n} \xi_i)$, subject to the constraint that $y_i(w^T\phi(\mathbf{x_i}) + b) \geq 1 - \xi_i$ and $\xi_i \geq 0$, where w is the normal vector of the separating hyperplane and C is the penalty parameter of the distance of each $\mathbf{x_i}$ to the separating hyperplane. The kernel measures the similarity between two gene expression profiles $\mathbf{x_i}$ and $\mathbf{x_j}$.

Three types of kernels, linear, polynomial, and radial basis, were studied; they are denoted by SVM(L), SVM(P), and SVM(R), respectively. They assume the following forms: $k(\mathbf{x_i}, \mathbf{x_j}) = \mathbf{x_i} \cdot \mathbf{x_j}$, $k(\mathbf{x_i}, \mathbf{x_j}) = (\mathbf{x_i} \cdot \mathbf{x_j} + 1)^d$, $d = 2, \ldots$ and 10, and $k(\mathbf{x_i}, \mathbf{x_j}) = \exp(-\|\mathbf{x_i} - \mathbf{x_j}\|)^2/2\sigma^2)$, where σ is the standard deviation of the Gaussian distribution. The parameter σ in the radial kernel is a scaling parameter that penalizes the dissimilarity, namely a small value of σ gives a big dissimilarity value and vice versa.

We applied a multiclass SVM algorithm (Hsu and Lin, 2002) to classify the three datasets; the algorithm incorporated the "one-against-one" method to extend the 2-class SVMs to multiclass SVMs. To tune the parameters C and degree d in SVM(P), we applied the algorithm grid.py from http:/www.csie.ntu.edu.tw/~cjlin/lib svm/index.html and the package gnuplot from http://www.gnuplot.info/. Grid.py utilizes n-fold cross validation accuracy of test set to select the tuning parameters (C, σ^2) of SVM(R). We wrote a program to do stratified sampling and 3-fold cross-validation, and we further modified grid.py for SVM(P) to tune (C, d).

3.2. *Quest and Cruise*

QUEST (Loh and Shih, 1997) and CRUISE (Kim and Loh, 2001) are two classification trees that were proposed as alternatives to CART (Breiman et al., 1984).

Classification trees are capable of exploiting and revealing interactions between features (attributes). QUEST yields trees with binary splits. It uses statistical tests, (ANOVA F-test and Pearson chi-square test) to select a splitting feature, so that the method is almost unbiased when a single feature is preferred to form the split. QUEST applies a 2-class LDA to generate linear splits which usually yield better accuracy. On the other hand, CRUISE yields trees with multiway splits. It also uses Pearson chi-square tests to select splitting features. There are two univariate split methods in CRUISE called 1D and 2D. The former is similar to that of QUEST and the later can detect pairwise interactions between features. CRUISE provides linear splits using multiclass LDA. Both QUEST and CRUISE use the pruning method in Breiman et al. (1984) to select a subtree as the final classification model. The parameters used in QUEST and CRUISE are set to be the same as those in Lim et al. (2000) and Kim and Loh (2001). For both trees, the linear split was shown to be better than the univariate splits in terms of accuracy.

3.3. *Criterion for Variable (Feature) Selection*

We used the same variable (gene) selection criterion as Dudoit et al. (2002), namely the ratio of the genes' between-group to the within-group sum of squares (BSS/WSS) with training data. This ratio compares 'the distance of the center of each class to the over-all center' to 'the distance of each gene to its class center'. BSS/WSS for a given gene j has the form:

$$\frac{BSS(j)}{WSS(j)} = \frac{\sum_i \sum_k I(y_i = k)(\bar{x}_{kj} - \bar{x}_{\cdot j})^2}{\sum_i \sum_k I(y_i = k)(x_{ij} - \bar{x}_{kj})^2}, \tag{3.1}$$

where $\bar{x}_{\cdot j}$ denotes the average expression level of gene j across all samples and $\bar{x}_{kj}$ denotes the average expression level of gene j belonging to class k. The performances of SVMs, QUEST, and CRUISE, based on the top-p genes selected by BSS/WSS using training data, where $p = 10, 30, 40$ or 200, are compared in Sec. 4.1.

3.4. *Study Design*

The main purpose of this study is to compare the performance of classifiers, not to estimate error rates, so we followed the 2:1 scheme in Dudoit et al. (2002) to increase the test set size to one third of the data instead of commonly used 10-fold cross validation. In each class, we employed a stratified sampling scheme to generate a set of data from the original. The stratified sampling scheme randomly samples two thirds of the class members as the learning set, while the rest are left in the test set. Thus we can compare performances of the three classifiers to those studied in Dudoit et al. (2002), and can avoid extremely small learning sets.

Next, the top-p ranked variables (genes) were selected using training data and then they were applied to the test data. For example, $p = 10$, 40, and 200 were used for leukemia data set. To study the effect of decreasing and increasing number of variables, $p = 10$ and $p = 200$ were also studied in Sec. 4.

We repeated the above procedures 150 times. The median and upper quantile of 150 error rates were reported in Table 1. Each test set error rate was obtained after applying the predictors trained from learning data.

Table 1
Medians and 75 percentiles of the 150 error rates of SVMs, QUEST, CRUISE, and DLDA applied to the leukemia, breast cancer, and NCI 60 datasets, where K denotes the number of classes

		SVMs		QUEST		CRUISE			
Dataset		Linear	Radial	Linear	Univariate	Linear	1D	2D	DLDA
Leukemia $\begin{pmatrix}K = 2\\ p = 10\end{pmatrix}$	Median	0.04	0.04	0.04	0.08	0.04	0.08	0.08	0.04
	Upper quartile	0.08	0.05	0.08	0.13	0.08	0.13	0.13	0.04
$\begin{pmatrix}K = 2\\ p = 40\end{pmatrix}$	Median	0.04	0.04	0.13	0.08	0.13	0.08	0.13	0.04
	Upper quartile	0.08	0.07	0.21	0.13	0.17	0.13	0.17	0.04
$\begin{pmatrix}K = 2\\ p = 200\end{pmatrix}$	Median	0.04	0.04	0.04	0.08	0.04	0.08	0.21	0.04
	Upper quartile	0.05	0.05	0.04	0.13	0.04	0.13	0.29	0.04
Leukemia $\begin{pmatrix}K = 3\\ p = 10\end{pmatrix}$	Median	0.05	0.05	0.04	0.08	0.04	0.08	0.04	0.04
	Upper quartile	0.08	0.08	0.08	0.13	0.04	0.13	0.13	0.08
$\begin{pmatrix}K = 3\\ p = 40\end{pmatrix}$	Median	0.04	0.04	0.17	0.08	0.13	0.13	0.08	0.04
	Upper quartile	0.05	0.05	0.25	0.17	0.17	0.17	0.13	0.08
$\begin{pmatrix}K = 3\\ p = 200\end{pmatrix}$	Median	0.04	0.04	0.17	0.13	0.04	0.13	0.13	0.04
	Upper quartile	0.05	0.06	0.37	0.17	0.04	0.17	0.17	0.04
Breast cancer $\begin{pmatrix}K = 6\\ p = 10\end{pmatrix}$	Median	0.28	0.24	0.28	0.31	0.21	0.34	0.34	0.28
	Upper quartile	0.35	0.30	0.31	0.34	0.24	0.38	0.41	0.34
$\begin{pmatrix}K = 6\\ p = 40\end{pmatrix}$	Median	0.21	0.19	0.52	0.31	0.41	0.34	0.38	0.21
	Upper quartile	0.26	0.24	0.58	0.34	0.48	0.38	0.45	0.28
$\begin{pmatrix}K = 6\\ p = 200\end{pmatrix}$	Median	0.16	0.19	0.48	0.34	0.10	0.34	0.41	0.17
	Upper quartile	0.21	0.24	0.52	0.41	0.14	0.38	0.47	0.23
NCI 60 $\begin{pmatrix}K = 8\\ p = 10\end{pmatrix}$	Median	0.59	0.65	0.70	0.55	0.55	0.70	0.70	0.50
	Upper quartile	0.68	0.74	0.80	0.65	0.60	0.75	0.80	0.60
$\begin{pmatrix}K = 8\\ p = 30\end{pmatrix}$	Median	0.47	0.53	0.55	0.55	0.73	0.70	0.70	0.40
	Upper quartile	0.55	0.63	0.65	0.60	0.80	0.75	0.75	0.45
$\begin{pmatrix}K = 8\\ p = 200\end{pmatrix}$	Median	0.36	0.50	0.50	0.60	0.35	0.70	0.70	0.35
	Upper quartile	0.46	0.55	0.55	0.65	0.40	0.80	0.80	0.40

4. Classifier Performance Comparisons

We applied SVMs, QUEST, and CRUISE to the leukemia, breast cancer, and NCI 60 datasets. For the leukemia dataset, the classifiers' ability to distinguish ALL from AML (a 2-class problem) and to distinguish among ALL B-cell, ALL T-cell, and AML (a 3-class problem) are compared. For the breast cancer and the NCI 60 datasets, we compared the classifiers' ability to correctly separate members of test sets into 6 and 8 classes, respectively. We further compared the three classifiers with the results of diagonal linear discriminant analysis (DLDA). DLDA is the best in terms of low prediction error rates among the three types of classifiers studied in Dudoit et al. (2002). Let $pr(\mathbf{x} \mid y = k)$ be the conditional density of gene $\mathbf{x}$ with class label k. DLDA is a special case of maximum likelihood (ML) discriminant rule. Once the conditional density is assumed (for example multivariate normal), ML discriminant rule predicts a gene ($\mathbf{x}$)'s class by that which maximizes the likelihood of $\mathbf{x}$. DLDA further assumes that the class densities have the same diagonal covariance matrix and uses the pooled estimate of the covariance matrix from a learning set. For details, we refer to Dudoit et al. (2002).

4.1. *Comparison of SVMs, QUEST, and CRUISE*

Box plots, displaying the distributions, of the 150 error rates for each classifier using a 2:1 scheme in 150 stratified samplings, can be obtained from the corresponding

Table 2
The average of 150 error rates of SVMs with different kernels applied to stratisfied samplings from the leukemia dataset

Number	(2-class)				(3-class)			
of genes	Linear	Radial	**d**	Polynomial	Linear	Radial	**d**	Polynomial
10	0.04	0.03	2	0.03	0.02	0.01	3	0.02
20	0.04	0.02	4	0.03	0.04	0.02	2	0.00
30	0.04	0.03	3	0.03	0.03	0.02	2	0.03
40	0.04	0.03	2	0.03	0.04	0.02	2	0.01
50	0.04	0.02	4	0.02	0.03	0.02	4	0.02
60	0.04	0.02	3	0.02	0.03	0.02	2	0.03
70	0.04	0.02	2	0.02	0.03	0.02	2	0.03
80	0.04	0.02	3	0.03	0.03	0.02	2	0.01
90	0.04	0.02	2	0.02	0.03	0.02	4	0.01
100	0.04	0.02	2	0.02	0.03	0.02	2	0.02
120	0.03	0.02	4	0.02	0.02	0.02	4	0.02
140	0.04	0.02	2	0.01	0.04	0.02	2	0.02
160	0.03	0.02	3	0.02	0.03	0.02	3	0.02
180	0.04	0.02	4	0.02	0.04	0.02	4	0.03
200	0.04	0.02	4	0.03	0.04	0.02	4	0.03
1786	0.02	0.02	5	0.02	0.03	0.02	5	0.03
3571	0.02	0.02	5	0.02	0.03	0.03	4	0.03

d: the tuned value of the degree parameter in polynomial kernel.

Table 3
The average of 150 error rates of SVMs with different kernels applied to stratified samplings from the breast cancer dataset

Number of genes	Linear	Radial	**d**	Polynomial
10	0.18	0.23	2	0.28
20	0.11	0.16	2	0.18
30	0.07	0.10	2	0.16
40	0.05	0.09	2	0.14
50	0.08	0.08	2	0.10
60	0.06	0.09	2	0.07
100	0.03	0.06	2	0.07
150	0.03	0.07	2	0.10
200	0.03	0.05	2	0.09
250	0.04	0.05	2	0.08
300	0.04	0.05	2	0.08
400	0.04	0.04	2	0.07
500	0.03	0.03	2	0.06
600	0.03	0.03	2	0.06
700	0.05	0.03	2	0.07
800	0.03	0.04	2	0.07
900	0.04	0.04	2	0.07
1000	0.04	0.04	2	0.08

d: the tuned value of the degree parameter in polynomial kernel.

author. Media and 75th-percentiles of the 150 error rates are summarized in Table 1. We shall compare SVM(L) to QUEST and CRUISE with the linear combination splitting method since their error rates decrease significantly as the number of features (p) increases.

Let $A \succ B$ ($A \equiv B$) denote that the prediction error rates of A are lower than (or equivalent to) those of B. In general, SVMs $\succ$ CRUISE $\succ$ QUEST, except in eight cases where SVMs $\equiv$ CRUISE for all datasets with $p = 10$ and $p = 200$, and QUEST $\succ$ CRUISE for the NCI 60 dataset with $p = 30$. The error rates for the leukemia (both 2-class and 3-class), breast cancer, and NCI 60 datasets were very low, moderate, and very high, respectively. When 10 features (genes) were used, the medians of 150 experiments using the three classifiers ranged from 0.04 to 0.08 (for the leukemia data), 0.28 to 0.34 (for the breast cancer data), and 0.55 to 0.68 (for the NCI 60 data). When 200 genes were used, the medians of 150 experiments using the three classifiers ranged from 0.04 to 0.21, from 0.16 to 0.48, and from 0.36 to 0.50, respectively. The error rates for the NCI 60 data are in general very high. This is perhaps due to the large number of classes with a limited number of training data. We observed the trend from the three datasets, showing that the prediction error increased as the average sub-sample size decreased.

4.2. *Comparison of SVMs, QUEST, and CRUISE to DLDA*

We further compared SVM(L), QUEST, and CRUISE to DLDA in Dudoit et al. (2002). Taking two random errors (8%) into account, we can conclude that for

Table 4
The average of 150 error rates of SVMs with different kernels applied to stratified samplings from the NCI 60 dataset

Number of genes	Linear	Radial	**d**	Polynomial
10	0.57	0.31	2	0.45
20	0.43	0.28	3	0.37
30	0.37	0.22	3	0.29
40	0.34	0.17	2	0.28
50	0.33	0.17	2	0.22
60	0.32	0.20	2	0.22
70	0.29	0.18	2	0.23
80	0.30	0.21	2	0.24
90	0.29	0.19	2	0.23
100	0.28	0.21	2	0.24
120	0.26	0.17	2	0.24
140	0.26	0.16	2	0.18
160	0.27	0.20	2	0.23
180	0.27	0.18	2	0.26
200	0.28	0.19	2	0.26
708	0.26	0.18	2	0.24
1416	0.27	0.22	2	0.32

d: the tuned value of the degree parameter in polynomial kernel.

the NCI 60 dataset, when $p = 10$, DLDA $\succ$ SVM(L) $\equiv$ CRUISE $\succ$ QUEST; when $p = 200$, DLDA $\equiv$ SVM(L) $\equiv$ CRUISE $\succ$ QUEST. When $p = 30$, the trend is similar to that of $p = 200$ except QUEST $\succ$ CRUISE. The bad performance of CRUISE with $p = 30$ may be due to perfect classification for training sets and poor classification for test sets. For the leukemia (both 2-class and 3-class cases) and the breast cancer datasets, DLDA $\equiv$ SVM(L) $\succ$ CRUISE $\succ$ QUEST when $p = 10, 40$ and 200.

4.3. *Performance of SVMs with Three Different Kernel Types*

There is no theoretical guidelines in choosing types of kernels when applying SVMs to classification problems. Thus we compared the performance of SVMs with different kernel types using three types of MGED. The results with various numbers of features (p) are shown in Tables 2, 3, and 4. For a fixed p, we applied SVMs to 150 sets of data simulated from the stratified sampling scheme in Sec. 3.4. For the leukemia dataset, SVMs performed the same with three different types of kernel. The error rates were very low (ranging from 0.01 to 0.04) even when $p = 10$. This suggests that using the top-10 ranked genes is good enough for the SVMs to classify the leukemia data into either 2 classes or 3 classes.

In the case of the breast cancer dataset, when $10 \leq p \leq 40$, SVM(L) $\equiv$ SVM(R) $\succ$ SVM(P). When $p \geq 50$, all three kernels performed about the same. For the NCI 60 dataset, when $p \leq 60$, SVM(R) $\succ$ SVM(P) $\succ$ SVM(L), and when $p \geq 70$, SVM(R) $\succ$ SVM(P) $\equiv$ SVM(L).

5. How Irrelevant and Correlated Features Affect SVMs, QUEST, CRUISE, and DLDA

The four classifiers with the top-10 ranked genes selected by *BSS*/*WSS* were shown to classify the leukemia dataset well (Table 1). Using these top-ranked genes, we conducted two sets of experiments by: (a) adding some housekeeping genes, and (b) adding genes that are correlated (collinear) to the top-10 ranked genes of the leukemia (2- and 3-class), breast cancer, and NCI 60 datasets. In each experiment, the procedure of adding genes was independently repeated 150 times. The housekeeping genes have constant gene expression levels with small noise fluctuations compared to the sample variances of the top-10 ranked genes.

In general, classifiers resistant to irrelevant features are better. We conducted Experiment a to check how robust the four classifiers are when irrelevant features are present. On the other hand, correlations among genes do exist as biological realities (Sec. 6 of Dudoit et al., 2002). Furthermore, these correlated features may affect the classifiers' estimation of parameters and thus may affect performance. Classification trees (QUEST and CRUISE) can exploit and reveal interactions between features; in particular, CRUISE(2D) can detect pairwise interaction between features. Thus, it is of interest to study how correlated features affect the performance of SVMs, QUEST, CRUISE, and DLDA.

In theory, the red and green fluorescence intensities of a housekeeping gene are the same, so their log ratio is 0. The empirical distributions of log(R/G) of the three data sets are roughly normal, thus we generated housekeeping genes independently from a Gaussian distribution with a mean 0 and small variances. For the NCI 60 dataset, 1/16 of the median of sample variances of the top-10 ranked genes was used for the variances of housekeeping genes, while for the other two datasets it was 1/64. In Experiment a(i)–(iii), we added 5, 10, and 190 housekeeping genes to the top-10 ranked genes in each data set. This made up 33.3%, 50%, and 95% of the data housekeeping genes. In Experiment a(iv), we added 10 groups of 19 housekeeping genes to the top-10 ranked genes. The variances of the 10 groups were the same but the means were all different. Comparing the results of Experiments a(iv) and (iii), we can discern whether housekeeping genes with different means (location parameters) affect classification results. Specifically, we conducted the following experiments:

(a)(i) $g_{10+i} \sim N(0, \sigma^2)$, $i = 1, \ldots, 5$, where $\sigma = 0.125$, 0.15, 0.25 for the leukemia, breast cancer, and NCI 60 dataset, respectively.

(a)(ii) $g_{10+i} \sim N(0, \sigma^2)$, $i = 1, \ldots, 10$.

(a)(iii) $g_{10+i} \sim N(0, \sigma^2)$, $i = 1, \ldots, 190$.

(a)(iv) $g_{[i-1]19+j} \sim c_i + N(0, \sigma^2)$, $i = 1, \ldots, 10$, $j = 1, \ldots, 19$ and c_i's are constants ranging from 0.0 to 1.0, −1.0, to 1.0, and −0.025 to 0.025 with equal spacings for the leukemia, breast cancer, and NCI 60 datasets, respectively.

Similarly, we conducted the following experiments which added 5, 10, 30, and 190 correlated features to the top-10 ranked genes in each dataset. This made up 33.3%, 50%, 75%, and 95% of correlated features (genes). In Experiments b(i)–(ii), we added 5 and 10 genes which have the same means and variances as the top-5 and top-10 genes, respectively. In Experiments b(iii)–b(iv), we added 10 groups of 3 genes (19 genes), where genes in the *i*th group have the same mean and variance as the top-*i* ranked gene, $i = 1, \ldots, 10$.

Table 5
The average of 150 error rates of SVMs applied, repeatedly for 150 times, to the experiment (a) adding housekeeping genes or (b) adding correlated genes to the top-10 ranked genes

	Leukemia							
	(2-class)		(3-class)		Breast cancer		NCI 60	
Dataset	L	R	L	R	L	R	L	R
a(0)	0.05	0.04	0.06	0.06	0.30	0.24	0.59	0.64
a(i)	0.05	0.04	0.07	0.06	0.30	0.24	0.59	0.67
a(ii)	0.05	0.04	0.06	0.06	0.30	0.24	0.59	0.69
a(iii)	0.05	0.05	0.06	0.11	0.26	0.41	0.64	0.88
a(iv)	0.05	0.05	0.06	0.11	0.26	0.41	0.65	0.88
b(0)	0.05	0.04	0.07	0.06	0.30	0.24	0.59	0.64
b(i)	0.06	0.05	0.07	0.08	0.33	0.32	0.64	0.73
b(ii)	0.05	0.05	0.08	0.10	0.34	0.36	0.66	0.77
b(iii)	0.05	0.05	0.07	0.06	0.30	0.25	0.59	0.77
b(iv)	0.05	0.05	0.06	0.11	0.26	0.41	0.64	0.88

L: SVMs with the linear kernal.
R: SVMs with the radial kernal.
a(0): only the top-10 ranked genes used.
b(0): only the top-10 ranked genes used.

Specifically, we conducted the following experiments:

(b)(i) $g_{10+i} \sim g_{(i)} + N(0, s^2(g_{(i)}))$, where $i = 1, \ldots, 5$ and $\bar{g}_{(i)}$ and $s^2(g_{(i)})$ denote the sample mean and variance of the top-i ranked gene $g_{(i)}$, respectively.

(b)(ii) $g_{10+i} \sim g_{(i)} + N(0, s^2(g_{(i)}))$, where $i = 1, \ldots, 10$.

(b)(iii) $g_{[i-1]3+j} \sim g_{(i)} + N(0, s^2(g_{(i)}))$, where $i = 1, \ldots, 10$, and $j = 1, \ldots, 3$.

(b)(iv) $g_{[i-1]19+j} \sim g_{(i)} + N(0, s^2(g_{(i)}))$, where $i = 1, \ldots, 10$, and $j = 1, \ldots, 19$.

Since default values of parameters, for instance node size and estimated priors, of QUEST and CRUISE were used, default values of parameters in the SVM(R) were also used; namely $C = 1$ and $\sigma^2 =$ (number of subclasses)/2.

Table 5 shows that for all three datasets, the performance of SVM(L) was not affected either by the presence of housekeeping genes or by the presence of correlated genes. However, with leukemia (3-class), breast cancer, and the NCI 60 datasets when 95% or more housekeeping genes were present, the performance of SVM(R) were worsened. The error rates of a(iii) and (iv) were greater than two random error 8%. Results from Experiment b show that when 33% or more correlated genes were present, the performance of SVM(R) was worsened with the breast cancer and NCI 60 datasets. There were 85 and 60 samples in the two datasets, but they were classified into 6- and 8-classes, respectively. With the averaged subclass samples being 14 or fewer, the classification performance of SVM(R) was worsened when 33% or more correlated features were present, and the error rates of SVM(R) increased as more correlated features present. We also conducted Experiment a and b using SVM(R) with tuned parameters was as good

Table 6
The average of 150 error rates of QUEST applied, repeatedly for 150 times, to the experiment (a) adding housekeeping genes or (b) adding correlated genes to the top-10 ranked genes

	Leukemia							
	(2-class)		(3-class)		Breast cancer		NCI 60	
Dataset	L	U	L	U	L	U	L	U
a(0)	0.06	0.09	0.09	0.12	0.36	0.38	0.68	0.61
a(i)	0.06	0.09	0.11	0.12	0.43	0.38	0.69	0.63
a(ii)	0.06	0.09	0.14	0.12	0.47	0.37	0.72	0.62
a(iii)	0.06	0.10	0.33	0.12	0.62	0.38	0.81	0.67
a(iv)	0.06	0.10	0.32	0.12	0.64	0.39	0.80	0.66
b(0)	0.06	0.09	0.09	0.12	0.36	0.38	0.68	0.61
b(i)	0.06	0.09	0.11	0.12	0.42	0.38	0.69	0.62
b(ii)	0.06	0.09	0.14	0.12	0.47	0.38	0.72	0.63
b(iii)	0.20	0.09	0.28	0.12	0.56	0.38	0.73	0.63
b(iv)	0.06	0.10	0.22	0.12	0.55	0.40	0.69	0.67

L: QUEST with the linear split.
U: QUEST with the univariate split.
a(0): only the top-10 ranked genes used.
b(0): only the top-10 ranked genes used.

as SVM(L) in all cases. Furthermore, SVM(R) with tuned parameters was as robust as SVM(L) against housekeeping genes and correlated genes.

Let QUEST(L) and QUEST(U) denote QUEST with linear and univariate split, respectively. Similarly, CRUISE(L), CRUISE(1D), and CRUISE (2D) denote CRUISE with linear, 1D and 2D split, respectively. Table 6 shows that for all four classification cases, QUEST(U) was not affected by adding either housekeeping genes or correlated genes (variables). However, the performance of QUEST(L) was deteriorated when 95% (50%) or more housekeeping genes were present, with leukemia 3-class and NCI60 (breast cancer) datasets. Results from Experiment b shows that with the leukemia 2-class and 3-class (breast cancer) datasets, the performance QUEST(L) was worsened when 95% (50%) or more correlated genes were present.

Similarly, Table 7 shows that the performance of CRUISE(1D) and CRUISE (2D) were not affected when housekeeping genes or correlated genes were present. However, the performance of CRUISE(L) was worsened when 50% (95%) house or more keeping genes were present with breast cancer (NCI60) dataset. Furthermore, the performance of CRUISE(L) deteriorated when 75% correlated genes were present with all datasets. We note that when 95% correlated genes were present, with all datasets, the performance of CRUISE(L) was not affected at all. This may be due to that the linear splits made by b(iv) and b(i) were similar.

Table 8 shows that the performance of DLDA was not worsened except cases in a(iii) and a(iv) (95% housekeeping genes in the data sets) with the breast cancer and the NCI60 datasets. DLDA was not affected by adding any correlated genes (variables).

Table 7
The average of 150 error rates of CRUISE applied, repeatedly for 150 times, to the experiment (a) adding housekeeping genes or (b) adding correlated genes to the top-10 ranked genes

	Leukemia											
	(2-class)			(3-class)			Breast cancer			NCI 60		
Dataset	L	1D	2D	L	1D	2D	L	1D	2D	L	1D	2D
a(0)	0.06	0.09	0.09	0.08	0.14	0.11	0.28	0.39	0.42	0.54	0.72	0.71
a(i)	0.06	0.09	0.09	0.08	0.14	0.11	0.30	0.39	0.43	0.59	0.72	0.73
a(ii)	0.07	0.09	0.09	0.08	0.14	0.11	0.34	0.39	0.43	0.61	0.74	0.73
a(iii)	0.05	0.09	0.09	0.12	0.14	0.13	0.45	0.38	0.44	0.77	0.75	0.75
a(iv)	0.06	0.10	0.10	0.11	0.14	0.13	0.46	0.39	0.44	0.79	0.76	0.75
b(0)	0.06	0.09	0.09	0.08	0.14	0.11	0.28	0.39	0.42	0.54	0.72	0.71
b(i)	0.06	0.09	0.09	0.08	0.14	0.11	0.30	0.39	0.43	0.59	0.72	0.72
b(ii)	0.07	0.09	0.09	0.08	0.14	0.12	0.34	0.39	0.43	0.61	0.73	0.73
b(iii)	0.19	0.09	0.10	0.22	0.14	0.12	0.53	0.39	0.43	0.65	0.73	0.74
b(iv)	0.05	0.10	0.10	0.07	0.14	0.14	0.26	0.38	0.46	0.55	0.75	0.75

L: CRUISE with the linear split.
1D: CRUISE with the univariate 1D split.
2D: CRUISE with the univariate 2D split.
a(0): only the top-10 ranked genes used.
b(0): only the top-10 ranked genes used.

Table 8
The average of 150 error rates of DLDA applied, repeatedly for 150 times, to the experiment (a) adding housekeeping genes or (b) adding correlated genes to the top-10 ranked genes

	Leukemia			
Dataset	(2-class)	(3-class)	Breast cancer	NCI 60
a(0)	0.03	0.06	0.30	0.50
a(i)	0.03	0.06	0.32	0.55
a(ii)	0.03	0.07	0.34	0.58
a(iii)	0.05	0.13	0.45	0.75
a(iv)	0.05	0.14	0.47	0.76
b(0)	0.03	0.06	0.30	0.50
b(i)	0.03	0.07	0.30	0.53
b(ii)	0.03	0.07	0.31	0.52
b(iii)	0.04	0.07	0.30	0.53
b(iv)	0.04	0.07	0.25	0.51

a(0) : only the top-10 ranked genes used.
b(0) : only the top-10 ranked genes used.

Overall, SVM(L) $\equiv$ SVM(R) with tuned parameters. Furthermore, SVM(L) $\succ$ DLDA in a(iii) and a(iv), and they both performed better than SVM(R), QUEST, and CRUISE in a(iii)–(iv) and b(iii)–(iv) with breast cancer and NCI60 datasets. In terms of robust against correlated genes, SVM(L) $\equiv$ DLDA with the breast cancer dataset and DLDA $\succ$ SVM(L) with the NCI60 dataset. These results are consistent with Table 1.

6. Conclusion

In general, SVMs outperformed both QUEST and CRUISE for all three types of microarray datasets whose averaged subclass sample size range from 36 (relatively large), to 14 (medium), and 7 (small). Taking two random errors (8%) into account, SVM(L) performed as well as DLDA, the best classifier in Dudoit et al. (2002) with leukemia and breast cancer datasets. Furthermore, SVM(L) performed slightly worse than DLDA with NCI60 dataset. We note that when the number of subclass is medium or large compared to the sample size, the between class boundaries of datasets tend to be complex. In this case DLDA is recommended in terms of its low classification error rates, as shown in the results of Table 1.

SVM(L) and SVM(R) with tuned parameters were stable when many housekeeping genes or correlated genes were present in MGED. However, SVM(R) with default parameters, QUEST(L) and CRUISE(L) were affected. Housekeeping genes or correlated genes do exist in microarray data. Hence when applying SVM(R), we should use tuned parameters. Furthermore, pairwise correlation of features (genes) should be calculated first when applying classification trees to MGED. Although QUEST(U), CRUISE(1D), and CRUISE(2D) were also robust against any housekeeping (irrelevant) or correlated genes that are present in MGED. However, these classification trees have higher error rates than SVM(L) when the between-class boundaries of datasets are complex. Thus we suggest that SVMs and DLDA be applied to multiclass classification of MGED. We do not know what caused the non monotone behaviors of QUEST and CRUISE in Table 1 (some of their error rates went up when p was increased); we leave this as an open problem. Recently, random forest (RF) (Breiman, 2001) has been applied to microarray data for classification of tumors (Zhang et al., 2003). It will be interesting to compare performance of RF to the classifiers studied in this article; we leave this for future study.

Acknowledgments

We thank Chi Lee for writing codes for Tables 2–5 and Drs. Y. C. Chang, C. J. Lin, and Sandrine Dudoit for helpful discussions. Thanks also go to the anonymous referee for constructive suggestions. Grace S. Shieh's research was supported in part by NSC grants 90-2118-M001-022 and 91-3112-P001-037.

References

Breiman, L. (2001). Random forests. *Mac. Learn.* 45:5–32.

Breiman, L., Friedman, J. H., Olshen, R. A., Stone, C. J. (1984). *Classification and Regression Trees*. New York: Chapman and Hall.

Brown, M. P., Grundy, W. N., Lin, D., Cristianini, N., Sugnet, C. W., Furey, T. S., Ares, M., Haussler, D. (2000). Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl. Acad. Sci. USA* 97:262–267.

Cristianini, N., Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines*. Cambridge: Cambridge University Press.

DeRisi, J., Penland, L., Brown, P., Bittner, M., Meltzer, P., Ray, M., Chen, Y., Su, Y., Trent, J. (1996). Use of a cDNA microarray to analyze gene expression patterns in human cancer. *Nature Genet.* 4:457–460.

Dudoit, S., Fridlyand, J., Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Amer. Stat. Assoc.* 97:77–86.

Furey, T. S., Cristianini, N., Duffy, N., Bednarski, D. W., Schummer, M., Haussler, D. (2000). Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics* 16:906–914.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gassenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., Lander, E. S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286:531–537.

Hsu, C. W., Lin, C. J. (2002). A comparison of methods for multiclass support vector machines. *IEEE Trans. Neural Networks* 13:415–425.

Kim, H., Loh, W.-Y. (2001). Classification trees with unbiased multiway splits. *J. Amer. Statist. Assoc.* 96:589–604.

Lim, T.-S., Loh, W.-Y., Shih, Y.-S. (2000). A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. *Mac. Learn.* 40:203–228.

Loh, W.-Y., Shih, Y.-S. (1997). Split selection methods for classification trees. *Statistica Sinica* 7:815–840.

Scherf, U., Ross, D. T., Eisen, M. B., Perou, C. M., Spellman, P., Iyer, V., Jeffrey, S. S., Van de Rijn, M., Waltham, M., Pergamenschikov, A., Lee, J. C. F., Lashkari, D., Shalon, D., Myers, T. G., Weinstein, J. N., Botstein, D., Brown, P. O. (2000). System variation in gene expression patterns in human cancer cell lines. *Nature Genet.* 24:227–234.

Sorlie, T., Perou, C. M., Tibshirani, R. et al. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl. Acad. USA* 98:10869–10874.

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

Zhang, H., Yu, C. Y., Singer, B. (2003). Cell and tumor classification using gene expression data: construction of forests. *Proc. Natl. Acad. Sci. USA* 100:4168–4172.