JOURNAL OF BIOPHARMACEUTICAL STATISTICS
Vol. 14, No. 4, pp. 1065–1084, 2004

Evaluating Methods for Classifying Expression Data

Michael Z. Man,[1,*] Greg Dyson,[2] Kjell Johnson,[1] and Birong Liao[3]

[1]Nonclinical Statistics, Pfizer Global Research and Development – Ann Arbor Laboratories, Ann Arbor, Michigan, USA
[2]Biostatistics, Allergan, Irvine, California, USA
[3]Department of Systems Biology, Lilly Research Laboratories, Indianapolis, Indiana, USA

ABSTRACT

An attractive application of expression technologies is to predict drug efficacy or safety using expression data of biomarkers. To evaluate the performance of various classification methods for building predictive models, we applied these methods on six expression datasets. These datasets were from studies using microarray technologies and had either two or more classes. From each of the original datasets, two subsets were generated to simulate two scenarios in biomarker applications. First, a 50-gene subset was used to simulate a candidate gene approach when it might not be practical to measure a large number of genes/biomarkers. Next, a 2000-gene subset was used to simulate a whole genome approach. We evaluated the relative performance of several classification methods by using leave-one-out cross-validation and bootstrap cross-validation. Although all methods perform well in both subsets for a relative easy dataset with two classes, differences in performance do exist among methods for other datasets. Overall, partial least squares discriminant analysis (PLS-DA) and support vector machines (SVM) outperform all other methods. We suggest a practical approach to take advantage of multiple methods in biomarker applications.

*Correspondence: Michael Z. Man, Nonclinical Statistics, Pfizer Global Research and Development – Ann Arbor Laboratories, 2800 Plymouth Road, Ann Arbor, MI 48105, USA; Fax: (734) 622-3153; E-mail: michael.mann@pfizer.com.

DOI: 10.1081/BIP-200035491
1054-3406 (Print); 1520-5711 (Online)

Key Words: Biomarkers; Gene approach; Whole genome approach; Leave-one-out cross-validation; Bootstrap cross-validation.

INTRODUCTION

Because of the ability to monitor the levels of thousands of steady-state RNA molecules in a high-throughput fashion, microarray technology has been widely used in genomic research (Alizadeh et al., 2000; Alon et al., 1999; Golub et al., 1999; Ross et al., 2000). Quantitative polymerase chain reaction (QPCR) and various proteomic technologies complement microarrays to give a more complete picture of expression. In the drug discovery and development process, applications of these expression technologies include target identification and validation, model characterization, lead optimization, mechanism of action study, drug metabolism study, and new indication seeking (Bumol and Watanabe, 2001; Ricci and El Deiry, 2000; Stratowa et al., 2001).

One of the most attractive applications of expression technologies in pharmaceutical industry is to predict drug efficacy or safety using biomarkers (Gunther et al., 2003). Biomarkers have great value in guiding clinical trials and developing inclusion/exclusion criteria. For a typical clinical trial, only a small number of proteins/genes/single nucleotide polymorphisms (SNP) are used as biomarkers basing on expert knowledge, disease relevance, prior validation, etc. They must be carefully chosen and validated in a large patient population (http://grants.nih.gov/grants/guide/pa-files/PA-01-043.html) because mere statistical association is not sufficient to qualify a gene/protein/SNP as a biomarker (Cantor, 1999). Financial and other constraints make it unlikely to measure large numbers of genes or proteins in a large clinical trial. For example, the limited amount of biopsy sample precludes technologies that measure large numbers of genes/proteins/SNPs. In addition, the risk of getting spurious correlation and inflated false-positive rates also increases when measuring large numbers of markers (Petricoin et al., 2002). Because of these concerns, it is prudent to plan a clinical trial with a list of relevant and validated biomarkers.

A comparison of classification methods is necessary to make accurate prediction using the expression data of biomarkers. Previous studies on classification methods are either not directly applicable to biomarker applications in clinical trials due to confounding with variable selection or do not include all the methods we intend to use, such as partial least square (PLS), random forest (RF), and support vector machines (SVM) (Ben Dor et al., 2000; Dudoit et al., 2002; Gunther et al., 2003; Lee and Lee, 2002; Ramaswamy et al., 2001; Wu et al., 2003; Yeang et al., 2001).

With applications in clinical trials using expression data of biomarkers in mind, we design our comparison in two scenarios. The first scenario models the usual biomarker application: Using small number of biomarkers that have been validated previously via multiple technologies such as microarray, QPCR, proteomic microarray, tissue array, and fluorescent in situ hybridization (Nishizuka et al., 2003; Wilkinson, 1994). The second scenario models future biomarker applications by including large numbers of "level II" biomarkers (with apparent association, but less

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

reliable or unclear pathophysiological connection) (Petricoin et al., 2002). In both scenarios, the same set of classification methods are applied to expression datasets from various public sources and are compared in terms of overall prediction errors using cross-validation techniques (see details in Methods).

METHODS

This section includes a high level overview of each of the methods used in our analyses. For the interested readers, we include standard references for the theoretical foundation of each technique.

K-Nearest Neighbors

K-Nearest Neighbors (KNN) is a classification method that predicts class membership for a new observation based on its distance to observations in the training data set. The basic idea behind KNN is that samples that fall close together in the feature space are likely to belong to the same class (Cortijo et al., 1993). Upon selecting a distance metric (e.g., Euclidean), the algorithm proceeds by selecting the k-nearest observations in the training set to the new observation. The modal class of the k-nearest observations is designated as the predicted class for the new observation (Fix and Hodges, 1951). K-Nearest Neighbors is computationally efficient and is easy to visualize and understand. Dudiot et al. (2002) demonstrated the use of KNN for lymphoma, leukemia, and cancer microarray data. For our analysis, we used the Euclidean distance metric. The number of nearest neighbors (k) was chosen via cross-validation on the ''training data set.'' For LOOCV, a ''training dataset'' was the *N-1* sample in each leave-one-out cycle; for bootstrap cross-validation, a ''training dataset'' was generated in each cycle of bootstrap sampling. Parameters were chosen from the ''training set'' in each cycle of LOOCV or bootstrap sampling (Table 1).

Principal Component Analysis—Discriminant Analysis

Discriminant analysis (DA) traces its roots as a statistical classification method to R. A. Fisher in 1936. There are numerous variants of the original idea of Fisher, including the form that we implemented: maximum likelihood. Accordingly, we assume that the predictor variables have a multivariate normal distribution within each class. A new observation is classified as belonging to the kth class if the distance of that observation is the closest to the kth class. However, it is difficult to implement this method when there are a large number of input variables. Therefore, dimension reduction technique is performed prior to DA. Principal components analysis (PCA) is a commonly used dimension reduction technique, which finds the linear combinations of input variables that produce new uncorrelated scores with maximum variance.

Similar to Xiong et al., we combine these two approaches, which performs discriminant analysis using m principal components as the inputs (Xiong et al., 2000).

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

Table 1. Implementation of classification methods. PLS was performed in SAS® (SAS Institute, Cary, NC); all other methods were implemented in R (http://www.r-project.org).

Method	50-Gene dataset	2000-Gene dataset
A. AML/ALL		
NN	1 hidden layer	NA
KNN	k chosen by cross-validation	k chosen by cross-validation
PCA–LDA	2 PCs	8 PCs
PCA–QDA	2 PCs	8 PCs
PLS	2 latent variables	6 latent variables
SVM	RBF ($\gamma = 2^{-12}$, cost $= 2^{10}$)	RBF ($\gamma = 2^{-8}$, cost $= 2^{20}$)
B. NC160		
NN	1 hidden layer	NA
KNN	k chosen by cross-validation	k chosen by cross-validation
PCA–LDA	13 PCs	13 PCs
PLS	7 latent variables	15 latent variables
SVM	RBF ($\gamma = 2^{-9}$, cost $= 2^{4}$)	RBF ($\gamma = 2^{-24}$, cost $= 2^{16}$)

Two variants of DA were used in this paper: linear discriminant analysis (LDA), which attempts to find the best linear separation in the data and (2) quadratic discriminant analysis (QDA), which attempts to find the best quadratic separation in the data. In addition, we chose the top m principal components based on cross-validation of the training data set. However, we can only perform PCA–QDA on the AML/ALL data sets due to difficulties estimating a within-class covariance matrix for the NCI60 data sets (refer to Datasets section).

Partial Least Squares

Partial least squares has been studied extensively both in theory (Frank and Friedman, 1993; Stone and Brooks, 1990) and in practice (Wold, 1995). Because of Stone and Brooks' work, PLS can be viewed as simultaneous dimension reduction and regression (Stone and Brooks, 1990). Hence, this technique is particularly useful for modeling data in which the number of descriptors (p) is greater than samples (n), or the descriptors are linearly dependent.

Technically, for data with a univariate response, PLS seeks to find successively defined latent variables (linear combinations of the original descriptors) that have maximum covariance with the response, subject to a user-specified set of constraints (Rayens, 2000). For data with a multivariate response, PLS seeks to find successively defined pairs of latent variables from the descriptor and response spaces such that each pair of latent variables has maximum covariance.

Although PLS was initially applied for the purpose of regression, it has also been applied for the purpose of discrimination (Alsberg et al., 1998; Saaksjarvi et al., 1989). Recently, Nguyen and Rocke (2002) applied PLS to several microarray gene expression data sets classifying human tumor types (Nguyen and Rocke, 2002).

For a p-group discrimination problem, a p-dimensional binary response matrix is typically used to distinguish the classes, where the jth column represents the class

membership for the *j*th group. Cross-validation is then used to determine the number of latent variables required to adequately relate the descriptors to the response. Upon determining the number of necessary latent variables, a model is built on a training set and applied to the dataset of interest. For each observation, the predicted group membership corresponds to the group associated with the column with largest predicted absolute value (Alsberg et al., 1988). As an alternative to classifying observations based on magnitude of predictions, Nguyen and Rocke (2002) applied LDA to the predicted response matrix to obtain predicted classifications (Nguyen and Rocke, 2002). For the analyses in this work, observations were classified on the basis of the largest predicted absolute value.

Neural Networks

Neural networks (NN) emerged as a statistical pattern recognition tool in the 1980s designed to mimic the fault-tolerance and learning capacity of biological neural systems (Patterson, 1996). An NN takes the input variables and uses prespecified activator functions to build a network to predict either a categorical or a continuous response (Ripley, 1996). Neural networks offer a flexible method of prediction for large datasets. In addition, NN are robust to moderate amounts of noise in the data. However, this method is computationally intensive, limiting users to approximately 1000 input variables under certain circumstances (see below). Despite possible computational difficulties, NN have been used to analyze microarray data (Gruvberger et al., 2001; Selaru et al., 2002).

In our analysis, each NN had one hidden layer and the decay parameter was chosen via cross-validation using the training dataset. To avoid local minima, we iterated the fitting procedure. The weighted average (reciprocal of the fitting criteria) of the iterated networks was taken as the output network. This output network was used to produce predictions for the test dataset. Due to computational difficulties, only the 50-gene datasets were investigated here (see Datasets section).

One-to-One (Pairwise Coupling) Algorithm

The one-to-one algorithm was designed to improve the performance of neural networks in multiclass problems (Moreira and Mayoraz, 1998; Price et al., 1994). The algorithm begins by partitioning the original multiclass problem into all possible two-class problems. Then, for each two-class problem, a classifier is built on the training data. Next, a prediction is computed for each observation in the test dataset. Finally, the scores for each observation across all two-class problems are aggregated. The maximum of the aggregated scores is the predicted class.

Random Forest

Random forest (RF) improves the classification tree method by implementing random split selection in combination with bootstrap aggregating (Breiman, 2001).

Class membership is determined by popular votes from each of the large number of trees generated from RF. The generalization error can be estimated because the misclassification is assessed out of bag (OOB). In addition, the variable importance can be estimated for each variable.

Support Vector Machines

Support vector machines methodology is based on statistical learning theory and can be traced back to Rosenblatt's perceptron (Cristianini and Shawe-Taylor, 2000; Rosenblatt, 1958). Developments in SVM generalization theory and computation optimization by Vapnik and others have helped to advance SVM as a preferred method in difficult classification problems such as voice recognition (Cristianini and Shawe-Taylor, 2000; Vapnik, 1998). Support vector machines are defined as "learning systems that use a hypothesis space of linear functions in a high dimensional feature space, trained with a learning algorithm from optimization theory" (Cristianini and Shawe-Taylor, 2000). The SVM is suitable for classifying high dimensional data, even though SVM does not completely escape from the curse of dimensionality (Hastie et al., 2001).

Although SVM was initially developed for binary classification, it is possible to analyze multiclass data using SVM. The one-to-one algorithm, initially developed for NN, is one of the best performers in classifying multiclass data (Hsu and Lin, 2002). Recently, SVM was used to classify tumors, cancers, and cell lines using microarray expression data (Ben Dor et al., 2000; Brown et al., 2000; Chow et al., 2001; Furey et al., 2000; Lee and Lee, 2002; Moler et al., 2000; Ramaswamy et al., 2001; Yeang et al., 2001).

In analyzing the AML/ALL and NCI60 datasets, we used the radial basis function (rbf) as an SVM kernel and chose the best combination of parameters via grid search ($\gamma = [2^{-20}, 2^{2}]$, cost $= [2^{-2}, 2^{20}]$). Because the NCI60 data have multiple classes, we used the one-to-one algorithm to obtain prediction for new observations.

Evaluation Criteria

Because of the small sample size of the chosen datasets, we decided to employ leave-one-out cross-validation (LOOCV) and a resampling method based on the bootstrap (BS) to estimate prediction error rates for each classification technique. In LOOCV, one observation is removed from the original dataset and a model is built on the remaining $n-1$ observations. Subsequently, the model is used to predict the response for the held-out observation. This process is repeated for each of the remaining $n-1$ observations. The n leave-one-out predictions can then be compared with the observed responses to assess the predictive ability of the method. This method provides a near unbiased estimate of the prediction error rate (Aeberhard et al., 2002; Ambroise and McLachlan, 2002; Hastie et al., 2001). See Neter et al. for more specific details about LOOCV (Neter et al., 1996). Although more complicated, the bootstrap approach enables us to also obtain an estimate of the prediction error rate as well as confidence bounds about the rate

(Efron and Tibshirani, 1993; Wu et al., 2003). The BS method for estimating prediction error rates for each model can be found in Efron and is described here (Efron and Tibshirani, 1993). The first step in obtaining an estimate of prediction error is to compute the apparent error rate. The apparent error rate is the proportion of misclassified observations, when applying the original dataset to a model built from the original dataset. After obtaining the apparent error rate*, the following steps are performed 2000 times:

(1) Generate a stratified bootstrap sample of the original data by fixing the number of observation in each class constant.
(2) Build a model on the bootstrap sample.
(3) Apply the model to the original data set and obtain the error rate.
(4) Apply the model to the bootstrap sample and obtain the apparent error rate for the sample.
(5) Subtract the apparent error rate obtained in Step 4 from the error rate obtained in Step 3. This value is known as the optimism.

Upon iterating through Steps 1–5, the average of the 2000 estimates of optimism are added to the apparent error rate* to obtain an estimate of prediction error.

For the random forest, the prediction error is calculated as the out of bag estimate based on large number of trees (500 in our study) (Breiman, 2001). The RF procedure is repeated 200 times to estimate the error variability.

The BS method also enables us to keep track of the proportion of times each observation is incorrectly predicted. As we show in our examples, these frequencies can help to identify observations that may have been incorrectly labeled.

Datasets

To model two scenarios in clinical studies using biomarkers, we chose to analyze two defined subsets for each dataset described below. The first subset contains 50 genes that are chosen on the basis of the ratios of between sum-of-squares and within sum-of-squares

$$\sum_i \sum_k I(y_i = k)(\bar{x}_k - \bar{x}_\cdot)^2 \Big/ \sum_i \sum_k I(y_i = k)(x_i - \bar{x}_k)^2,$$

where x_i and y_i are the gene expression level and classification for sample i, $\bar{x}$ is the mean of all samples, and $\bar{x}_k$ is the means samples in class k (Dudoit et al., 2002). In practice, a defined set of genes can be chosen on the basis of expert knowledge, relevance to pathophysiology, prior validation. The second subset contains 2000 genes randomly selected from the corresponding whole dataset.

Six datasets from the literature were chosen to evaluate the selected classification methods. The Leukemia data (AML/ALL) are from MIT (http://www-genome.wi.mit.edu/mpr/data_set_ALL_AML.html). These data, generated by using Affymetrix oligonucleotide microarray (Hu6800), categorized acute myeloid leukemia

(AML), and acute lymphoblastic leukemia (ALL) (Golub et al., 1999). There are 72 arrays (47 for ALL and 25 for AML).

In addition, we analyzed the multiclass NCI60 cancer cell-line data (http://genome-www.stanford.edu/nci60/). This dataset was generated from spotted microarray on cell lines commonly used for cancer drug screening. The data from 57 arrays in eight classes were used for our comparison because two prostate lines and 1 unknown line were excluded, because of small class size.

The TUMOR data were from a molecular profiling study of metastasis in primary solid tumors (http://www-genome.wi.mit.edu/cgi-bin/cancer/datasets.cgi) (Ramaswamy et al., 2003). These data from 76 samples were generated by using Affymetrix oligonucleotide microarray for the comparison of humor primary adenocarcinomas and adenocarcinoma metastases (64 vs. 12 arrays).

The DLFS and DLBC data were from a study on diffuse large B-Cell Lymphomas (http://www-genome.wi.mit.edu/cgi-bin/cancer/datasets.cgi) (Shipp et al., 2002). These data were generated by using Affymetrix oligonucleotide microarray for classifying Diffuse Large B-Cell Lymphoma (DLBC) vs. Follicular Lymphoma (FL) in DLFS dataset, and cured vs. fatal Diffuse Large B-Cell Lymphoma in DLBC dataset. DLFS dataset has 77 arrays with 58 DLBC and 19 FL; DLBC dataset has 58 arrays with 32 cured and 26 fatal.

We also used a multiclass ALIZ data (commonly known as *lymphoma data*), which were from a study of Diffuse Large B-Cell Lymphoma (http://geaw.nci.nih.gov/classification/lymphomaData) (Alizadeh et al., 2000). Of the 96 samples used in the article, we chose 88 samples that included six classes (each class with at least six samples). The data were generated from spotted microarray.

Data were neither filtered, nor censored, to preserve the inherent noise in the data. However, data were standardized to have a mean of zero and a variance of one for each gene. The standardization was necessary to give equal weight for each variable (gene) and avoid convergence problems in methods such as SVM and NN. For the NCI60 dataset, any gene that had more than one missing value was excluded. Genes that had only one missing value were imputed by using KNN with the correlation as the distance metric (Dudoit et al., 2002).

Implementation

The PLS was performed in SAS® (SAS Institute, Cary, NC); all other methods were implemented in R (http://www.r-project.org) (Ihaka and Gentleman, 1996).

RESULTS AND DISCUSSION

In general, all methods (NN, KNN, PCA–LDA, PCA–QDA, PLS–DA, RF, and SVM) performed well for AML/ALL data, as evidenced by low misclassification rates for both gene subsets and validation methods (Table 2 and Fig. 1). For the 50-gene dataset, the estimated error rate is from 2.78 to 4.17% with leave-one-out cross-validation (LOOCV) and 1.05–2.93% with bootstrap (BS). For the 2000 gene dataset, the error is a bit larger, ranging from 2.78% to 9.72% with LOOCV and from

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

Table 2. Comparison of different classification methods in analyzing six datasets. For each dataset, two subsets (50 genes and 2000 genes) were used (refer to Datasets section). The LOOCV (leave-one-out cross-validation) procedure was described in the Methods section. The 95% confidence intervals (CI) of error rates were obtained by using the bootstrap approach as described in the Methods section.

		LOOCV		Bootstrap		
Dataset	Method	Error rate (%)	Misclassified samples	Error rate (%)	Lower CI (%)	Upper CI (%)
AML/ALL						
AML/ALL-50	NN	4.17	17,54,66	1.08	0.00	4.17
	KNN	2.78	38,66	2.15	0.00	4.17
	PCA–LDA	2.78	17,66	2.81	0.79	5.56
	PCA–QDA	2.78	17,66	2.93	0.79	5.56
	PLS	2.78	17,66	2.78	0.00	4.17
	RF			1.68	1.39	2.78
	SVM	2.78	17,66	1.05	0.00	2.78
AML/ ALL-2000	KNN	9.72	28,40,52,54, 57,61,65	6.29	0.00	13.61
	PCA–LDA	8.33	35,54,57,60,61,66	3.67	0.00	8.84
	PCA–QDA	5.56	29,40,54,66	4.91	0.69	11.11
	RF			5.15	2.78	6.94
	PLS	4.17	17,61,66	1.39	0.00	6.25
	SVM	2.78	61,66	1.88	0.00	6.94
NCI60						
NCI60-50	NN	24.56	3,4,5,7–10,18,19, 27,32,33,44,47	7.61	1.75	14.04
	KNN	26.32	4,5,8–10,18,19, 22,27,28,44,46, 50–52	35.96	25.75	45.62
	PCA–LDA	22.81	4,5,8–10,18,23, 24,27,37,44, 50,51	9.37	3.55	15.79
	PLS	19.30	4,5,8–10,18,19,23, 24,27,51	12.28	8.77	17.54
	RF			31.64	26.32	38.60
	SVM	19.30	4,5,8-10,18,19,20, 27,44,51	7.59	1.75	14.04
NCI60-2000	KNN	31.58	4–10,18,19,21, 22,25,27,29,31, 36,44,46	51.87	39.19	63.95
	PCA–LDA	49.12	1–8,10,11,16, 18–22,24–27,31,36, 37,44–46,50,51	21.27	12.94	29.82
	PLS	28.07	3–5,7,9,10,18–22, 25,27,45,46,48	10.53	7.02	15.79
	RF			43.27	38.60	49.12
	SVM	33.33	3–10,18–22,25,27, 36,37,45,46	12.18	5.26	19.30

(*continued*)

Table 2. Continued.

Dataset	Method	LOOCV Error rate (%)	LOOCV Misclassified samples	Bootstrap Error rate (%)	Bootstrap Lower CI (%)	Bootstrap Upper CI (%)
TUMOR						
TUMOR-50	NN	13.16	65,66,67,69,70, 71,72,73,74,76	3.97	0.00	11.12
	KNN	7.89	65,66,70,71,72,76	10.12	4.99	14.98
	PCA–LDA	6.58	36,65,71,72,76	7.04	3.16	11.54
	PCA–QDA	5.26	65,71,72,76	7.06	2.52	13.16
	PLS	7.89	65,68,70-72,76	6.58	1.32	11.84
	RF			10.05	9.21	10.53
	SVM	5.26	65,71,72,76	6.56	2.63	10.53
TUMOR-2000	KNN	17.11	48,54,65,66,67, 68,69,70,71,72, 74,75,76	20.51	15.05	25.59
	PCA–LDA	17.11	48,65,66,67,68, 69,70,71,72,73, 74,75,76	14.76	10.66	19.08
	PCA–QDA	13.16	65,66,67,68,69, 70,72,74,75,76	13.16	9.63	19.85
	PLS	10.53	35,48,65,67, 69,71,72,76	6.58	2.63	11.84
	RF			16.54	14.47	17.11
	SVM	11.80	15,48,65,67,68, 69,71,72,76	4.96	1.32	9.21
ALIZ						
ALIZ-50	NN	32.95	1,2,3,4,6,7,8,9, 10,11,12,13,14,16, 17,18,20,23,25, 68,71,74,75,77, 81,83,85,87,88	6.20	1.52	11.36
	KNN	7.95	8,17,34,66,79,83,84	6.44	0.06	12.40
	PCA–LDA	15.91	6,8,62,66,74,75, 76,77,80,81,83, 84,87,88	5.72	1.65	10.61
	PLS	9.09	6,17,20,62,75, 76,83,87	9.09	5.11	13.64
	RF			14.70	12.50	17.05
	SVM	4.50	8,34,83,88	3.48	0.00	7.95
ALIZ-2000	KNN	15.91	3,4,6,8,9,17,19,21, 24,34,69,83,87,88	10.14	2.95	17.44
	PCA–LDA	19.32	1,3,4,5,6,16,17, 19,20,21,24,34, 73,74,83,87,88	7.37	3.12	12.50
	PLS	2.27	83,87	2.27	0.00	6.82
	RF			28.76	26.14	30.68
	SVM	4.50	8,34,83,88	3.69	0.00	9.09

(*continued*)

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

Table 2. Continued.

		LOOCV		Bootstrap		
Dataset	Method	Error rate (%)	Misclassified samples	Error rate (%)	Lower CI (%)	Upper CI (%)
DLFS						
DLFS-50	NN	18.18	55,61,62,63,64, 65,68,69,70,71, 72,73,75,77	2.04	0.00	6.49
	KNN	2.60	26,56	0.99	0.00	3.90
	PCA–LDA	3.90	26,46,68	2.99	0.24	6.49
	PCA–QDA	2.60	26,68	1.61	0.36	3.70
	PLS	5.19	26,46,67,68	5.19	2.59	7.79
	RF			8.69	6.49	10.39
	SVM	1.30	68	1.00	0.00	3.90
DLFS-2000	KNN	15.58	15,27,29,39,52,54, 55,61,62,67,68,76	12.04	5.56	17.64
	PCA–LDA	6.49	59,61,67,68,70	6.08	1.40	11.50
	PCA–QDA	9.09	29,59,60,67,68, 70,72	NA[a]		
	PLS	5.19	29,55,67,68	2.59	0.00	6.49
	RF			13.14	10.39	15.58
	SVM	5.19	29,55,67,68	1.91	0.00	5.19
DLBC						
DLBC-50	NN	10.34	13,16,23,43,48,56	4.44	0.00	10.35
	KNN	24.14	1,12,13,17,28,33, 35,39,40,47,51,54, 56,57	6.15	0.00	14.72
	PCA–LDA	13.79	1,13,16,18,23,45, 56,58	12.21	5.75	19.71
	PCA–QDA	17.24	13,16,23,35,43,45, 54,56,57,58	10.66	4.33	17.88
	PLS	15.52	1,13,16,18,23, 35,39,54,56	17.24	10.35	24.14
	RF			18.95	13.79	24.14
	SVM	8.60	13,23,33,47,56	4.17	0.00	10.34
DLBC-2000	KNN	46.55	1,4,8,12,14,17,20, 23,25,26,28,32,33, 35,37,38,39,40,43, 44,46,49,50,52,53, 54,56	46.88	33.08	57.47
	PCA–LDA	56.90	3,6,10,12,13,22,31, 33,34,35,36,37,38, 39,40,41,42,43,44, 45,46,47,48,49,50, 51,52,53,54,55,56, 57,58	42.53	35.26	51.15

(*continued*)

Table 2. Continued.

Dataset	Method	LOOCV		Bootstrap		
		Error rate (%)	Misclassified samples	Error rate (%)	Lower CI (%)	Upper CI (%)
	PCA–QDA	51.72	9,11,13,14,15,16, 17,19,21,22,27,29, 31,33,34,37,38,40, 41,42,43,44,46,47, 48,51,52,53,54,56	32.05	22.75	41.48
	PLS	NA[a]		NA[b]		
	RF			45.12	37.93	51.72
	SVM	39.66	1,12,13,14,17,19, 20,23,28,33,35,37, 39,43,45,46,47,49, 52,53,54,56,58	16.17	8.62	24.14

[a]Failed to obtain results due to imbalance between two classes.
[b]Cross-validation indicated that no PLS model was appropriate for this data.

1.39% to 6.29% with BS. The ease of classification with the AML/ALL datasets is, in part, due to the separation of leukemia types in the descriptor space. In fact, one gene (Zyxin) provides near perfect classification (Li and Yang, 2002). Regardless of subset, methods perform nearly the same within subset, except for KNN with 2000 genes. Noting this exception, the choice of classification technique is less important for data in which classes are well separated in descriptor space.

However, the error rates were higher and more variable for the NCI60 data sets (Table 2 and Fig. 1). This difference is partly due to multiple classes and few replicates per sample. For these data, there appears to be a performance difference among methods within each gene subset. Specifically, KNN and RF perform significantly worse than the other methods. At the other extreme, the best performers are SVM and PLS. Hence, the choice of classification method appears to be important when analyzing data that are more complicated.

We observed similar results in four additional microarray datasets (Table 2 and Fig. 1). The SVM remained to be the top performer regardless of using the 50-gene or the 2000-gene subset for each of the four datasets. In a similar comparative study of different classification methods on proteomic data, SVM was also one of the best methods using markers having the highest *t*-statistic (Wu et al., 2003). The PLS, while achieving respectable results on 50-gene subsets, performed extremely well on 2000-gene subsets, which reflects PLS's resilience to noise in data.

It is also interesting to note that SVM and PLS handled the multiclass ALIZ dataset very well, in particular on the 2000-gene subset, which is more realistic because the inherited noise is preserved. On the other hand, RF, which gave reasonable accuracies for two-class datasets, committed much higher errors for the multiclass ALIZ datasets.

Although SVM was the top performer for the DLBC 2000-gene subset, all methods gave high error rates. This might suggest either very weak class difference

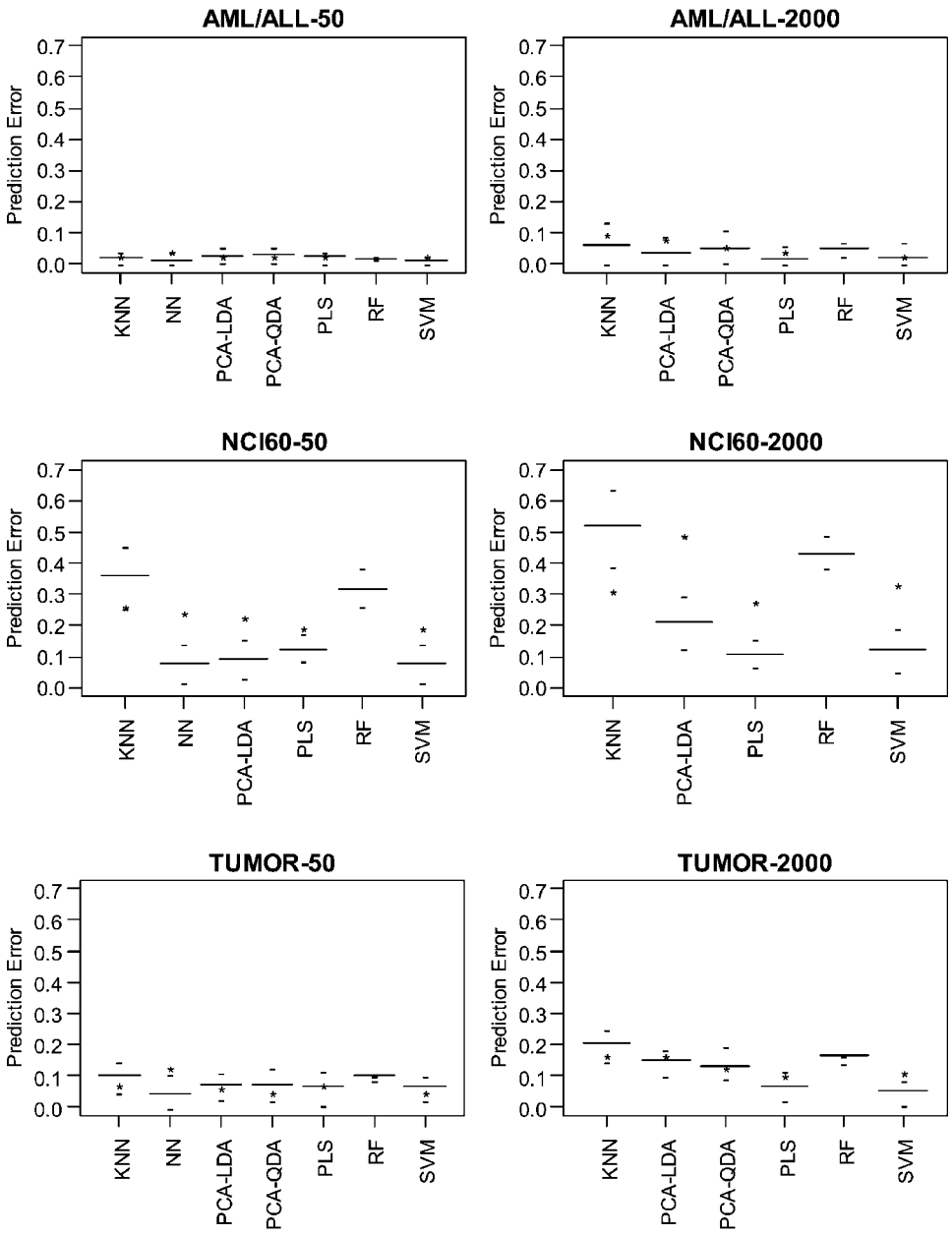

Figure 1. Graphical comparison of classification methods. For each method and each dataset, the prediction error from bootstrap CV is reported as mean (——) with 95% confidence interval (-) and that from LOOCV is labeled as $*$ on the plot.

(*continued*)

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

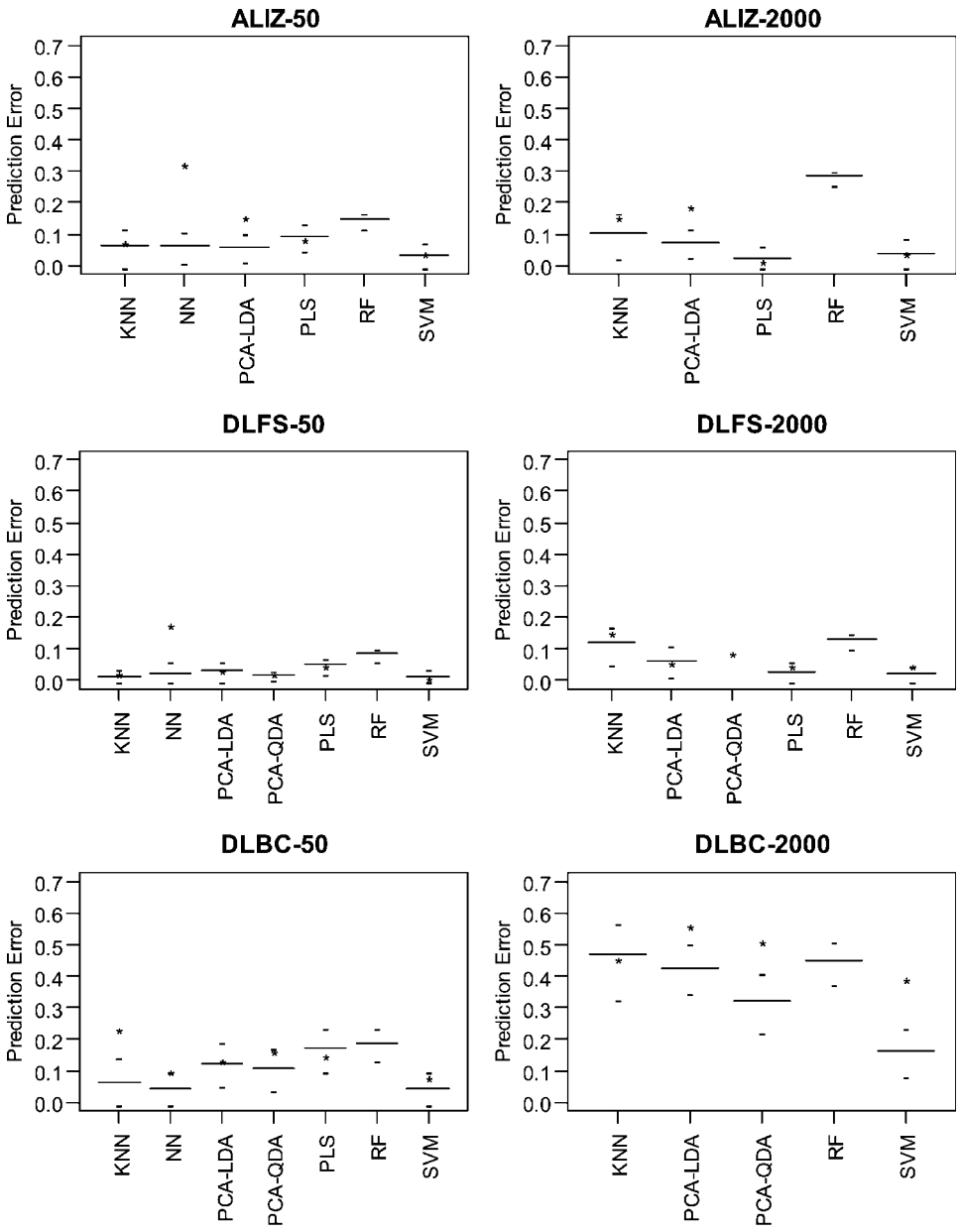

Figure 1. Continued.

or extremely high noise in the data. Indeed, DLBC data were noted to have little correlation with survival (Kaplan–Meier log-rank p-value is 0.497) and high misclassification rate (26/58) (Shipp et al., 2002).

Among the statistical approaches that we used, PLS and PCA-DA are parametric methods and pay more attention on assumptions about the underlying

structure of the data. When using PCA-DA, the dimension reduction occurs prior to finding an optimal separating hyper-plane. The independence of the dimension reduction and discrimination can often lead to a less-than-desirable model. Alternatively, PLS simultaneously reduces dimension and discriminates, thus producing a more optimal model.

The KNN is a nonparametric statistical method and is a special case of kernel density estimator (Scott, 1992). Hence, KNN requires no underlining distribution assumption and is used frequently as an exploratory tool. The KNN is simple to implement and was noted for excellent performance in classifying image datasets in the StatLog project (Michie et al., 1994). However, it does not explain the underlining structure in the data.

Among the learning algorithms, the field of NN is very diverse. Neural networks, in one of its simplest forms, is equivalent to logistical regression in statistics. Details of statistical model fitting in NN are hidden from the user. The downside of NN is the lack of diagnostics, which are important for assumption checking and model fine-tuning. Neural networks was noted for the best or near the best performance in StatLog project for most of the datasets (Michie et al., 1994). However, NN was not noted for speed, due to slow convergence (Michie et al., 1994). Multiple starting points and fine-tuning might be required to build a ''good'' NN architecture (Ripley, 1996). Another drawback is the difficulty of interpreting the model generated by NN.

Support vector machines, a recent phenomenon in learning algorithm circles, benefited from development in other learning algorithms (such as NN). Yet, SVM has strong theoretical foundations in mathematics and statistics (Vapnik, 1998). More importantly, SVM, formulated as a dual problem, is easy to solve and flexible to use for a wide range of data, given the choice of many different kernels. Initially developed for binary classification, SVM has evolved to accommodate multiclass problems (Hsu and Lin, 2002). The SVM presents great promise for genomic applications, such as microarray data analysis (Ben Dor et al., 2000; Brown et al., 2000; Chow et al., 2001; Furey et al., 2000; Lee and Lee, 2002; Moler et al., 2000; Ramaswamy et al., 2001; Yeang et al., 2001).

Potential pitfalls exist when one uses the result from a subset generated in a variable selection study to infer the generalized error rate of the whole dataset, due to the selection bias. Although remedies (using external selection coupled with cross-validation or bootstrap) exist if the assessment of the selection bias is desired, they are not foolproof (Ambroise and McLachlan, 2002). Therefore, bias could still be introduced, when making direct comparison of any two subsets, regardless of what selection criteria were used (either scientific or statistical), how many genes were chosen, and what remedy was used.

Given this concern, even though it is tempting to compare the performance of each method in the 50-gene subset with that in the 2000-gene subset for each dataset, the comparison would not be appropriate. In addition, each subset has a different purpose. Our intention is to compare the different methods on the same dataset instead of on different datasets (the 50-gene subset and 2000-gene subset are different datasets). The 50-gene subsets can be thought of as a complete mock dataset to reflect one scenario of a biomarker study (e.g., measure small number of genes using QPCR or custom microarray due to limited sample or resource). The subsequent analyses

were performed without selecting a ''subset'' from this ''complete dataset.'' Although we would prefer to use the whole dataset to reflect another scenario in biomarker studies where we could measure all genes, we had to settle with 2000 genes to ensure most methods would be computationally feasible given our computation resources.

Although we did not perform a direct computational comparison among methods, we can make general conclusions about computational concerns. Because our bootstrap iteration size is 2000, all of the methods shown here are computationally intensive. This issue could become problematic when a larger (more genes) data set is used. Running BS as a parallel process may reduce the computation time. Unfortunately, certain methods, including NN, cannot handle large datasets when the bootstrap methodology is used.

In biomarker applications using expression data, there may be some added benefit by taking a combination approach. Using top performers (PLS and SVM) in combination with other methods (such as KNN, NN, RF, etc.) may increase the confidence in prediction and/or generate more interpretable model. For future research, we are interested in studying the combination approach in more detail. In addition, using a variety of classification techniques in combination with cross-validation provides a way to identify consistently misclassified samples, which initially may have been mislabeled. For example, in the AML/ALL dataset, observation 66 was consistently misclassified (Table 2). This was one of the observations that Golub initially marked as uncertain (Golub et al., 1999). In the NCI60 dataset, some observations (4, 5, 8, 9, 10, 19, 20, 27, 44, 50, 51) could also be mislabeled. Ross et al. (2000) previously suggested that two breast cancer cell lines (50, 51) might have been mislabeled and were in fact melanoma cell lines. The use of multiple methods here coincides with that opinion. However, it is unlikely that all of those observations are mislabeled; rather, it is more plausible that some of those observations are difficult to classify.

ACKNOWLEDGMENTS

We thank Ken Hu, Michael Gieseg, Christy Chuang-Stein, Tom Vidmar, and David Potter for providing valuable feedback, and David Pyne for support and encouragement.

REFERENCES

Aeberhard, S., De Vel, O. Y., Coomans, D. H. (2002). New fast algorithms for error rate-based stepwise variable selection in discriminant analysis. *SIAM J. Sci. Comput.* 22:1036–1052.

Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, X., Powell, J. I., Yang, L., Marti, G. E., Moore, T., Hudson, J. Jr., Lu, L., Lewis, D. B., Tibshirani, R., Sherlock, G., Chan, W. C., Greiner, T. C., Weisenburger, D. D., Armitage, J. O., Warnke, R., Staudt, L. M. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 403:503–511.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., Levine, A. J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci. USA* 96:6745–6750.

Alsberg, B. K., Douglas, B. K., Goodacre, R. (1998). Variable selection in discriminant partial least-squares analysis. *Anal. Chem.* 70:4126–4133.

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

Ben Dor, A., Bruhn, L., Friedman, N., Nachman, I., Schummer, M., Yakhini, Z. (2000). Tissue classification with gene expression profiles. *J. Comput. Biol.* 7:559–583.

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

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

Bumol, T. F., Watanabe, A. M. (2001). Genetic information, genomic technologies, and the future of drug discovery. *JAMA* 285:551–555.

Cantor, C. R. (1999). Pharmacogenetics becomes pharmacogenomics: Wake up and get ready. *Mol. Diagn.* 4:287–288.

Chow, M. L., Moler, E. J., Mian, I. S. (2001). Identifying marker genes in transcription profiling data using a mixture of feature relevance experts. *Physiol Genom.* 5:99–111.

Cortijo, F. J., Perez de la Blanca, N., Molina, R., Abad, J. (1993). *On the Combination of Nonparametric Nearest Neighbor Classification and Contextual Correction*. Universidad de Granada Technical Report.

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

Dudoit, S., Fridlyand, J., Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Am. Statist. Assoc.* 97:77–87.

Efron, B., Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Boca Raton, FL: Chapman & Hall/CRC.

Fix, E., Hodges, J. L. (1951). *Discriminatory Analysis. Nonparametric Estimation: Consistency Properties*. Technical Report No. 4, Project No. 21-49-004, Randolph Field, TX: USAF School of Aviation Medicine.

Frank, I., Friedman, J. (1993). Statistical view of chemometric regression tools. *Technometrics* 35:109–135.

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., Gaasenbeek, 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.

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

Gruvberger, S., Ringner, M., Chen, Y., Panavally, S., Saal, L. H., Borg, A., Ferno, M., Peterson, C., Meltzer, P. S. (2001). Estrogen receptor status in breast cancer is associated with remarkably distinct gene expression patterns. *Cancer Res.* 61:5979–5984.

Gunther, E. C., Stone, D. J., Gerwien, R. W., Bento, P., Heyes, M. P. (2003). Prediction of clinical drug efficacy by classification of drug-induced genomic expression profiles in vitro. *Proc. Natl. Acad. Sci. USA* 100:9608–9613.

Hastie, T., Tibshirani, R., Friedman, J. H. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer.

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

Ihaka, R., Gentleman, R. (1996). R: A language for data analysis and graphics. *J. Comput. Graph. Statist.* 5:299–314.

Lee, Y., Lee, C. K. (2002). *Classification of Multiple Cancer Types by Multicategory Support Vector Machines Using Gene Expression Data*. Technical Report 1051r ftp://ftp. stat. wisc. edu/pub/wahba/tr1051rr. ps.

Li, W., Yang, Y. (2002). How many genes are needed for a discriminant microarray data analysis. In: Lin, S. M., Johnson, K. F., eds. *Methods of Microarray Data Analysis*. Boston: Kluwer Academic, pp. 137–150.

Michie, D., Spiegelhalter, D. J., Taylor, C. C. (1994). *Machine Learning, Neural and Statistical Classification.* Michie, D., Spiegelhalter, D. J., Taylor, C. C., eds. New York: Ellis Horwood.

Moler, E. J., Chow, M. L., Mian, I. S. (2000). Analysis of molecular profile data using generative and discriminative methods. *Physiol. Genom.* 4:109–126.

Moreira, M., Mayoraz, E. (1998). Improved pairwise coupling classification with correcting classifiers. In: *Proceedings of the Tenth European Conference on Machine Learning*. London, Springer-Verlag, pp. 160–171.

Neter, J., Kutner, M. H., Nachtsheim, C. J., Wasserman, W. (1996). *Applied Linear Statistical Models*. Boston: WCB/McGraw-Hill.

Nguyen, D. V., Rocke, D. M. (2002). Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics* 18:39–50.

Nishizuka, S., Chen, S. T., Gwadry, F. G., Alexander, J., Major, S. M., Scherf, U., Reinhold, W. C., Waltham, M., Charboneau, L., Young, L., Bussey, K. J., Kim, S., Lababidi, S., Lee, J. K., Pittaluga, S., Scudiero, D. A., Sausville, E. A., Munson, P. J., Petricoin, E. F., Liotta, L. A., Hewitt, S. M., Raffeld, M., Weinstein, J. N. (2003). Diagnostic markers that distinguish colon and ovarian adenocarcinomas: Identification by genomic, proteomic, and tissue array profiling. *Cancer Res.* 63:5243–5250.

Patterson, D. (1996). *Artificial Neural Networks*. Singapore: Prentice Hall.

Petricoin, E. F. III, Hackett, J. L., Lesko, L. J., Puri, R. K., Gutman, S. I., Chumakov, K., Woodcock, J., Feigal, D. W. Jr., Zoon, K. C., Sistare, F. D. (2002). Medical applications of microarray technologies: A regulatory science perspective. *Nat. Genet.* 32(Suppl):474–479.

Price, D., Knerr, S., Personnaz, L., Dreyfus, G. (1994). Pairwise neural network classifiers with probabilistic outputs. *Neural Inf. Process. Sys.* 7.

Ramaswamy, S., Tamayo, P., Rifkin, R., Mukherjee, S., Yeang, C. H., Angelo, M., Ladd, C., Reich, M., Latulippe, E., Mesirov, J. P., Poggio, T., Gerald, W.,

Loda, M., Lander, E. S., Golub, T. R. (2001). Multiclass cancer diagnosis using tumor gene expression signatures. *Proc. Natl. Acad. Sci. USA* 98:15149–15154.

Ramaswamy, S., Ross, K. N., Lander, E. S., Golub, T. R. (2003). A molecular signature of metastasis in primary solid tumors. *Nat. Genet.* 33:49–54.

Rayens, W. S. (2000). *The Art of Maximizing Covariance*. University of Kentucky Technical Report.

Ricci, M. S., El Deiry, W. S. (2000). Novel strategies for therapeutic design in molecular oncology using gene expression profiles. *Curr. Opin. Mol. Ther.* 2:682–690.

Ripley, B. D. (1996). *Pattern Recognition and Neural Networks*. Cambridge: Cambridge University Press.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychol. Rev.* 65:386–408.

Ross, D. T., Scherf, U., Eisen, M. B., Perou, C. M., Rees, C., Spellman, P., Iyer, V., Jeffrey, S. S., van de, R. M., Waltham, M., Pergamenschikov, A., Lee, J. C., Lashkari, D., Shalon, D., Myers, T. G., Weinstein, J. N., Botstein, D., Brown, P. O. (2000). Systematic variation in gene expression patterns in human cancer cell lines. *Nat. Genet.* 24:227–235.

Saaksjarvi, E., Khalighi, M., Minkkinen, P. (1989). Waste water pollution modeling in the southern area of lake Saimaa, Finland, by the SIMCA pattern recognition model. *Chemomet Intell Lab Sys.* 7:171–180.

Scott, D. W. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization.* New York: John Wiley.

Selaru, F. M., Xu, Y., Yin, J., Zou, T., Liu, T. C., Mori, Y., Abraham, J. M., Sato, F., Wang, S., Twigg, C., Olaru, A., Shustova, V., Leytin, A., Hytiroglou, P., Shibata, D., Harpaz, N., Meltzer, S. J. (2002). Artificial neural networks distinguish among subtypes of neoplastic colorectal lesions. *Gastroenterology* 122:606–613.

Shipp, M. A., Ross, K. N., Tamayo, P., Weng, A. P., Kutok, J. L., Aguiar, R. C., Gaasenbeek, M., Angelo, M., Reich, M., Pinkus, G. S., Ray, T. S., Koval, M. A., Last, K. W., Norton, A., Lister, T. A., Mesirov, J., Neuberg, D. S., Lander, E. S., Aster, J. C., Golub, T. R. (2002). Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nat. Med.* 8:68–74.

Stone, M., Brooks, R. J. (1990). Continuum regression: Cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares, and principal component regression. *J. R. Statist. Soc. Ser. B.* 52:237–269.

Stratowa, C., Loffler, G., Lichter, P., Stilgenbauer, S., Haberl, P., Schweifer, N., Dohner, H., Wilgenbus, K. K. (2001). CDNA microarray gene expression analysis of B-cell chronic lymphocytic leukemia proposes potential new prognostic markers involved in lymphocyte trafficking. *Int. J. Cancer* 91: 474–480.

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

Wilkinson, D. G. (1994). *Situ Hybridization. A practical Approach*. Oxford: Oxford University Press.

Wold, S. (1995). PLS for multivariate linear modeling. In: Han van de Waterbeemd, ed. *Chemometric Methods in Molecular Design*. Weinheim: VCH Verlagsgesellschaft mbH, pp. 195–218.

Wu, B., Abbott, T., Fishman, D., McMurray, W., Mor, G., Stone, K., Ward, D., Williams, K., Zhao, H. (2003). Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. *Bioinformatics* 19:1636–1643.

Xiong, M., Jin, L., Li, W., Boerwinkle, E. (2000). Computational methods for gene expression-based tumor classification. *Biotechniques* 29:1264–1268, 1270.

Yeang, C. H., Ramaswamy, S., Tamayo, P., Mukherjee, S., Rifkin, R. M., Angelo, M., Reich, M., Lander, E., Mesirov, J., Golub, T. (2001). Molecular classification of multiple tumor types. *Bioinformatics* 17(Suppl 1):S316–S322.

Received March 2004
Accepted May 2004

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016