Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data

Sandrine Dudoit*
Mathematical Sciences Research Institute, Berkeley, CA.

Jane Fridlyand*
Department of Statistics, UC Berkeley

Terence P. Speed
Department of Statistics, UC Berkeley

Technical report # 576, June 2000

Address for correspondence:
Sandrine Dudoit
Department of Statistics
University of California, Berkeley
Berkeley, CA 94720-3860
sandrine@stat.berkeley.edu

* *These authors contributed equally to this work.*

Abstract

A reliable and precise classification of tumors is essential for successful treatment of cancer. cDNA microarrays and high-density oligonucleotide chips are novel biotechnologies which are being used increasingly in cancer research. By allowing the monitoring of expression levels for thousands of genes simultaneously, such techniques may lead to a more complete understanding of the molecular variations among tumors and hence to a finer and more informative classification. The ability to successfully distinguish between tumor classes (already known or yet to be discovered) using gene expression data is an important aspect of this novel approach to cancer classification.

In this paper, we compare the performance of different discrimination methods for the classification of tumors based on gene expression data. These methods include: nearest neighbor classifiers, linear discriminant analysis, and classification trees. In our comparison, we also consider recent machine learning approaches such as bagging and boosting. We investigate the use of prediction votes to assess the confidence of each prediction. The methods are applied to datasets from three recently published cancer gene expression studies.

Keywords: Discriminant analysis; machine learning; variable selection; microarray experiment; cancer; tumor class.

1 Introduction

A reliable and precise classification of tumors is essential for successful treatment of cancer. Current methods for classifying human malignancies rely on a variety of morphological, clinical, and molecular variables. In spite of recent progress, there are still uncertainties in diagnosis. Furthermore, it is likely that the existing classes are heterogeneous and comprise diseases which are molecularly distinct and follow different clinical courses. cDNA microarrays and high-density oligonucleotide chips are novel biotechnologies which are being used increasingly in cancer research [2, 3, 17, 23, 24, 26]. By allowing the monitoring of expression levels for thousands of genes simultaneously [1, 12, 13, 19, 28, 29], such techniques may lead to a more complete understanding of the molecular variations among tumors and hence to a finer and more reliable classification. In the words of Alizadeh *et al.* [2]: "Indeed, the new methods of gene expression profiling call for a revised definition of what is deemed a 'disease'."

The wealth of gene expression data now available poses numerous statistical questions ranging from the analysis of images produced by microarray experiments and the study of the variability of measured gene expression levels [11, 22], to the elucidation of biochemical pathways. Here we focus on the classification of tumors using gene expression data. There are three main types of statistical problems associated with tumor classification: (i) the identification of new/unknown tumor classes using gene expression profiles - *cluster analysis/unsupervised learning*, (ii) the classification of malignancies into known classes - *discriminant analysis/supervised learning*, and (iii) the identification of "marker" genes that characterize the different tumor classes - *variable selection*. An unusual feature of this new type of data is the very large number of variables (genes) relative to the number of observations (tumor samples); the publicly available datasets currently contain gene expression data for 5,000-10,000 genes on less than 100 observations. Both numbers are expected to grow, the number of genes reaching around 100,000, an estimate for the total number of genes in the human genome.

Recent publications on cancer classification using gene expression data have mainly focused on the cluster analysis of both tumor samples and genes, and include applications of hierarchical clustering methods [2, 3, 23, 24, 26, 30] and partitioning methods such as self-organizing maps [17]. Using acute leukemias as a test case, Golub *et al.* [17] looked into both the cluster analysis and the discriminant analysis of tumors using gene expression data. For cluster analysis, or "class discovery", self-organizing maps (SOM) were applied to the gene expression data and the tumor groups revealed by this method were compared to known classes. For discriminant analysis, or "class prediction", Golub *et al.* proposed a weighted gene voting scheme which turns out to be a minor variant of a special case of linear discriminant analysis for multivariate normal class densities. Alizadeh *et al.* [2] studied gene expression in the three most prevalent adult lymphoid malignancies. Two previously unrecognized types of diffuse large B-cell lymphoma, with distinct clinical behaviors, were identified based on gene expression data. Average linkage hierarchical clustering was used to identify the two tumor subclasses as well as to group genes with similar expression patterns across the different samples. Ross *et al.* [26] used cDNA microarrays to study gene expression in the 60 cell lines from the National Cancer Institute's anti-cancer drug screen (NCI 60). Hierarchical clustering of the cell lines based on gene expression data revealed a correspondence between gene expression and tissue of origin of the tumors. Hierarchical clustering was also used to group genes with similar expression patterns across the cell lines.

The three recent studies just cited are instances of a growing body of research, in which gene expression profiling is used to distinguish between known tumor classes and to identify previously unrecognized and clinically significant subclasses of tumors. Indeed, Golub *et al.* [17] conclude that "This experience underscores the fact that leukemia diagnosis remains imperfect and could benefit from a battery of expression-based predictors for various cancers. Most importantly, the technique of class prediction can be applied to distinctions relating to future clinical outcomes, such as drug response or survival". In the same vein, Alizadeh *et al.* [2] "... anticipate that global surveys of gene expression in cancer, such as we present here, will identify a small number of marker genes that will be used to stratify patients into molecularly relevant categories which will improve the precision and power of clinical trials". The ability to successfully distinguish between tumor classes (already known or yet to be discovered) using gene expression data is thus an important aspect of this novel approach to cancer classification. So far, most published papers on tumor classification have applied a single technique to a single gene expression dataset. It is hard to assess the merits of each technique in the absence of a comprehensive comparative study.

In this paper, we compare the performance of different discrimination methods for the classification of tumors based on gene expression profiles. These methods include traditional ones such as nearest neighbors and linear discriminant analysis, as well as more modern ones such as classification trees. In our comparison, we also consider recent machine learning approaches such as bagging and boosting. We investigate the use of prediction votes to assess the confidence of each prediction. The methods are applied to three recently published datasets: the leukemia (ALL/AML) dataset of Golub *et al.* [17], the lymphoma dataset of Alizadeh *et al.* [2], and the 60 cancer cell line (NCI 60) dataset of Ross *et al.* [26].

The paper is organized as follows. Section 2 contains a brief introduction to the biology and technology of cDNA microarrays. Section 3 discusses the discrimination methods considered in the paper. The datasets are described in Section 4, along with preliminary data processing steps. The study design for the comparison of the discrimination methods is discussed in Section 5 and the results of the study are presented in Section 6. Finally, Section 7 summarizes our findings and outlines open questions.

2 Background on cDNA microarrays

The ever increasing rate at which genomes are being sequenced has opened a new area of genome research, functional genomics, which is concerned with assigning biological function to DNA sequences. With the complete DNA sequences of many genomes already known (*e.g.* the yeast *S. cerevisae*, the round worm *C. elegans*, the fruit fly *D. melanogaster*, and many bacteria) and the human genome well on its way to being fully sequenced, an essential and formidable task is to define the role of each gene and understand how the genome functions as a whole. Innovative approaches, such as the cDNA and oligonucleotide microarray technologies, have been developed to exploit DNA sequence data and yield information about gene expression levels for entire genomes. Next, we briefly review basic genetic notions useful for an understanding of microarray experiments.

A *gene* consists of a segment of DNA which codes for a particular *protein*, the ultimate expression of the genetic information. A *deoxyribonucleic acid* or *DNA* molecule is a double-stranded polymer composed of four basic molecular units called nucleotides. Each *nucleotide* comprises a phosphate group, a deoxyribose sugar, and one of *four nitrogen bases*. The four different bases found in DNA are adenine (A), guanine (G), cytosine (C), and thymine (T). The two chains are held together by hydrogen bonds between nitrogen bases, with base-pairing occurring according to the following rule: G pairs with C, and A pairs with T. While a DNA molecule is built from a four-letter alphabet, proteins are sequences of twenty different types of *amino acids*. The expression of the genetic information stored in the DNA molecule occurs in two stages: (i) *transcription*, during which DNA is transcribed into *messenger ribonucleic acid* or *mRNA*, a single-stranded complementary copy of the base sequence in the DNA molecule, with the base uracil (U) replacing thymine; (ii) *translation*, during which mRNA is translated to produce a protein. The correspondence between DNA's four-letter alphabet and a protein's twenty-letter alphabet is specified by the *genetic code*, which relates nucleotide triplets to amino acids. See Griffiths *et al.* [18] for an introduction to the relevant biology.

Different properties of gene expression can be studied using microarrays, such as expression at the transcription or translation level, and subcellular localization of gene products. Microarrays derive their power and universality from a key property of DNA molecules described above: *complementary base-pairing*. The term *hybridization* refers to the annealing of nucleic acid strands from different sources according to the base-pairing rules. To date, attention has focussed primarily on expression at the transcription stage, *i.e.*, on mRNA levels. Although the regulation of protein synthesis in a cell is by no means regulated solely by mRNA levels, mRNA levels sensitively reflect the type and state of the cell. To utilize the hybridization property of DNA, *complementary* DNA or *cDNA* is obtained from mRNA by reverse transcription. There are different types of microarray systems, including cDNA microarrays [29, 12, 13] and high-density oligonucleotide arrays [19]; the description below focuses on the former.

cDNA microarrays consist of thousands of individual DNA sequences printed in a high density array on a glass microscope slide. The relative abundance of these DNA sequences in two DNA or cDNA samples may be assessed by monitoring the differential hybridization of the two samples to the sequences on the array. To this end, the two DNA samples or *targets* are labeled using different fluorescent dyes (*e.g.* a red-fluorescent dye Cy5 and a green-fluorescent dye Cy3), then mixed and hybridized with the arrayed DNA sequences or *probes* (following the definition of probe and target adopted in the January 1999 supplement to Nature Genetics [1]). After this competitive hybridization, fluorescence measurements are made separately for each dye at each spot on the array. The ratio of the fluor intensity for each spot is indicative of the relative abundance of the corresponding DNA sequence in the two samples (see `http://rana.Stanford.EDU/software/` for more information on the measurement of fluorescence intensities).

Aside from the enormous scientific potential of DNA microarrays to help in understanding gene regulation and interactions, microarrays have very important applications in pharmaceutical and clinical research. By comparing gene expression in normal and disease cells, microarrays may be used to identify disease genes and targets for therapeutic drugs. The supplement to Nature Genetics [1] and the book *DNA Microarrays : A Practical Approach* [28] provide general overviews of microarray technologies and of different areas of application of microarrays.

Microarrays are being applied increasingly in cancer research to study the molecular variations among tumors [2, 3, 17, 23, 24, 26]. This should lead to an improved classification of tumors, which in turn should result in progresses in the prevention and treatment of cancer. An important aspect of this endeavor is the ability to predict tumor types on the basis of gene expression data. We review below a number of prediction methods and assess their performance on the three cancer datasets described in Section 4.

3 Discrimination methods

For our purpose, gene expression data on p genes for n mRNA samples may be summarized by an $n \times p$ matrix $X = (x_{ij})$, where x_{ij} denotes the expression level of gene (variable) j in mRNA sample (observation) i. The expression levels might be either absolute (*e.g.* oligonucleotide arrays used to produce the leukemia dataset) or relative with respect to the expression levels of a suitably defined common reference sample (*e.g.* cDNA microarrays used to produce the lymphoma and NCI 60 datasets). When the mRNA samples belong to known classes (*e.g.* follicular lymphoma), the data for each observation consist of a gene expression profile $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$ and a class label y_i, *i.e.*, of predictor variables $\mathbf{x}_i$ and response y_i. For K classes, the class labels y_i are defined to be integers ranging from 1 to K. We let n_k denote the number of observations belonging to class k.

A *predictor* or *classifier* for K tumor classes partitions the space $\mathcal{X}$ of gene expression profiles into K disjoint subsets, $A_1, \ldots, A_K$, such that for a sample with expression profile $\mathbf{x} = (x_1, \ldots, x_p) \in A_k$ the predicted class is k.

Predictors are built from past experience, *i.e.*, from observations which are known to belong to certain classes. Such observations comprise the *learning set (LS)*
$\mathcal{L} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_{n_L}, y_{n_L})\}$. Predictors may then be applied to a *test set (TS)*
$\mathcal{T} = \{\mathbf{x}_1, \ldots, \mathbf{x}_{n_T}\}$, to predict for each observation $\mathbf{x}_i$ in the test set its class y_i. In the event that the y_i are known, the predicted and true classes may be compared to estimate the error

rate of the predictor.

We denote a classifier built from a learning set $\mathcal{L}$ by $C(\cdot, \mathcal{L})$; the predicted class for an observation $\mathbf{x}$ is $C(\mathbf{x}, \mathcal{L})$. Below, we review briefly a number of well-known discrimination methods. General references on the topic of discriminant analysis include [20, 21, 25].

3.1 Fisher linear discriminant analysis

First applied in 1935 by M. Barnard [4] at the suggestion of R. A. Fisher [14], *Fisher linear discriminant analysis (FLDA)* is based on finding linear combinations $\mathbf{xa}$ of the gene expression levels $\mathbf{x} = (x_1, \ldots, x_p)$ with large ratios of between-groups to within-groups sum of squares [20, 21, 25]. This criterion is intuitively appealing, because it is easier to tell the classes apart using variables (functions of variables) for which the between-groups sum of squares is large relative to the within-groups sum of squares.

For an $n \times p$ learning set data matrix X, the linear combination Xa of the columns of X has ratio of between-groups to within-groups sum of squares given by $\mathbf{a}'B\mathbf{a}/\mathbf{a}'W\mathbf{a}$, where B and W denote respectively the $p \times p$ matrices of between-groups and within-groups sum of squares. The extreme values of $\mathbf{a}'B\mathbf{a}/\mathbf{a}'W\mathbf{a}$ are obtained from the eigenvalues and eigenvectors of $W^{-1}B$. The matrix $W^{-1}B$ has at most $s = \min(K-1, p)$ non-zero eigenvalues, $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_s$, with corresponding linearly independent eigenvectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_s$. The *discriminant variables* are defined to be $u_l = \mathbf{x}\mathbf{v}_l$, $l = 1, \ldots, s$, and in particular, $\mathbf{v}_1$ maximizes $\mathbf{a}'B\mathbf{a}/\mathbf{a}'W\mathbf{a}$.

For an observation $\mathbf{x} = (x_1, \ldots, x_p)$, let

$$d_k(\mathbf{x}) = \sum_{l=1}^{s} ((\mathbf{x} - \bar{\mathbf{x}}_k)\mathbf{v}_l)^2$$

denote its (squared) Euclidean distance, in terms of the discriminant variables, from the $1 \times p$ vector of class k averages $\bar{\mathbf{x}}_k$ for the learning set $\mathcal{L}$. The predicted class for observation $\mathbf{x}$ is

$$\mathcal{C}(\mathbf{x}, \mathcal{L}) = \text{argmin}_k\, d_k(\mathbf{x}),$$

that is, the class whose mean vector is closest to $\mathbf{x}$ in the space of discriminant variables.

FLDA is a non-parametric method which also arises in a parametric setting. For $K = 2$ classes, FLDA yields the same classifier as the maximum likelihood discriminant rule for multivariate normal class densities with the same covariance matrix [20] (see below case 1 for $K = 2$).

3.2 Maximum likelihood discriminant rules

In a situation where the class conditional densities $pr(\mathbf{x}|y = k)$ are known, the *maximum likelihood (ML) discriminant rule* predicts the class of an observation $\mathbf{x} = (x_1, \ldots, x_p)$ by that which gives the largest likelihood to $\mathbf{x}$, *i.e.*, by $\mathcal{C}(\mathbf{x}) = \text{argmax}_k\, pr(\mathbf{x}|y = k)$. (When the class conditional densities are fully known, a learning set is not needed and the classifier is simply $\mathcal{C}(\mathbf{x})$.) In practice, however, even if the forms of the class conditional densities are known, their parameters must be estimated from a learning set. Using parameter estimates in place of the unknown parameters yields the *sample maximum likelihood discriminant rule.*

For multivariate normal class densities, *i.e.*, for $\mathbf{x}|y = k \sim N(\mu_k, \Sigma_k)$, the ML discriminant rule is

$$\mathcal{C}(\mathbf{x}) = \text{argmin}_k \left\{ (\mathbf{x} - \mu_k)\Sigma_k^{-1}(\mathbf{x} - \mu_k)' + \log|\Sigma_k| \right\}.$$

In general, this is a quadratic discriminant rule. Interesting special cases include:

1. When the class densities have the same covariance matrix, $\Sigma_k = \Sigma$, the discriminant rule is based on the square of the Mahalanobis distance and is linear

$$\mathcal{C}(\mathbf{x}) = \text{argmin}_k \ (\mathbf{x} - \mu_k)\Sigma^{-1}(\mathbf{x} - \mu_k)'.$$

2. When the class densities have diagonal covariance matrices, $\Delta_k = \text{diag}(\sigma_{k1}^2, \ldots, \sigma_{kp}^2)$, the discriminant rule is given by additive quadratic contributions from each variable

$$\mathcal{C}(\mathbf{x}) = \text{argmin}_k \sum_{j=1}^{p} \left\{ \frac{(x_j - \mu_{kj})^2}{\sigma_{kj}^2} + \log \sigma_{kj}^2 \right\}.$$

3. In this simplest case, when the class densities have the same diagonal covariance matrix $\Delta = \text{diag}(\sigma_1^2, \ldots, \sigma_p^2)$, the discriminant rule is linear and given by

$$\mathcal{C}(\mathbf{x}) = \text{argmin}_k \sum_{j=1}^{p} \frac{(x_j - \mu_{kj})^2}{\sigma_j^2}.$$

We refer to special cases 2 and 3 as diagonal quadratic (DQDA) and linear (DLDA) discriminant analysis, respectively. For the corresponding sample ML discriminant rules, the population mean vectors and covariance matrices are estimated from a learning set $\mathcal{L}$, by the sample mean vectors and covariance matrices, respectively: $\hat{\mu}_k = \bar{\mathbf{x}}_k$ and $\hat{\Sigma}_k = S_k$. For the constant covariance matrix case, the pooled estimate of the common covariance matrix is used: $\hat{\Sigma} = \sum_k (n_k - 1)S_k/(n - K)$.

Similar discriminant rules as above arise in a Bayesian context, where the predicted class is chosen to maximize the posterior class probabilities $pr(y = k|\mathbf{x})$ [20].

In one of the first applications of a discrimination method to gene expression data, Golub *et al.* [17] proposed a "weighted voting scheme" for binary classification. This method turns out to be a minor variant of the sample ML rule corresponding to special case 3. For two classes $k = 1$ and 2, the sample ML rule classifies an observation $\mathbf{x} = (x_1, \ldots, x_p)$ as 1 iff

$$\sum_{j=1}^{p} \frac{(x_j - \bar{x}_{2j})^2}{\hat{\sigma}_j^2} \geq \sum_{j=1}^{p} \frac{(x_j - \bar{x}_{1j})^2}{\hat{\sigma}_j^2},$$

that is, iff

$$\sum_{j=1}^{p} \frac{(\bar{x}_{1j} - \bar{x}_{2j})}{\hat{\sigma}_j^2}\Big(x_j - \frac{(\bar{x}_{1j} + \bar{x}_{2j})}{2}\Big) \geq 0.$$

The discriminant function can be rewritten as $\sum_j v_j$, where $v_j = a_j(x_j - b_j)$, $a_j = (\bar{x}_{1j} - \bar{x}_{2j})/\hat{\sigma}_j^2$, and $b_j = (\bar{x}_{1j} + \bar{x}_{2j})/2$. This is almost the same function as used in Golub *et al.*, except for a_j which Golub *et al.* define as $a_j = (\bar{x}_{1j} - \bar{x}_{2j})/(\hat{\sigma}_{1j} + \hat{\sigma}_{2j})$. Not only is $\hat{\sigma}_{1j} + \hat{\sigma}_{2j}$ an unusual way to calculate the standard error of a difference, but having standard deviations instead of variances in the denominator of a_j produces the wrong units.

For each prediction made by the classifier, Golub *et al.* also define a prediction strength, PS, which indicates the "margin of victory"

$$PS = \frac{\max(V_1, V_2) - \min(V_1, V_2)}{\max(V_1, V_2) + \min(V_1, V_2)},$$

where $V_1 = \sum_j \max(v_j, 0)$ and $V_2 = \sum_j \max(-v_j, 0)$. Golub *et al.* choose a conservative prediction strength threshold of .3 below which no predictions are made. The prediction strengths of Golub *et al.* are related to the prediction votes defined in Section 3.5 for aggregated predictors. An analogue of the votes of Section 3.5 for the gene voting predictor is given by $\max(V_1, V_2)/(V_1 + V_2) = \max(V_1, V_2)/(\max(V_1, V_2) + \min(V_1, V_2)) \geq PS$. Note that here the voting is over genes rather than predictors.

3.3 Nearest neighbor classifiers

Nearest neighbor (NN) methods are based on a distance function for pairs of observations, such as the Euclidean distance or one minus the correlation. For the gene expression data considered here, the distance between two mRNA samples, with gene expression profiles $\mathbf{x} = (x_1, \ldots, x_p)$ and $\mathbf{x}' = (x_1', \ldots, x_p')$, is based on the correlation between their two gene expression profiles:

$$r_{\mathbf{x},\mathbf{x}'} = \frac{\sum_{j=1}^{p}(x_j - \bar{x})(x_j' - \bar{x}')}{\sqrt{\sum_{j=1}^{p}(x_j - \bar{x})^2}\sqrt{\sum_{j=1}^{p}(x_j' - \bar{x}')^2}}.$$

The k *nearest neighbor* rule, due to Fix and Hodges [15], proceeds as follows to classify test set observations on the basis of the learning set. For each element in the test set: (i) find the k closest observations in the learning set, and (ii) predict the class by majority vote, *i.e.*, choose the class that is most common among those k neighbors.

The number of neighbors k is chosen by cross-validation, that is, by running the nearest neighbor classifier on the learning set only. Each observation in the learning set is treated in turn as if it were in the test set: its distance to all of the other learning set observations (except itself) is computed and it is classified by the nearest neighbor rule. The classification for each learning set observation is then compared to the truth to produce the cross-validation error rate. This is done for a number of k's (here $k \in \{1, ..., 21\}$) and the k for which the cross-validation error rate is smallest is retained for use on the test set.

3.4 Classification trees

Binary tree structured classifiers are constructed by repeated splits of subsets (nodes) of the measurement space $\mathcal{X}$ into two descendant subsets, starting with $\mathcal{X}$ itself. Each terminal subset is assigned a class label and the resulting partition of $\mathcal{X}$ corresponds to the classifier.

There are three main aspects to tree construction: (i) the selection of the splits; (ii) the decision to declare a node terminal or to continue splitting; (iii) the assignment of each terminal node to a class.

Different tree classifiers use different approaches to deal with these three issues. Here, we use the *CART - Classification And Regression Trees* - method described in Breiman *et al.* [9] and implemented in the CART software version 1.310 (also implemented in the S-Plus function `tree()`). Single pruned trees are grown using 10-fold cross-validation for estimating the classification error.

3.5 Aggregating classifiers

Breiman [5, 7] found that gains in accuracy could be obtained by *aggregating predictors* built from perturbed versions of the learning set (see Sections 3.5.1 and 3.5.2 for different methods for generating perturbed learning sets). In classification, the multiple versions of the predictors can be aggregated by *plurality voting*, *i.e.*, the "winning" class is the one being predicted by the largest number of predictors. The bias and variance properties of aggregated predictors were studied in Breiman [7]. The key to improved accuracy is the possible instability of the prediction method, *i.e.*, whether small changes in the learning set result in large changes in the predictor. Unstable procedures tend to benefit the most from aggregation. Classification trees tend to be unstable while, for example, nearest neighbor methods tend to be stable. We will thus aggregate only the CART predictors. The trees used for aggregation are maximal "exploratory" trees, in the sense that they are grown until each terminal node contains observations from only a single class.

More precisely, let $C(\cdot, \mathcal{L}_b)$ denote the classifier built from the bth perturbed learning set $\mathcal{L}_b$ and let w_b denote the weight given to predictions made by this classifier. The predicted class for an observation $\mathbf{x}$ is given by

$$\text{argmax}_k \sum_b w_b \, I(C(\mathbf{x}, \mathcal{L}_b) = k),$$

where $I(.)$ denotes the indicator function, equaling 1 if the condition in parentheses is true, and 0 otherwise.

For aggregated classifiers, *prediction votes* (PV) assessing the strength of a prediction may be defined for each observation. The prediction vote for an observation $\mathbf{x}$ is defined to be

$$PV(\mathbf{x}) = \frac{\max_k \sum_b w_b \, I(C(\mathbf{x}, \mathcal{L}_b) = k)}{\sum_b w_b}.$$

When the perturbed learning sets are given equal weights, *i.e.* $w_b = 1$, the prediction vote is simply the proportion of votes for the "winning" class, regardless of whether it is correct or not. Prediction votes belong to $[0, 1]$. The hope is that the magnitude of the prediction vote will reflect how confident we can be of a given prediction. Note that Schapire *et al.* [27] use prediction votes, which they call "weights", to calculate their classification margins.

There are two main classes of methods for generating perturbed versions of the learning set: bagging and boosting. In the *b*ootstrap *agg*regat*ing* or *bagging* procedure [5], perturbed learning sets of the same size as the original learning set are formed by forming bootstrap replicates of the learning set. In *boosting* the data are re-sampled *adaptively* and the predictors are aggregated by *weighted* voting [16].

3.5.1 Bagging

Non-parametric bootstrap. In the simplest form of bagging, perturbed learning sets of the same size as the original learning set are formed by drawing at random with replacement from the learning set, *i.e.*, by forming non-parametric bootstrap replicates of the learning set. Predictors are built for each perturbed dataset and aggregated by plurality voting ($w_b = 1$).

As demonstrated in Breiman [6], the bagging procedure has valuable by-products. At each iteration, about 37% of the observations in the original learning set do not appear in the bootstrap learning set $\mathcal{L}_b$. These *out-of-bag* observations yield an unused test set for the predictor $C(\cdot, \mathcal{L}_b)$ with no additional computing cost. The out-of-bag observations can be used to estimate misclassification rates of bagged predictors as follows:

$$\frac{1}{n_L} \sum_i I\Big(y_i \;\neq\; \text{argmax}_k \sum_{\{b: (\mathbf{x}_i, y_i) \notin \mathcal{L}_b\}} I(C(\mathbf{x}_i, \mathcal{L}_b) = k)\Big).$$

That is, for each observation $(\mathbf{x}_i, y_i)$ in the learning set $\mathcal{L}$, a prediction is obtained by aggregating the classifiers $C(\cdot, \mathcal{L}_b)$ such that $(\mathbf{x}_i, y_i) \notin \mathcal{L}_b$. The out-of-bag estimate of the error rate is the error rate for these out-of-bag predictions.

A general problem of the non-parametric bootstrap for small datasets is the discreteness of the sampling space. We describe next two methods for getting around this problem: the parametric bootstrap and the use of convex pseudo-data.

Parametric bootstrap. In this parametric form of bagging, perturbed learning sets are generated according to a mixture of multivariate normal (MVN) distributions (personal communication with Mark van der Laan). For each class k, the mean vector and covariance matrix of the multivariate normal distribution are taken to be the class sample mean vector and covariance matrix, respectively. The class mixing probabilities are taken to be the class proportions in the actual learning set. We require that at least one observation be sampled from each class. Predictors are built for each perturbed dataset and aggregated by plurality voting ($w_b = 1$). We consider sampling from multivariate normal distributions with both diagonal and non-diagonal covariance matrices.

Convex pseudo-data. Given a learning set $\mathcal{L} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_{n_L}, y_{n_L})\}$, Breiman [8] suggests creating perturbed learning sets based on *convex pseudo-data (CPD)*. Each perturbed learning set is generated by repeating the following n_L times:

1. Select two instances $(\mathbf{x}, y)$ and $(\mathbf{x}', y')$ at random from the learning set $\mathcal{L}$.
2. Select at random a number v from the interval $[0, d]$, $0 \leq d \leq 1$, and let $u = 1 - v$.
3. The new instance is $(\mathbf{x}'', y'')$ where $y'' = y$ and $\mathbf{x}'' = u\mathbf{x} + v\mathbf{x}'$.

As in bagging, multiple altered learning sets $\mathcal{L}_b$, of the same size as the original learning set $\mathcal{L}$, are generated and aggregated by plurality voting. Note that when d is 0, CPD reduces to bagging, and that the larger d, the greater the amount of smoothing. In practice, when a test set is not available d could be chosen by cross-validation.

3.5.2 Boosting

Boosting was first proposed by Freund and Schapire [16]. Here, the data are re-sampled *adaptively* so that the weights in the re-sampling are increased for those cases most often misclassified. The aggregation of predictors is done by *weighted* voting. More precisely, for a learning set $\mathcal{L} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_{n_L}, y_{n_L})\}$, let $\{p_1, \ldots, p_{n_L}\}$ denote the re-sampling probabilities, initialized to be equal. The bth step of the boosting algorithm [7] (an adaptation of `AdaBoost`) is

1. Using the current re-sampling probabilities $\{p_1, \ldots, p_{n_L}\}$, sample with replacement from $\mathcal{L}$ to get a learning set $\mathcal{L}_b$ of size n_L.
2. Build a classifier $C(\cdot, \mathcal{L}_b)$ based on $\mathcal{L}_b$.
3. Run the learning set $\mathcal{L}$ through the classifier $C(\cdot, \mathcal{L}_b)$ and let $d_i = 1$ if the ith case is classified incorrectly and $d_i = 0$ otherwise.
4. Define
$$\epsilon_b = \sum_i p_i d_i \qquad \text{and} \qquad \beta_b = (1 - \epsilon_b)/\epsilon_b$$
and update the re-sampling probabilities for the $(b+1)$st step by
$$p_i = \frac{p_i \beta_b^{d_i}}{\sum_i p_i \beta_b^{d_i}}.$$

After B steps, the classifiers $C(\cdot, \mathcal{L}_1), \ldots, C(\cdot, \mathcal{L}_B)$ are aggregated by weighted voting, with $C(\cdot, \mathcal{L}_b)$ having weight $w_b = \log(\beta_b)$. In the event that $\epsilon_b \geq 1/2$ or $\epsilon_b = 0$ the re-sampling probabilities are reset to be equal. Note that bagging is a special case of boosting, where the p_i's are uniform at each step and the perturbed predictors are given equal weight in the voting.

4 Data and pre-processing

4.1 Datasets

4.1.1 Lymphoma dataset

This dataset comes from a study of gene expression in the three most prevalent adult lymphoid malignancies: B-cell chronic lymphocytic leukemia (B-CLL), follicular lymphoma (FL), and diffuse large B-cell lymphoma (DLBCL) (Alizadeh *et al.* [2], `http://genome-www.stanford.edu/lymphoma`). Gene expression levels were measured using a specialized cDNA microarray, the Lymphochip, containing genes that are preferentially expressed in lymphoid cells or which are of known immunological or oncological importance. In each microarray experiment, fluorescent cDNA targets were prepared from an experimental mRNA sample (red-fluorescent dye Cy5) and a reference mRNA sample derived from a pool of 9 different lymphoma cell lines (green-fluorescent dye Cy3). This study produced gene expression data for $p = 4,682$ genes in $n = 81$ mRNA samples. The mRNA samples comprise 29 cases of B-CLL (class 1), 9 cases of FL (class 2), and 43 cases of DLBCL (class 3). The data are collected into an 81×4682 matrix $X = (x_{ij})$, where x_{ij} denotes the base 2 logarithm of the Cy5/Cy3 fluorescence ratio for gene j in mRNA sample i. Figure 1 displays images of the 81×81 correlation matrix between gene expression profiles for the 81 tumors. As demonstrated by Alizadeh *et al.* [2], the DLBCL class is heterogeneous and comprises two distinct subclasses with different clinical behaviors. However, it is nonetheless distinct from the other two classes, B-CLL and FL, and in Section 6 we compare the ability of different discrimination methods to distinguish between these three classes.

*** Place Figure 1 about here ***

4.1.2 Leukemia dataset

The leukemia dataset is described in the recent paper of Golub *el al.* [17] and available at `http://www.genome.wi.mit.edu/MPR`. This dataset comes from a study of gene expression in two types of acute leukemias: acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML). Gene expression levels were measured using Affymetrix high-density oligonucleotide arrays containing $p = 6,817$ human genes. The data comprise 47 cases of ALL (38 B-cell ALL and 9 T-cell ALL) and 25 cases of AML.

The following pre-processing steps were applied (personal communication, Pablo Tamayo): (i) thresholding: floor of 100 and ceiling of 16,000; (ii) filtering: exclusion of genes with $\max / \min \leq 5$ and $(\max - \min) \leq 500$, where max and min refer respectively to the maximum and minimum expression levels of a particular gene across mRNA samples; (iii) base 10 logarithmic transformation.

The data are then summarized by a 72×3571 matrix $X = (x_{ij})$, where x_{ij} denotes the base 10 logarithm of the expression level for gene j in mRNA sample i. Figure 2 displays images of the 72×72 correlation matrix between gene expression profiles for the 72 tumors.

In this study, the data are already divided into a learning set of 38 mRNA samples and a test set of 34 mRNA samples. The observations in the two sets came from different labs and were collected at different times. The test set is actually more heterogeneous than the learning set as it comprises a broader range of samples, including samples from peripheral blood as well as bone marrow, from childhood AML patients, and from laboratories that used different sample preparation protocols. In Section 6 we address the impact of this heterogeneity on the performance of the predictors.

*** Place Figure 2 about here ***

4.1.3 NCI 60 dataset

In this study, cDNA microarrays were used to examine the variation in gene expression among the 60 cell lines from the National Cancer Institute's anti-cancer drug screen (NCI 60) (Ross *et al.* [26], `http://genome-www.stanford.edu/nci60`). The cell lines are derived from tumors with different sites of origin: 7 breast, 5 central nervous system (CNS), 7 colon, 6 leukemia, 8 melanoma, 9 non-small-cell-lung-carcinoma (NSCLC), 6 ovarian, 2 prostate, 9 renal, 1 unknown. Gene expression was studied using cDNA microarrays with 9,703 spotted cDNA sequences. For each of the 60 cell lines, fluorescent cDNA targets were prepared from an mRNA sample (red-fluorescent dye Cy5). The reference sample (green-fluorescent dye Cy3) used in all hybridizations was prepared by pooling equal mixtures of mRNA from 12 of the cell lines. To investigate the reproducibility of the entire experiment (cell culture, mRNA isolation, labeling, hybridization, etc.), a leukemia (K562A) and a breast cancer (MCF7) cell line were the object of three independent microarray experiments. Because of their small class size, the two prostate cell line observations were excluded from our analysis, as well as the unknown cell line observation. After screening out genes with more than 2 missing data points, the data are collected into a 61×5244 matrix $X = (x_{ij})$, where x_{ij} denotes the base 2 logarithm of the Cy5/Cy3 fluorescence ratio for gene j in mRNA sample i. Figure 3 displays images of the 61×61 correlation matrix between gene expression profiles for the 61 cell lines.

*** Place Figure 3 about here ***

4.2 Imputation of missing data

For the lymphoma and NCI 60 data, some arrays contain a number of genes with unreliable or missing data (the mean percentage of missing data points per array is 6.6% for the lymphoma data and 3.3% for the NCI 60 data). Some of the discrimination methods examined here are able to deal with missing data (*e.g.* CART), however, others require complete data (*e.g.* Fisher linear discriminant analysis). For imputing the missing data, we use a simple k nearest neighbor algorithm, in which the neighbors are the genes and the distance between neighbors is based on their correlation. For each gene with missing data: (i) compute its correlation with all other $p-1$ genes, and (ii) for each missing entry, identify the k nearest genes having complete data for this entry and impute the missing entry by the average of the corresponding entries for the k neighbors. We used $k = 5$.

4.3 Standardization

It is common practice to use the correlation between the gene expression profiles of two mRNA samples to measure their similarity [2, 23, 26] . Consequently, we standardize the observations (arrays) to have mean 0 and variance 1 across variables (genes). With the data standardized in this fashion, the distance between two mRNA samples may be measured by their Euclidean distance.

4.4 Gene selection

A large number of genes exhibit near constant expression levels across samples. We thus perform a preliminary selection of genes on the basis of the ratio of their between-groups to within-groups sum of squares. For a gene j, this ratio is

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

where $\bar{x}_{.j}$ denotes the average expression level of gene j across all samples and $\bar{x}_{kj}$ denotes the average expression level of gene j across samples belonging to class k. We compare predictors based on the p genes with the largest BSS/WSS ratios. We also briefly consider slightly more complicated variable selection criteria (Sections 5 and 6).

Note. Golub *et al.* [17] use a different method for standardizing the data and for selecting genes than described above. Prior to the logarithmic transformation and after thresholding, a subset of p genes are selected on the basis of the statistics a_j: the $p/2$ genes with the largest a_j and the $p/2$ genes with the smallest a_j (see Section 3.2 for the definition of a_j). Golub *et al.* found $p = 50$ to be adequate for the ALL/AML dataset. The learning data matrix is then log-transformed and its columns (genes) are standardized to have mean 0 and variance 1. The quantities b_j are computed from this transformed and standardized matrix. The test set data are also log-transformed and genes are standardized using the learning set averages and variances.

For the sake of completeness, our comparison study includes the weighted voting scheme of Golub *et al.* with a_j calculated using standard deviations instead of variances and with the data processed as described in the previous paragraph. We refer to the resulting predictor as "Golub" and it is of interest to compare its performance to that of DLDA with the data processing steps of Sections 4.3 and 4.4.

5 Study design

In the absence of genuine test sets, the different predictors are compared based on random divisions of each dataset into a learning set $\mathcal{L}$ and a test set $\mathcal{T}$. There are no widely accepted guidelines for choosing the relative size of these artificial learning sets and test sets. A possible choice is to leave out a randomly selected 10% of the instances to use as a test set (*e.g.* Breiman [5]). However, in our case, test sets containing 10% of the data are not sufficiently large to provide adequate discrimination between the classifiers. Since our main purpose is to compare classifiers, rather than estimate generalization error rates, we choose to sacrifice training data and increase test set size to one third of the data ($2:1$ scheme).

In the principal comparison, for each learning set/test set (LS/TS) run, the p genes with the largest BSS/WSS are selected using the learning set: $p = 50$ for the lymphoma dataset, $p = 40$ for the leukemia dataset, and $p = 30$ for the NCI 60 dataset. (For a comparison involving all predictors, Fisher linear discriminant analysis sets an upper limit on the size of the gene set because of rank issues.)

Next, predictors are constructed using the learning set and test set error rates are obtained by applying the predictors to the test set. Aggregated predictors (bagging and boosting) are built from $B = 50$ "pseudo" learning sets (increasing the number of iterations didn't seem to affect the outcome). For CPD, several values of the parameter d are examined: $d = .1, .25, .5, .75, 1$.

This entire procedure is repeated $N = 150$ times.

Test set error rates. Each LS/TS run yields a test set error rate for each predictor; boxplots are used to summarize these error rates over the runs.

Observation-wise error rates. We also record observation-wise error rates, *i.e.*, for a given observation, the proportion of times it was classified incorrectly (out of the LS/TS runs in which it was in the test set). Observation-wise error rates can be summarized by means of survival plots, *i.e.*, by plotting against $V\%$ the fraction of observations with observation-wise error rate less than $1 - V\%$. For each observation, prediction votes for aggregated predictors are recorded each time that particular observation is in the test set. Votes for individual observations are summarized by boxplots and compared to the observation-wise error rates.

Variable selection. The effect of increasing ($p = 200$) or decreasing ($p = 10$) the number of variables is examined. A "smarter" BSS/WSS criterion is also applied to the lymphoma data. For $p = 10$ genes, this criterion consists of selecting the 5 genes with the largest BSS/WSS ratio (as before) and the 5 genes with the largest BSS/WSS ratio when the two largest classes (B-CLL and DLBCL) are pooled. Such a criterion should allow better discrimination of the smaller FL class (Figure 18).

*** Place Figure 18 about here ***

6 Results

6.1 Test set error rates

For the main comparison, Figures 4, 5, 6, and 7 display boxplots of the error rates for each classifier using the $2:1$ scheme. For the leukemia dataset, we compared classifiers based on their ability to distinguish ALL from AML (two-class problem) and to distinguish between ALL B-cell, ALL T-cell, and AML (three-class problem). The figures for each dataset contain three panels: boxplots of error rates for discriminant analysis (DA), boxplots of error rates for CART-based classifiers, and boxplots of error rates comparing the nearest neighbor classifier to the best DA and CART-based classifiers. It was often hard to distinguish between the best CART-based classifiers, as boosting and CPD tended to have very similar error rates. In cases when the best classifiers had the same median error rate, we chose the classifier with the smallest mean error rate. In general, the nearest neighbor and DLDA predictors had the smallest error rates, while FLDA had the highest error rates. With the exception of the NCI 60 data, the error rates seemed fairly low given the limited amount of data.

*** Place Figures 4, 5, 6, 7 about here ***

Nearest neighbors. The parameter k of the nearest neighbor predictor was selected by cross-validation and was usually quite small for each dataset: 1 or 2 for about half of the runs and generally less than 7. This suggests that very good predictions can be obtained from the class of the observation most highly correlated to the observation to be predicted. Examination of the correlation matrices between mRNA samples (Figures 1, 2, and 3) gives an explanation for the good performance of nearest neighbor classifiers. For the p genes with the largest ratio of BSS/WSS ($p = 50$ for lymphoma data, $p = 40$ for leukemia data, and $p = 30$ for NCI 60 data), observations within the same class tend to have high positive correlations (patches of red along the diagonal), while observations belonging to different classes tend to have high negative correlations (patches of green off the diagonal). This pattern is much more subtle for the correlation matrices based on all genes and it is to be expected that the nearest neighbor method benefits greatly from the initial selection of variables. The pattern is also much stronger for the lymphoma and leukemia data than for the NCI 60 data.

Fisher linear discriminant analysis. On the opposite end of the performance spectrum is FLDA. One possible reason for the poor performance of FLDA is that it is a "global" method, *i.e.*, it borrows strength from the bulk of the data, and as a result some observations may not be well represented by the discriminant variables (only one discriminant variable for the leukemia dataset and two for the lymphoma dataset). In contrast, nearest neighbor methods are "local". More importantly, with limited data and a fairly large number of genes p, the matrices of between-groups and within-groups sum of squares may be quite unstable and provide poor estimates of the corresponding population quantities. We show below that the performance of FLDA can dramatically improve when the number of variables is decreased to $p = 10$ and the variables are selected according to a "smarter"

BSS/WSS criterion. Also, as described next, DLDA, which ignores correlations between genes, results in predictors with much reduced misclassification rates.

Diagonal discriminant analysis. A simple ML discriminant rule for multivariate normal class densities with diagonal covariance matrices produced impressively low misclassification rates compared to more sophisticated predictors such as bagged classification trees. With the exception of the lymphoma dataset, linear classifiers (DLDA), which assume a common covariance matrix for the different classes, yielded lower error rates than quadratic classifiers (DQDA), which allow different class covariance matrices. The performance of DLDA was especially striking for the NCI 60 dataset, where it performed better than any of the other classifiers. Thus, for the datasets considered here, large gains in accuracy were obtained by ignoring correlations between genes. Such predictors are sometimes called "naive Bayes", as they arise in a Bayesian setting where the predicted class is the one with maximum posterior probability $pr(y = k|\mathbf{x})$.

Weighted voting scheme of Golub *et al.*. For the binary class leukemia dataset, we also examined the performance of the variant of DLDA implemented in Golub *et al.* [17], *i.e.*, DLDA with a_j calculated using standard deviations instead of variances and with the data processed as described at the end of Section 4.4. This method ("Golub") performed similarly to boosting and DQDA, but was inferior to nearest neighbors and especially DLDA which had a median error rate of zero. Note that in contrast to the aggregated predictors in bagging and boosting, the "voting" is over variables (here genes) rather than predictors.

Aggregated CART. CART-based predictors had performance intermediate between FLDA and DLDA and nearest neighbors. Aggregated predictors were generally more accurate than a single tree predictor. The parametric version of bagging (MVN) performed worse than the non-parametric forms (CPD and standard bagging). This may be for any of several reasons, including non-normality and low sample size for the estimation of the class means and covariance matrices. Using diagonal covariance matrices for the class densities didn't seemed to help (results not shown). CPD CART and boosting performed better than other aggregated predictors. Increasing the number of bagging or boosting iterations from 50 to 150 didn't affect the performance of the predictors. The convex pseudo-data method depends on a parameter $d \in [0, 1]$. Several values of d were tried, $d = .05$, .1, .25, .5, .75, 1, and the value of d with the smallest test set error rate was retained. For each dataset this value turned out to be between .5 and 1, suggesting that the performance of CPD is not very sensitive to the value of d controlling the degree of smoothing. We used $d = .75$ in the comparison.

6.2 Individual misclassification rates

We also kept track of observation-wise misclassification rates and prediction votes over the runs. This type of information could be useful for the identification of possibly mislabeled observations.

Figures 10, 11, 12, and 13 display survival plots for the fraction of observations with observation-wise error rate less than $1 - V\%$. These figures also illustrate the good performance of nearest neighbors and DLDA and the poor performance of FLDA.

*** Place Figures 10, 11, 12, 13 about here ***

For aggregated predictors, prediction votes may be used to summarize the strength of a prediction. Figures 14, 15, 16, and 17 display plots of the proportion of correct classifications and three number summaries (median and lower and upper quartiles) of the prediction votes for each observation (over N LS/TS runs). The qualitative correspondence between votes and proportions of correct classifications suggests that votes are good indicators of the ability of each predictor to classify a particular observation correctly. The prediction strengths of Golub *et al.* [17] seem to be highly variable and conservative in comparison to the proportions of correct predictions.

*** Place Figures 14, 15, 16, 17 about here ***

Lymphoma. For the lymphoma dataset, two observations tended to be more difficult to classify and had smaller prediction votes, as indicated in Figure 14. The first observation (index 1, "CLL-70;Lymph node") is a B-CLL case, but the mRNA sample was prepared from a lymph node biopsy specimen rather than from peripheral blood cells as for other B-CLL cases. This observation tended to be classified as an FL case, perhaps reflecting tissue sampling. The other observation (index 39, "DLCL-0042") is believed to be a DLBCL case and tends to be classified as an FL case. The observations from the second class, corresponding to FL cases, were generally harder to predict and had smaller prediction votes than observations from other classes. This may be due to the fact that class 2 only has 9 observations.

Leukemia. For the two-class leukemia dataset, three observations tended to be difficult to classify and had smaller prediction votes, as indicated in Figure 15. Two of these are thought to be AML cases (indices 28 and 66, corresponding to indices 48 and 72 in Figure 15) and the other an ALL T-cell case (index 67, corresponding to index 47 in Figure 15). Observations 66 and 67 were part of the test set in the Golub *et al.* paper and had low prediction strengths of .27 and .15, respectively. Observation 28 was part of the learning set and had a prediction strength of .44 in their cross-validation study.

NCI 60. The performance of the predictors was much worse for the NCI 60 dataset than for the other two datasets. This is probably due to the small class sizes and the heterogeneity of some of the classes (breast and NSCLC). Certain classes were easier to predict than others (*e.g.* melanoma, leukemia, colon). Observations from these classes exhibit strong correlation among themselves, as indicated by the patches of red along the diagonal of the correlation matrices (Figure 3). The triplicate leukemia (K562A) and breast cancer (MCF7) samples were also strongly correlated, suggesting good reproducibility of the experimental procedure.

6.3 Choice of predictor variables

In general, for the lymphoma or leukemia datasets, increasing the number of variables to $p = 200$ didn't affect greatly the performance of the various predictors (Figure 8). However, for the NCI 60 dataset, the error rates were generally lower for $p = 200$ (Figure 9). This is probably due to the larger number of classes and the fact that with a small p a crude BSS/WSS criterion isn't able to pick up variables that discriminate between all the classes.

Decreasing the number of variables to $p = 10$ resulted in an improved performance of FLDA. The increase in performance of FLDA was even more pronounced with the "smarter" BSS/WSS criterion (results shown only for the lymphoma dataset). The performance of DLDA and DQDA was not very sensitive to the number of predictor variables, although it improved slightly with the number of variables. The improvement was more pronounced for the NCI 60 data, for the reasons mentioned above.

*** Place Figures 8, 9 about here ***

6.4 Heterogeneity in Golub *et al.* learning and test sets

To address the impact of heterogeneity on the performance of the predictors, we constructed a nearest neighbor classifier based on the 38 observations belonging to the original learning set and predicted the remaining 34 test set observations using this classifier. The resulting error rate was similar to a typical error rate when 38 observations are sampled at random from the pooled data to form a learning set and the remaining 34 are used as a test set. Thus, heterogeneity in the original learning and test sets does not seem to have much of an impact on prediction power.

To examine whether there are systematic differences in gene expression between the original learning set and test set, the observations coming from the learning set were labeled as "0" and those from the test set as "1". We then performed a mini LS/TS random resampling study (2 : 1 scheme). Under the "no difference" hypothesis, a 50% misclassification rate is expected. In reality, we observed a mean misclassification rate of about 25%. To verify that 50% is indeed what we would expect under no difference, we permuted the labels at random with respect to the gene expression profiles and repeated the procedure on the permuted data. We did observe about 50% misclassification rate.

7 Discussion

We have compared the performance of different discrimination methods for the classification of tumors using gene expression data from three recent studies. The rankings of the classifiers were similar across datasets and the main conclusion, for these datasets, is that simple classifiers such as DLDA and nearest neighbors perform remarkably well compared to more sophisticated methods such as aggregated classification trees.

In our principal comparison, with an intermediate number of predictor variables selected according to a crude BSS/WSS criterion, nearest neighbor classifiers and DLDA had the lowest error rates, while FLDA had the highest. CART-based classifiers had performance intermediate between FLDA and nearest neighbors and DLDA, with aggregated classifiers being more accurate than a single tree. The greatest gains from aggregation were obtained by boosting and bagging with CPD. The improvement of CPD over standard bagging (non-parametric bootstrap samples) may be due to the fact that CPD gets around the discreteness of the sampling space by sampling from a smoothed version of the empirical c.d.f. For the datasets considered here, the degree of smoothing was fairly high ($d = .75$). The lack of accuracy of FLDA is likely due to the poor estimation of covariance matrices with a small training set and a fairly large number of genes p. Indeed, decreasing the number of variables resulted in an improved performance of FLDA. Also, ignoring correlations between genes as in DLDA produced impressively low misclassification rates compared to more sophisticated classifiers. For the binary class leukemia dataset, DLDA performed better than the related gene voting scheme of Golub *et al.* [17]. This is due to the corrected variance calculation and performing variable selection on already log-transformed data.

We briefly addressed the impact of variable selection on the relative performance of the classifiers. For the lymphoma and leukemia datasets, the performance of the discrimination methods other than FLDA was fairly insensitive to the number of predictor variables. The accuracy of FLDA improved dramatically for $p = 10$ variables selected according to the "smarter" BSS/WSS criterion. However, without careful pre-screening of the variables, we believe that nearest neighbor, DLDA, or CART-based classifiers are preferable to FLDA. For the NCI 60 data, increasing the number of predictor variables to $p = 200$ improved the accuracy of the classifiers.

Misclassification rates for the different classifiers were estimated based on random divisions of each dataset into a learning set and a test set comprising respectively two thirds and one third of the data ($2:1$ sampling scheme). One needs to distinguish between two tasks: estimating misclassification rates, *i.e.*, estimating the probability that a given classifier will misclassify a new sample drawn from the same distribution as the learning set (also called generalization error), and comparing the misclassification rates of two or more classifiers (see discussion in Chapter 2 of Ripley [25]). The second task, which is our main concern here, is rather easier as classifiers are compared using the same test set. We chose a $2:1$ scheme, rather than the perhaps more standard $9:1$ scheme in the machine learning literature, because for our datasets the later scheme resulted in very small test sets and more difficult discrimination between the classifiers due to the discreteness of the error rates. If our main concern was to estimate generalization error, a $2:1$ scheme would be wasteful of scarce data which could otherwise be used for training. Also, we would need much larger datasets to get reasonably accurate estimates of error rates. Note that since we are performing variable selection on the learning set (BSS/WSS criterion), the out-of-bag method (Breiman [6]) for estimating misclassification rates results in overly optimistic estimates of the error rates and hence is not appropriate here.

There are factors other than accuracy which contribute to the merits of a given classifier. These include simplicity and insight gained into the predictive structure of the data. DLDA is easy to implement and had remarkably low error rates in our study, but it ignores correlations between predictor variables (genes). These correlations are biological realities and when more data become available we may find that ignoring them is problematic. Also, LDA (with diagonal or arbitrary covariance matrix) is unable to handle interactions between predictor variables. Gene interactions are important biologically and may contribute to class distinctions; ignoring them is not desirable. Nearest neighbor classifiers are simple, intuitive and had impressively low error rates compared to more sophisticated classifiers. While they are able to handle interactions between genes, they do so in a "black-box" way and give very little insight into the structure of the data. By contrast, classification trees are capable to exploit and reveal interactions between variables. Trees are easy to interpret and yield information on the relationship between predictor variables and responses by performing stepwise variable selection. However, classification trees tend to be unstable and lacking in accuracy. Their accuracy can be greatly improved by aggregation (bagging or boosting). Although some simplicity is lost by aggregating trees, aggregation may be used as part of a variable selection approach (Fridlyand and Speed, work in progress). A useful by-product of aggregated tree classifiers are the prediction votes which can be used to assess the confidence of each prediction. We have only looked at prediction votes in a qualitative manner; it would be interesting to carry out a more quantitative analysis and explore the use of thresholds for making or not making a particular prediction. Note that the conclusions reached in our study were based on a comparison of classifiers on very small datasets by machine learning standards. As more data become available, we can expect to observe an improvement in the performance of aggregated classifiers relative to simpler classifiers, as trees should be able to correctly identify interactions. We may also be able to use these methods to gain a better understanding of the predictive structure of the data.

Our study did not include certain popular classifiers from the field of machine learning, such as neural networks (Ripley [25]) or support vector machines (SVM) (Vapnik [31]). We deliberately choose to look at simple methods which require little training. While SVMs are receiving a lot of attention and have been applied successfully to some problems (*e.g.* handwritten digit recognition), they require more training than the methods considered here (*e.g.* choice of kernel function K and scale factor λ). Also, the generalization of SVMs to more than two classes is not obvious. We are aware of a few applications of SVMs to gene expression data. SVMs were applied to the ALL/AML data, but didn't improve over a simple nearest neighbor or DLDA classifier (personal communication, Saira Mian). In another application, Brown *et al.* [10] used SVMs to classify genes, rather than mRNA samples. They considered only binary classification (*i.e.*, each class versus its complement) and found that SVMs outperformed cross–validated unaggregated classification trees and FLDA. We looked into applying logistic discrimination and a perceptron classifier (Ripley [25]) to these datasets, but our preliminary runs were not encouraging. For logistic discrimination, we encountered the well-known situation of infinite parameter estimates for perfect linear separation of the classes on the learning set. We then considered Rosenblatt's perceptron learning rule which is specifically designed for linearly separable classes. The perceptron test set error rates were disappointing. We have not considered more sophisticated perceptron algorithms.

A very important issue which remains to be addressed is the identification of "marker" genes for tumor classes. For more than two classes, a crude criterion like BSS/WSS is generally unable to pick up variables that discriminate between all the classes (*cf.* improvement using the "smarter" BSS/WSS criterion for the lymphoma dataset). It also tends to pick up genes that are highly correlated and doesn't reveal interactions between genes. With any variable selection approach, we must be aware of the issue of statistical *vs.* biological significance. A purely statistical approach may identify genes that reflect tissue sampling as opposed to biologically interesting and possibly unknown differences between the various tumors.

Acknowledgments

We are grateful to Ash Alizadeh, Pat Brown, Mike Eisen and Doug Ross for introducing us to this problem and for giving us access to their data. We have also appreciated Pablo Tamayo's assistance with the ALL/AML data. Finally, we would like to thank Leo Breiman, Yoram Gat, David Nelson, Mark van der Laan and Yee Hwa Yang for many helpful discussions and suggestions, and Sam Buttrey for his nearest neighbor routine.

This work was supported in part by an MSRI postdoctoral fellowship (SD), a PMMB Burroughs-Wellcome fellowship (JF), and by the NIH through grant 8R1GM59506A (TPS).

References

[1] The Chipping Forecast. *Supplement to Nature Genetics*, 21, 1999.

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

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

[4] M. Barnard. The secular variations of skull characters in four series of egyptian skulls. *Annals of Eugenics*, 6:352–371, 1935.

[5] L. Breiman. Bagging predictors. *Machine Learning*, 24:123–140, 1996.

[6] L. Breiman. Out-of-bag estimation. Technical report, Statistics Department, U.C. Berkeley, 1996.

[7] L. Breiman. Arcing classifiers. *Annals of Statistics*, 26:801–824, 1998.

[8] L. Breiman. Using convex pseudo-data to increase prediction accuracy. Technical Report 513, Statistics Department, U.C. Berkeley, March 1998.

[9] L. Breiman, J. H. Friedman, R. Olshen, and C. J. Stone. *Classification and regression trees*. The Wadsworth statistics/probability series. Wadsworth International Group, 1984.

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

[11] Y. Chen, E. R. Dougherty, and M. L. Bittner. Ratio-based decisions and the quantitative analysis of cdna microarray images. *Journal of Biomedical Optics*, 2:364–374, 1997.

[12] J. L. DeRisi, V. R. Iyer, and P. O. Brown. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, 278:680–685, 1997.

[13] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci.*, 95:14863–14868, 1998.

[14] R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annal of Eugenics*, 7:179–188, 1936.

[15] E. Fix and J. Hodges. Discriminatory analysis, nonparametric discrimination: consistency properties. Technical report, Randolph Field, Texas: USAF School of Aviation Medicine, 1951.

[16] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55:119–139, 1997.

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

[18] A. J. F. Griffiths, J. H. Miller, D. T. Suzuki, R. C. Lewontin, and W. M. Gelbart. *An Introduction to Genetic Analysis*. W. H. Freeman and Company, New York, 6th edition, 1996.

[19] D. J. Lockhart, H. L. Dong, M. C. Byrne, M. T. Follettie, M. V. Gallo, M. S. Chee, M. Mittmann, C. Wang, M. Kobayashi, and H. Horton. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology*, 14:1675–1680, 1996.

[20] K. V. Mardia, J. T. Kent, and J. M. Bibby. *Multivariate Analysis.* Academic Press, Inc., San Diego, 1979.

[21] G. J. McLachlan. *Discriminant analysis and statistical pattern recognition.* Wiley, New York, 1992.

[22] M. A. Newton, C. M. Kendziorski, C. S. Richmond, F. R. Blattner, and K. W. Tsui. On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data. Technical Report 139, Department of Biostatistics and Medical Informatics, UW Madison, 1999.

[23] C. M. Perou, S. S. Jeffrey, M. van de Rijn, C. A. Rees, M. B. Eisen, D. T. Ross, A. Pergamenschikov, C. F. Williams, S. X. Zhu, J. C. F. Lee, D. Lashkari, D. Shalon, P. O. Brown, and D. Botstein. Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proc. Natl. Acad. Sci.*, 96:9212–9217, 1999.

[24] J. R. Pollack, C. M. Perou, A. A. Alizadeh, M. B. Eisen, A. Pergamenschikov, C. F. Williams, S. S. Jeffrey, D. Botstein, and P. O. Brown. Genome-wide analysis of dna copy-number changes using cdna microarrays. *Nature Genetics*, 23:41–46, 1999.

[25] B. D. Ripley. *Pattern recognition and neural networks.* Cambridge University Press, Cambridge, New York, 1996.

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

[27] R. E. Schapire, Y. Freund, P. Bartlett, and W. S. Lee. Boosting the margin: a new explanation for the effectiveness of voting methods. *Annals of Statistics*, 26:1651–1686, 1998.

[28] M. Schena, editor. *DNA Microarrays : A Practical Approach.* Oxford University Press, 1999.

[29] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown. Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science*, 270:467–470, 1995.

[30] R. Tibshirani, T. Hastie, M. Eisen, D. Ross, D. Botstein, and P. Brown. Clustering methods for the analysis of dna microarray data. Technical report, Department of Health Research and Policy, Stanford University, 1999.

[31] V. N. Vapnik. *The nature of statistical learning theory.* Statistics for engineering and information science. Springer, New York, 2nd edition, 2000.

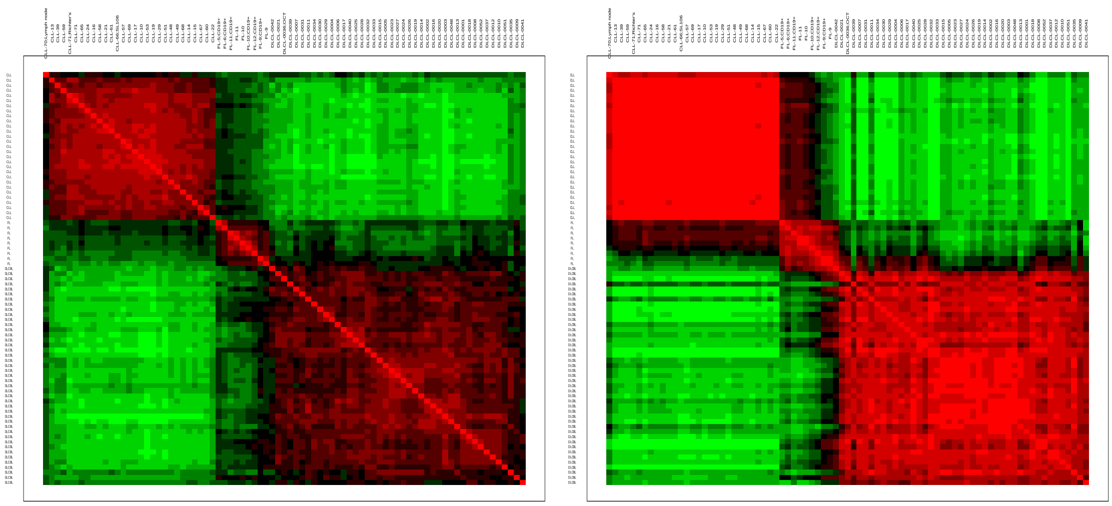


Figure 1: *Lymphoma data - Correlation matrix.* Images of correlation matrix for 81 mRNA samples, based on expression profiles for all $p = 4,682$ genes (left) and for the $p = 50$ genes with the largest BSS/WSS ratio (right). Correlations of 0 are represented in black, increasingly positive correlations are represented with reds of increasing intensity, and increasingly negative correlations are represented with greens of increasing intensity. The mRNA samples are ordered by class, first B-CLL, then FL, and finally DLBCL.

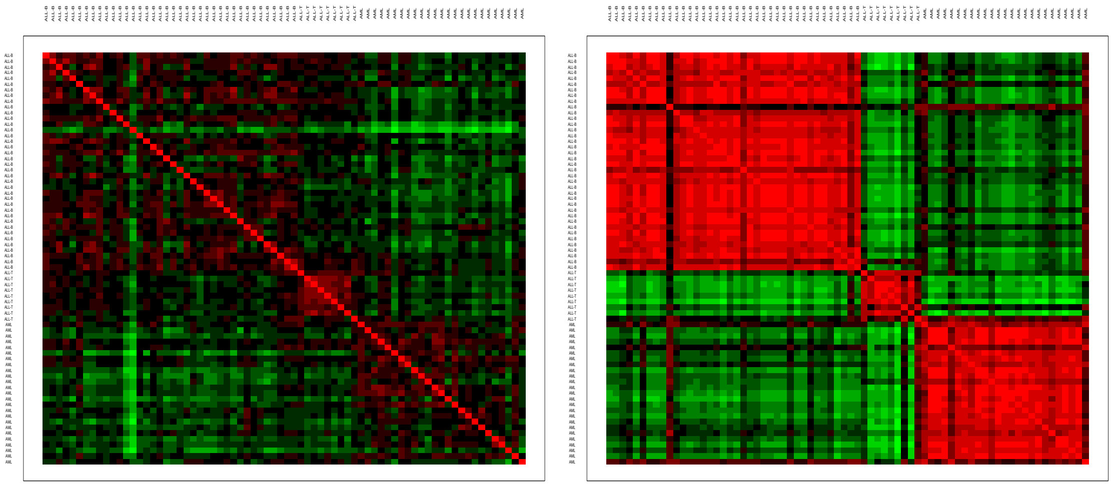


Figure 2: *Leukemia data - Correlation matrix.* Images of correlation matrix for 72 mRNA samples, based on expression profiles for $p = 3,571$ genes (left) and for the $p = 40$ genes with the largest BSS/WSS ratio for the three ALL B-cell, ALL T-cell and AML classes (right). Correlations of 0 are represented in black, increasingly positive correlations are represented with reds of increasing intensity, and increasingly negative correlations are represented with greens of increasing intensity. The mRNA samples are ordered by class, first ALL B-cell, then ALL T-cell, and finally AML.

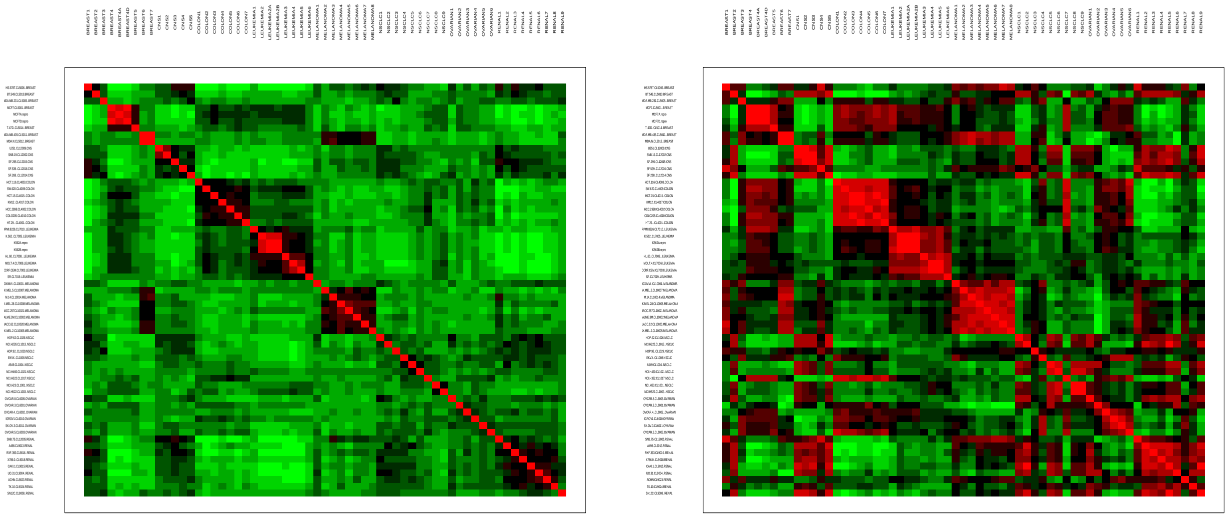


Figure 3: *NCI 60 data - Correlation matrix.* Images of correlation matrix for 61 mRNA samples, based on expression profiles for $p = 5,244$ genes (left) and for the $p = 30$ genes with the largest BSS/WSS ratio (right). Correlations of 0 are represented in black, increasingly positive correlations are represented with reds of increasing intensity, and increasingly negative correlations are represented with greens of increasing intensity. The mRNA samples are ordered by class: 7+2 breast, 5 CNS, 7 colon, 6+2 leukemia, 8 melanoma, 9 NSCLC, 6 ovarian, 9 renal.

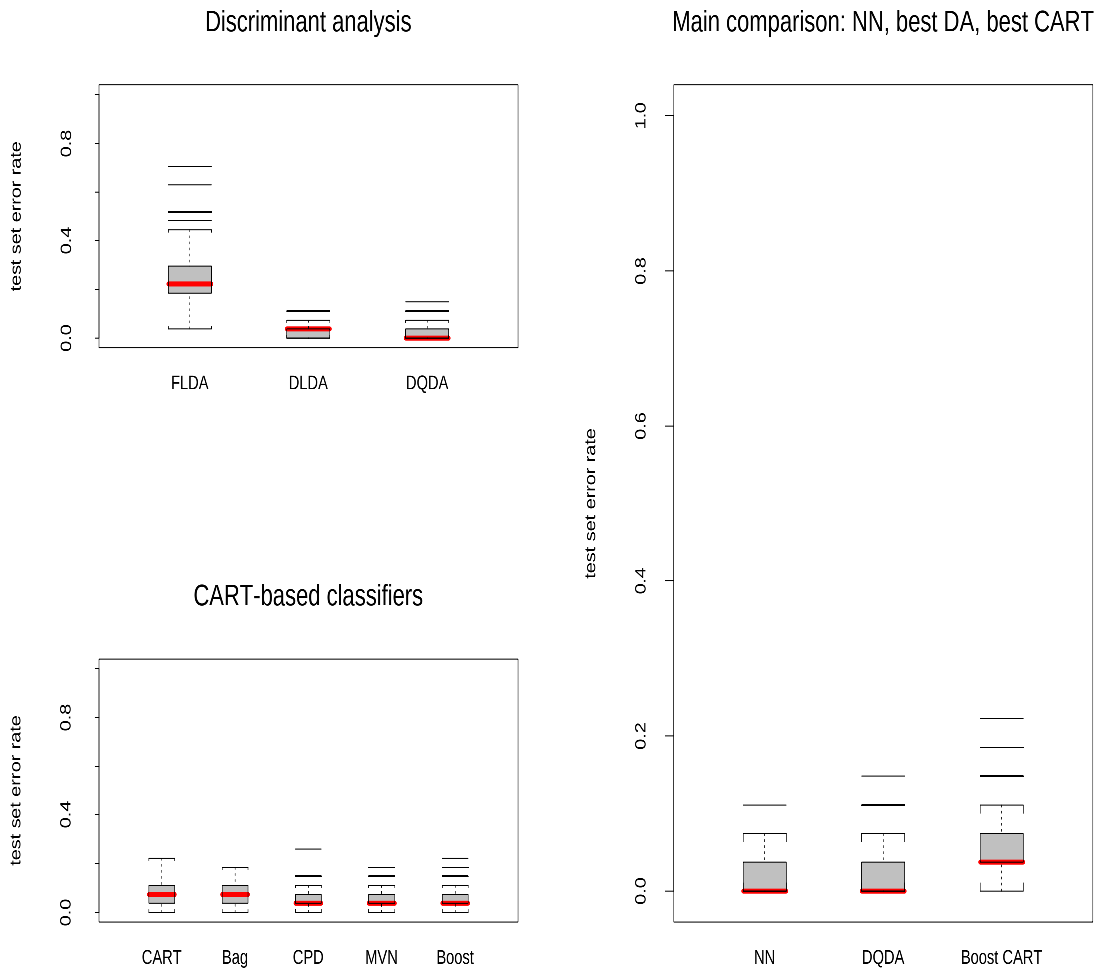


Figure 4: *Lymphoma data - Test set error rates.* Boxplots of test set error rates for classifiers built using the $p = 50$ genes with the largest BSS/WSS; $N = 150$ LS/TS runs for $2:1$ sampling scheme.

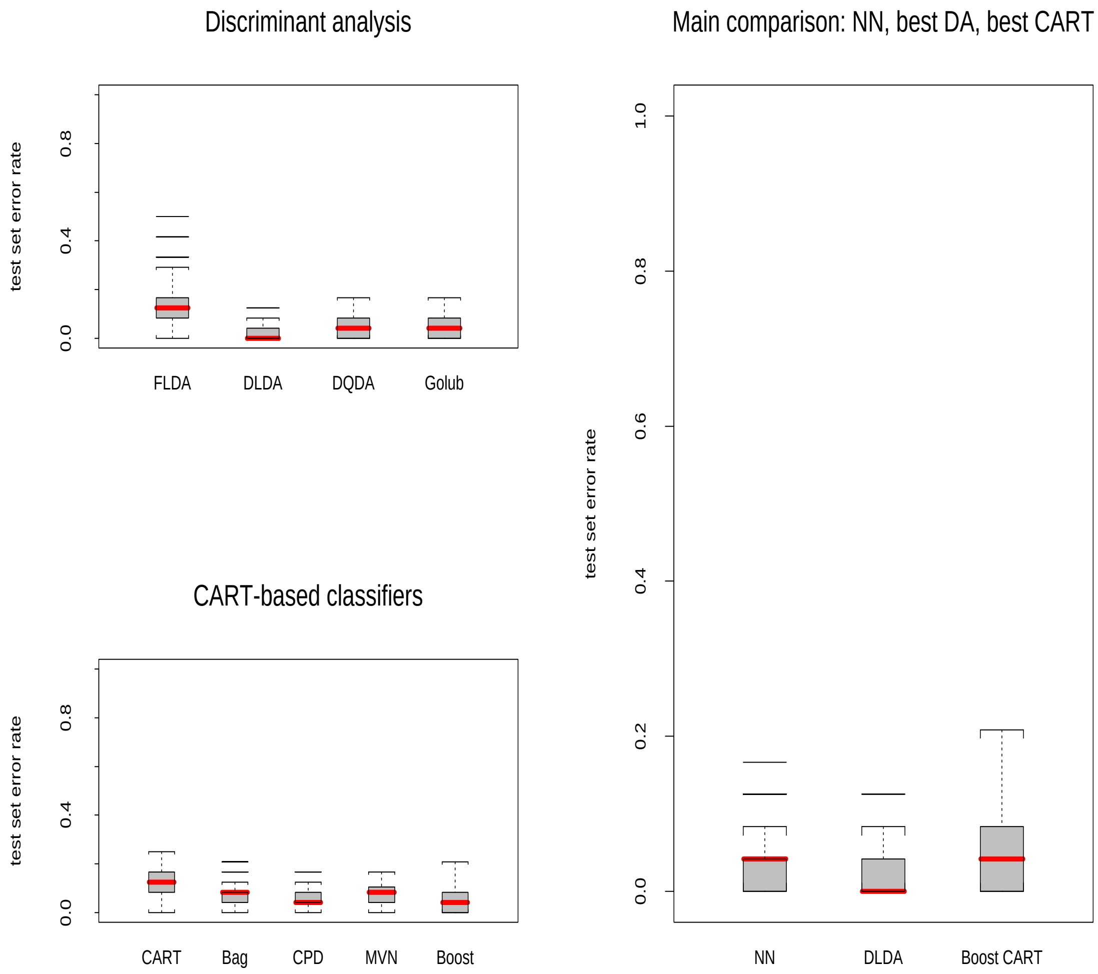


Figure 5: *Leukemia data, two classes - Test set error rates.* Boxplots of test set error rates for classifiers built using the $p = 40$ genes with the largest BSS/WSS; $N = 150$ LS/TS runs for $2:1$ sampling scheme.

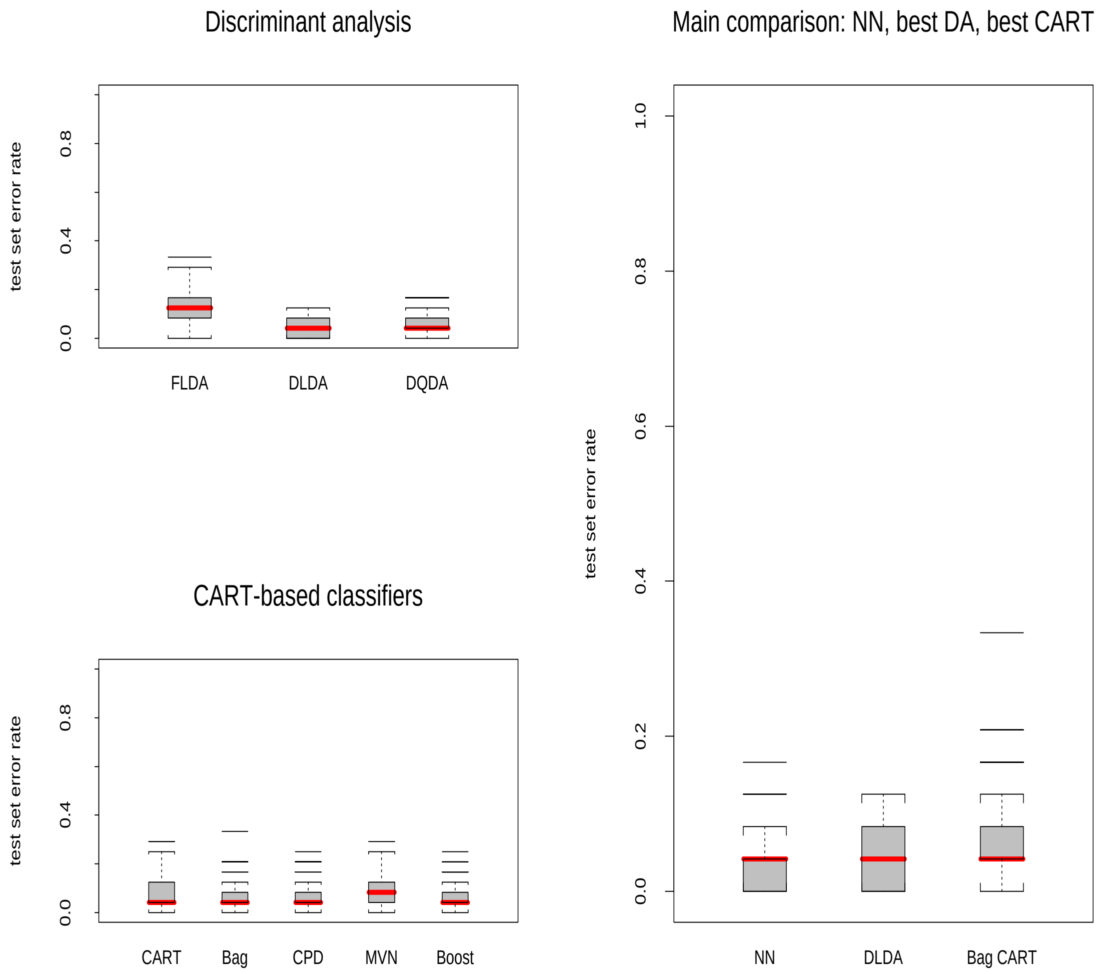


Figure 6: *Leukemia data, three classes - Test set error rates.* Boxplots of test set error rates for classifiers built using the $p = 40$ genes with the largest BSS/WSS; $N = 150$ LS/TS runs for $2:1$ sampling scheme.

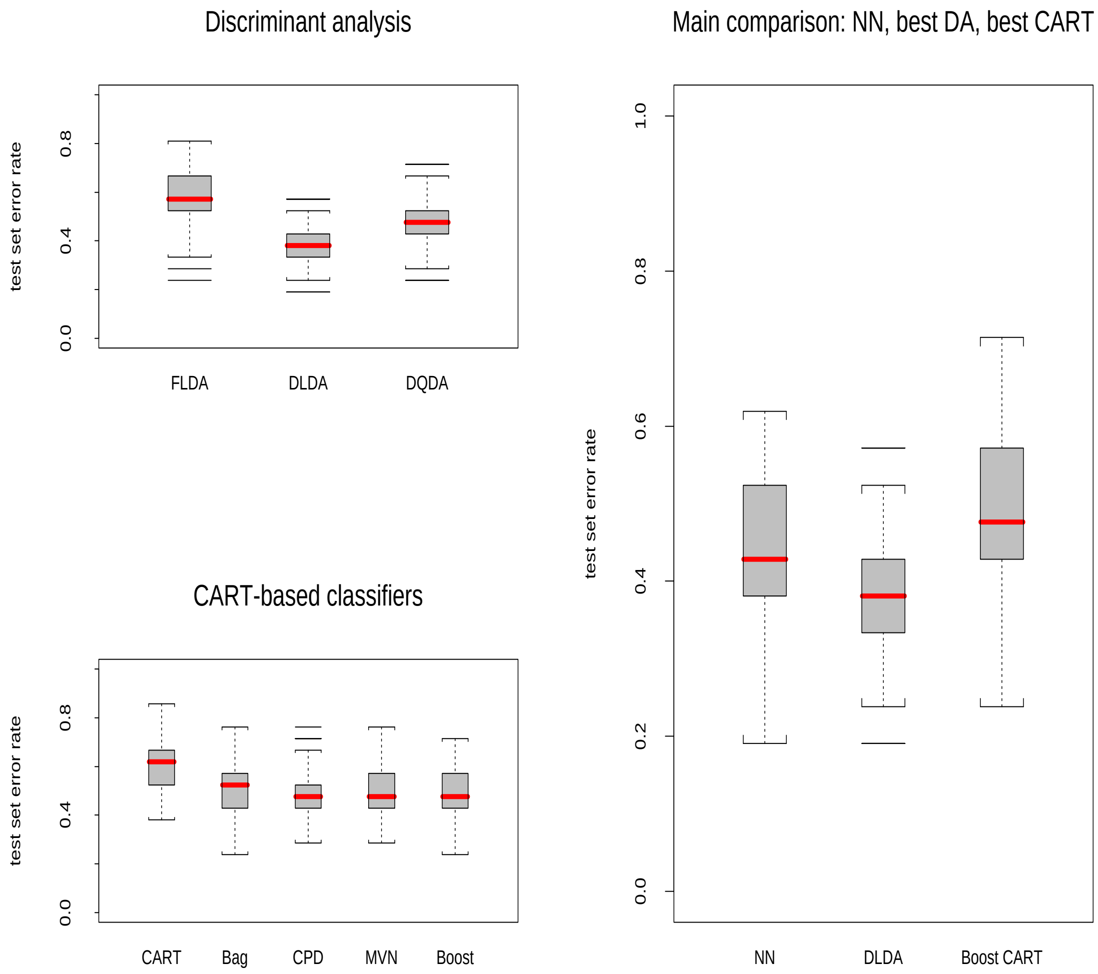


Figure 7: *NCI 60 data - Test set error rates.* Boxplots of test set error rates for classifiers built using the $p = 30$ genes with the largest BSS/WSS; $N = 150$ LS/TS runs for $2:1$ sampling scheme.

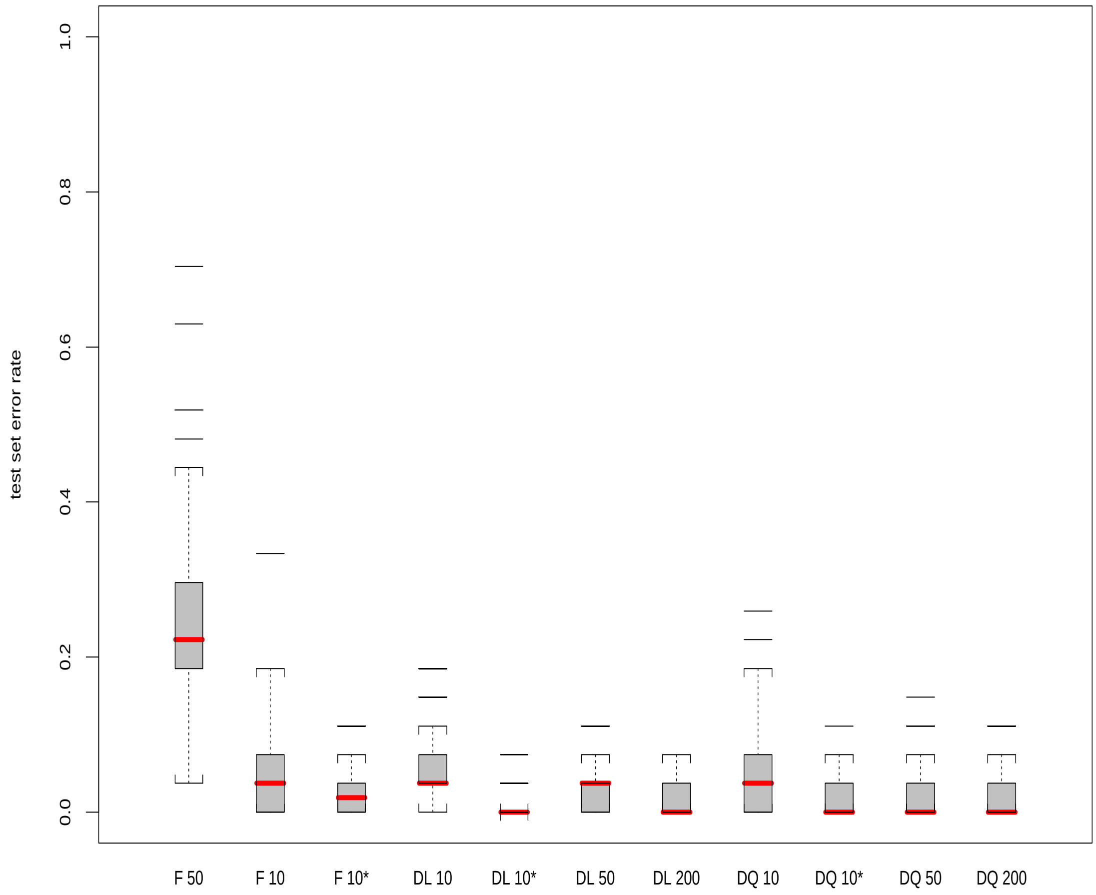


Figure 8: *Lymphoma data - Test set error rates.* Boxplots of test set error rates for Fisher linear discriminant analysis (F), diagonal linear discriminant analysis (DL) and diagonal quadratic discriminant analysis (DQ) based on the $p = 10, 50, 200$ genes with the largest BSS/WSS and $p = 10$ genes chosen according to the "smarter" BSS/WSS criterion; $N = 150$ LS/TS runs for $2:1$ scheme.

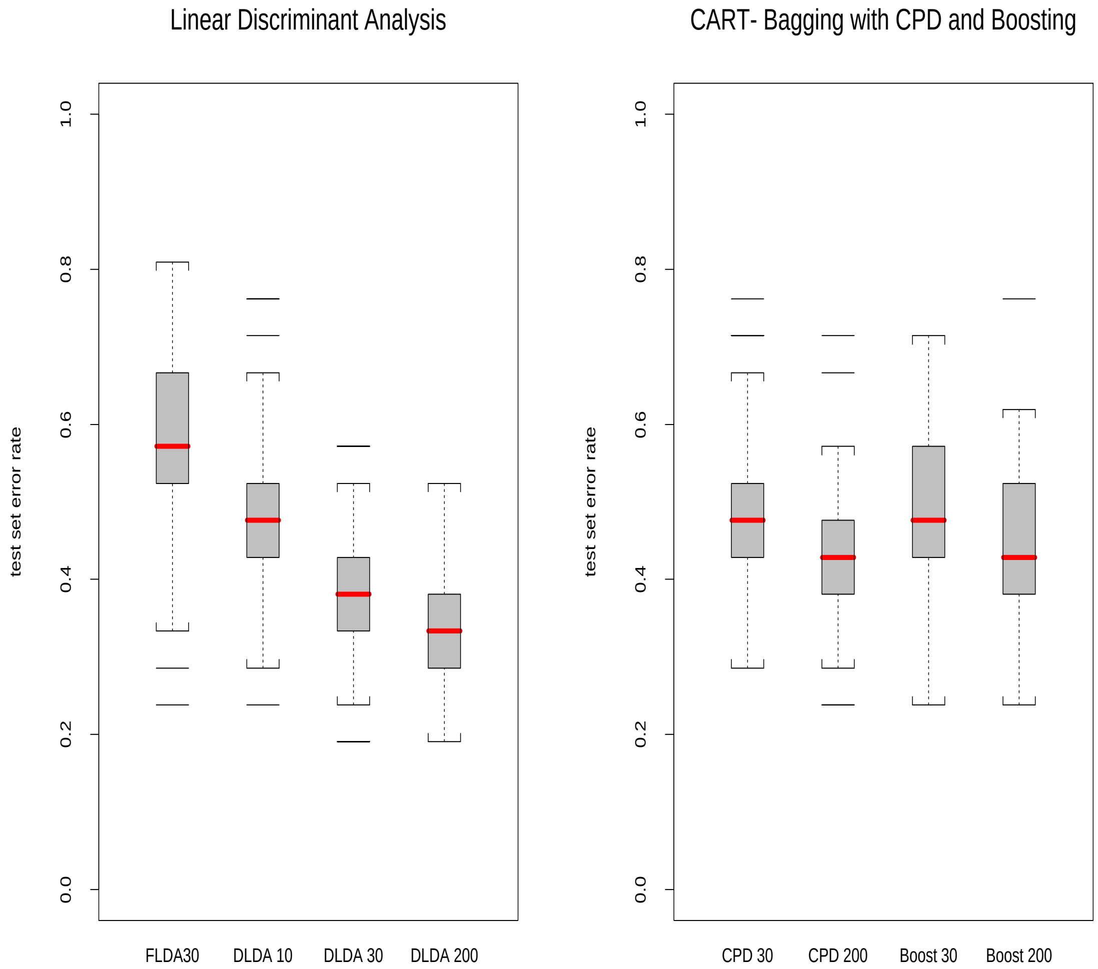


Figure 9: *NCI 60 data - Test set error rates.* Boxplots of test set error rates for discriminant analysis and CART-based classifiers based on $p = 10,\ 30,\ 200$ genes; $N = 150$ LS/TS runs for $2:1$ scheme.

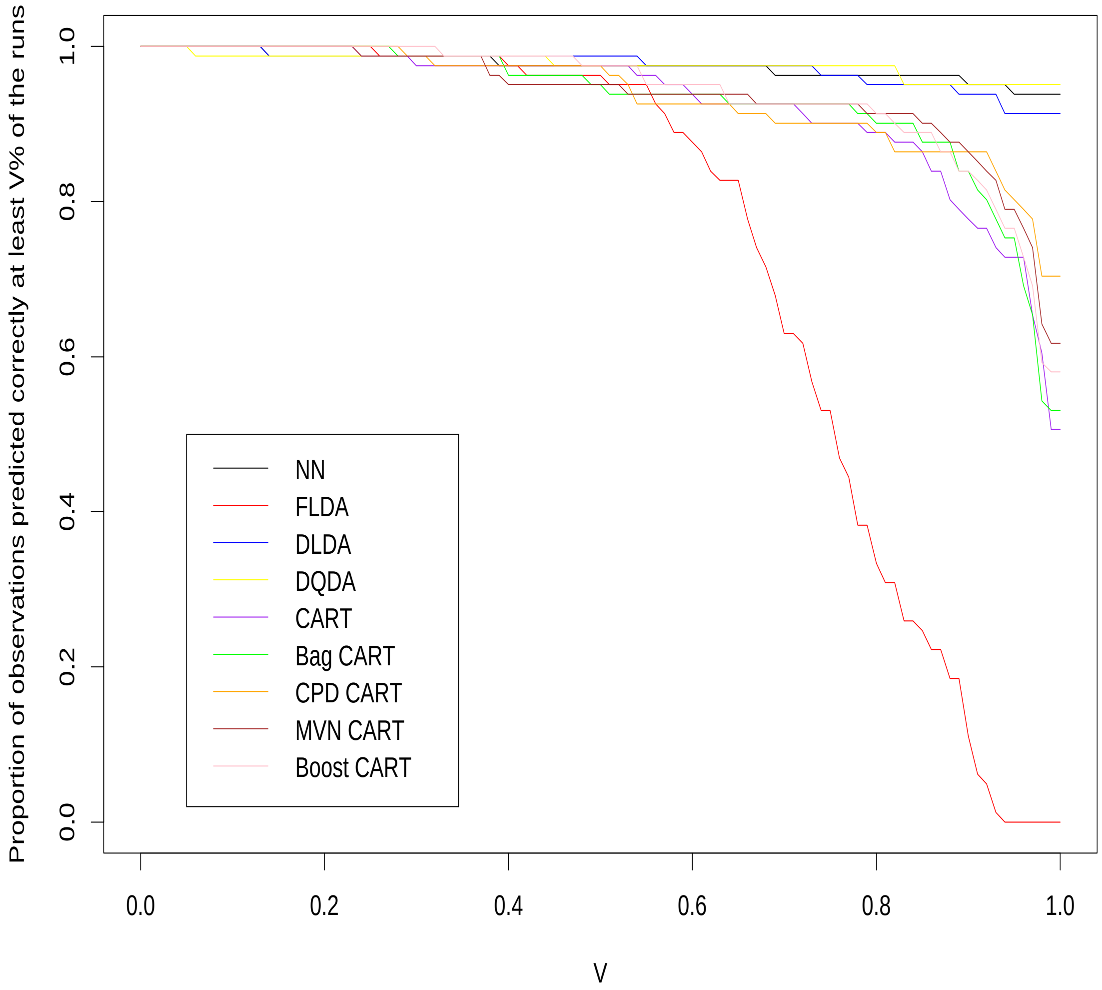


Figure 10: *Lymphoma data - observation-wise error rates.* Fraction of observations predicted correctly at least $V\%$ of the time (out of the runs for which a given observation belonged to the test set) for $p = 50$ genes; $N = 150$ LS/TS runs for $2:1$ sampling scheme.

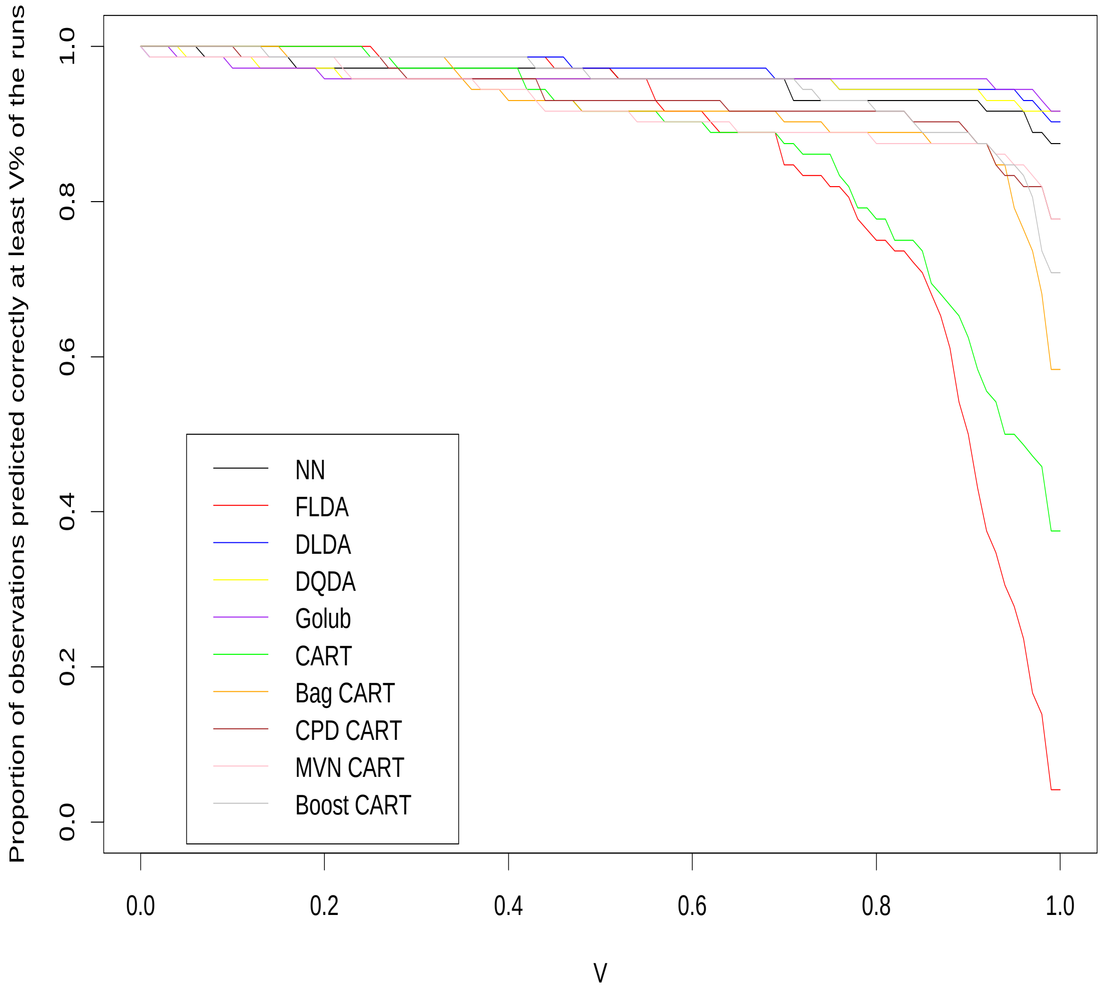


Figure 11: *Leukemia data, two classes - observation-wise error rates.* Fraction of observations predicted correctly at least $V\%$ of the time (out of the runs for which a given observation belonged to the test set) for $p = 40$ genes; $N = 150$ LS/TS runs for $2:1$ scheme.

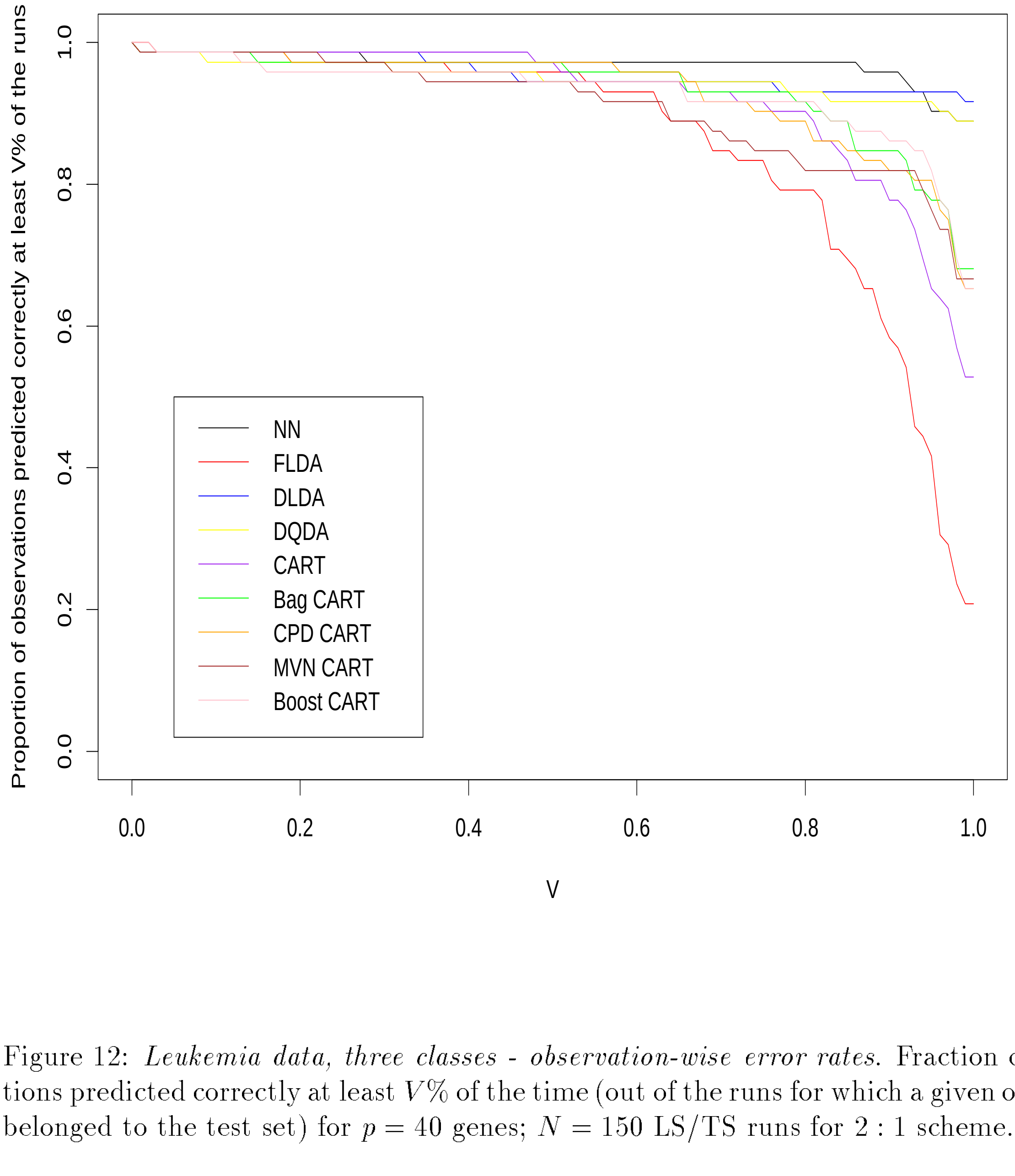


Figure 12: *Leukemia data, three classes - observation-wise error rates.* Fraction of observations predicted correctly at least $V\%$ of the time (out of the runs for which a given observation belonged to the test set) for $p = 40$ genes; $N = 150$ LS/TS runs for $2:1$ scheme.

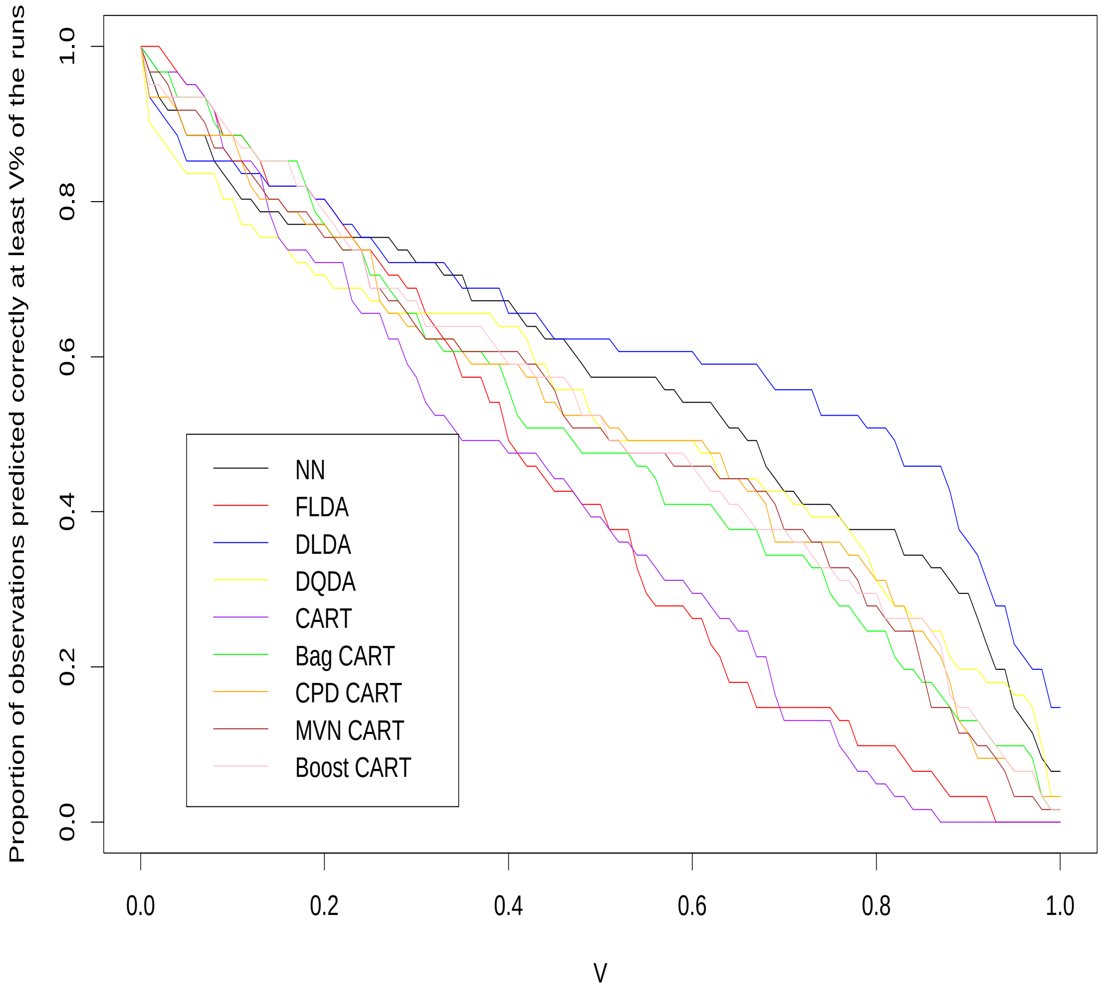


Figure 13: *NCI 60 data - observation-wise error rates.* Fraction of observations predicted correctly at least $V\%$ of the time (out of the runs for which a given observation belonged to the test set) for $p = 30$ genes; $N = 150$ LS/TS runs for $2:1$ scheme.

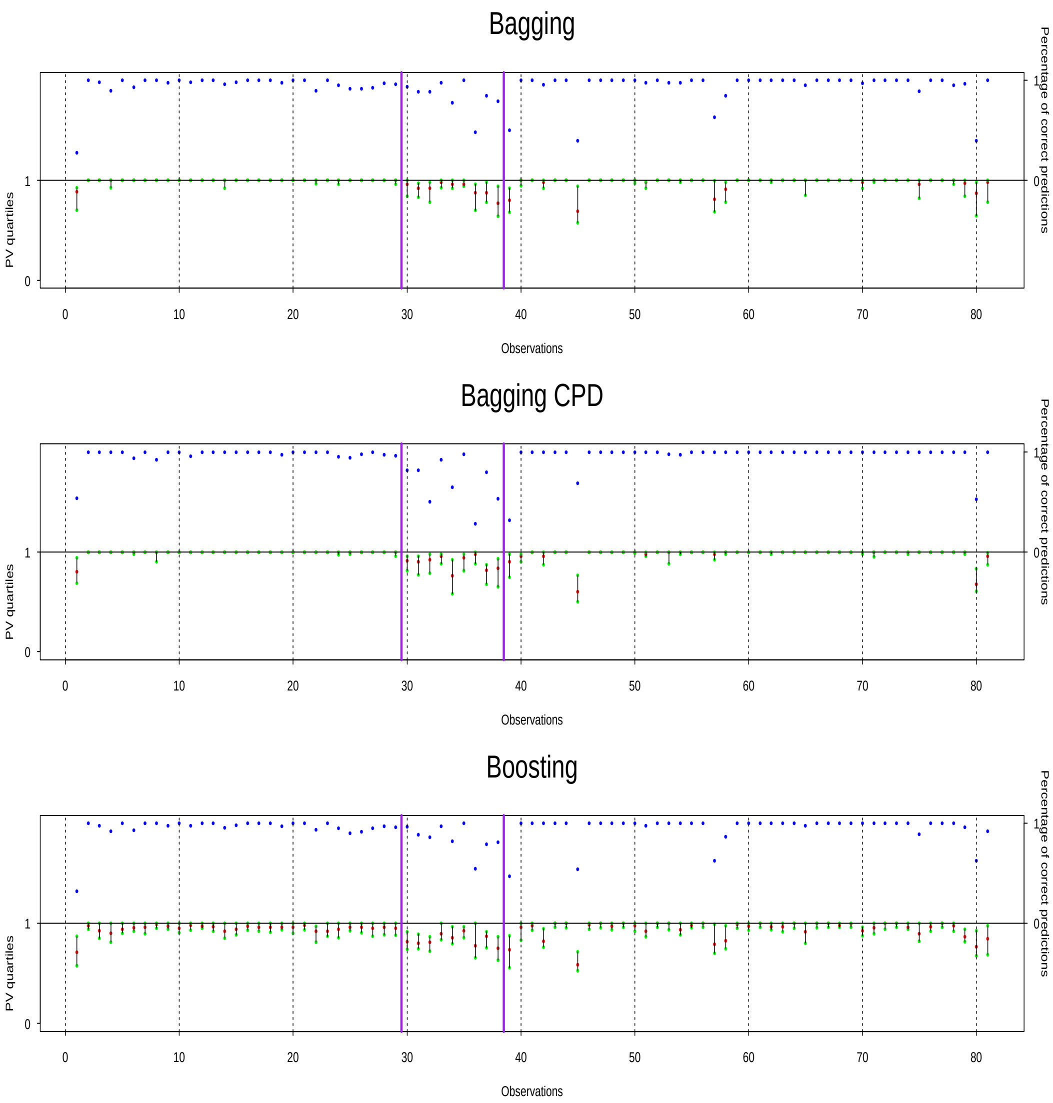


Figure 14: *Lymphoma data - Prediction votes.* Plots of percentage of correct predictions (top panel) and three number summaries (median, lower and upper quartile) of prediction votes (lower panel) for each observation. The observations are ordered by class: B-CLL, FL, DLBCL. Classifiers based on $p = 50$ genes; $N = 150$ LS/TS runs for $2:1$ scheme.

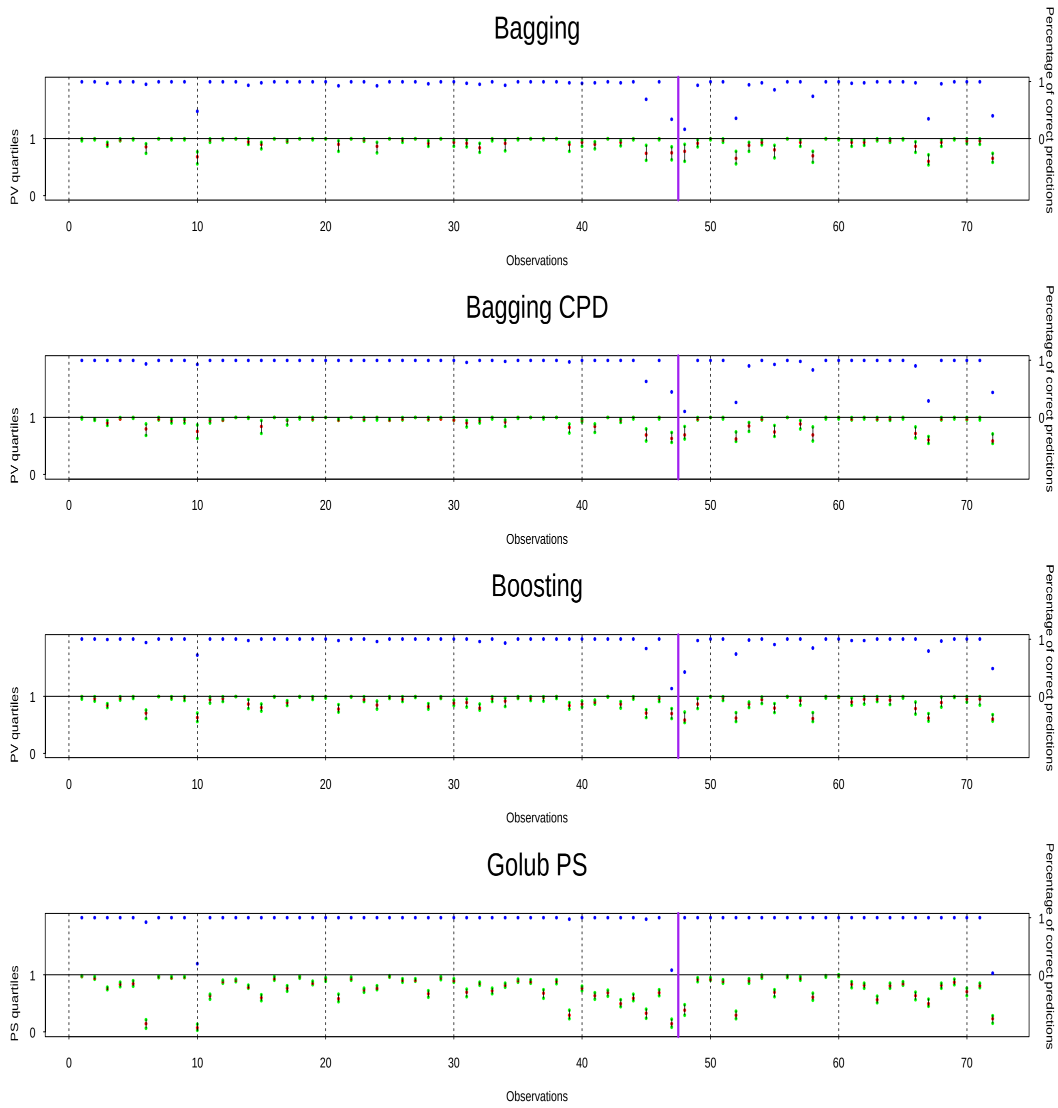


Figure 15: *Leukemia data, two classes - Prediction votes.* Plots of percentage of correct predictions (top panel) and three number summaries (median, lower and upper quartile) of prediction votes and prediction strengths (lower panel) for each observation. The observations are ordered by class: ALL followed by AML. Classifiers based on $p = 40$ genes; $N = 150$ LS/TS runs for $2:1$ scheme.

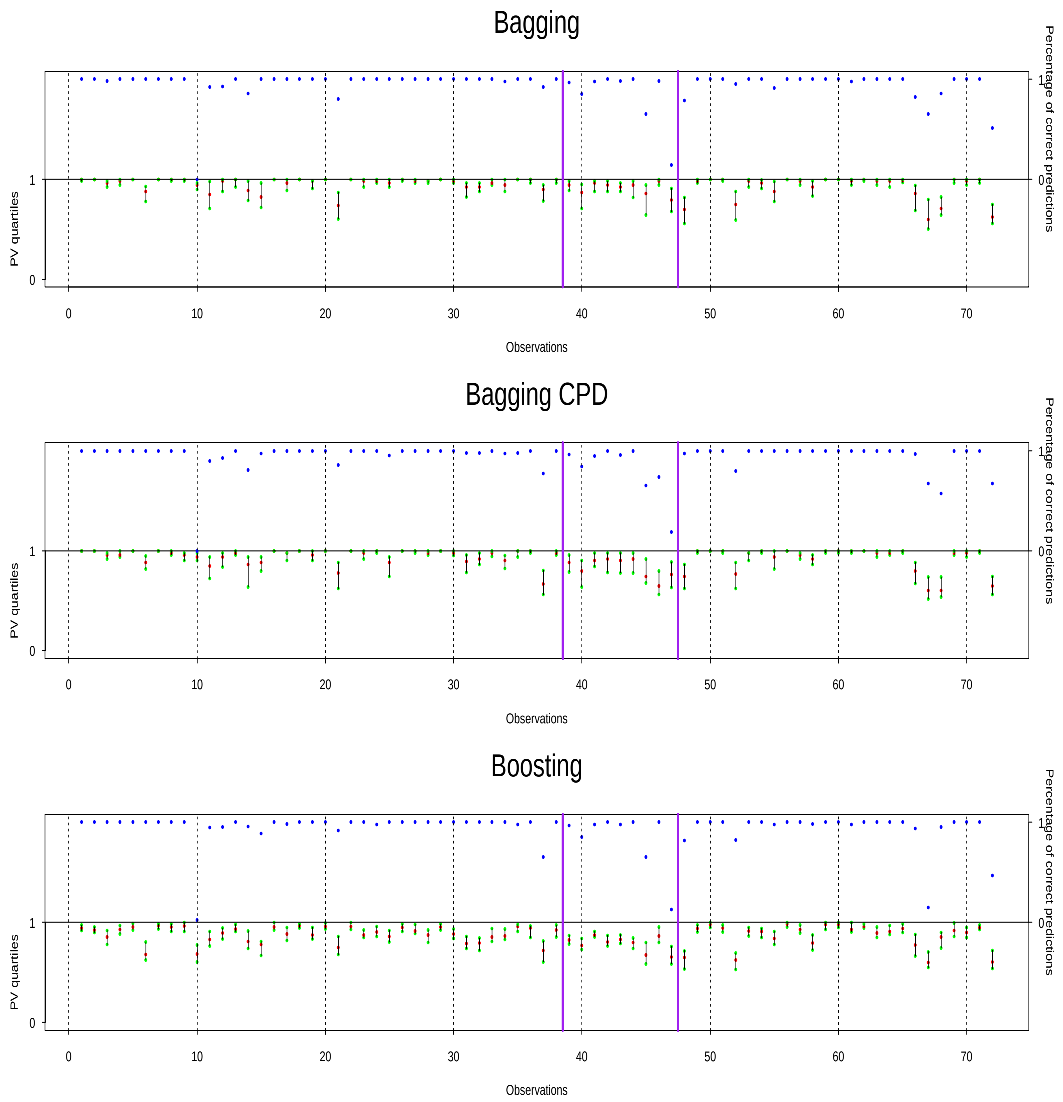


Figure 16: *Leukemia data, three classes - Prediction votes.* Plots of percentage of correct predictions (top panel) and three number summaries (median, lower and upper quartile) of prediction votes (lower panel) for each observation. The observations are ordered by class: ALL B-cell, ALL T-cell, AML. Classifiers based on $p = 40$ genes; $N = 150$ LS/TS runs for $2:1$ scheme.

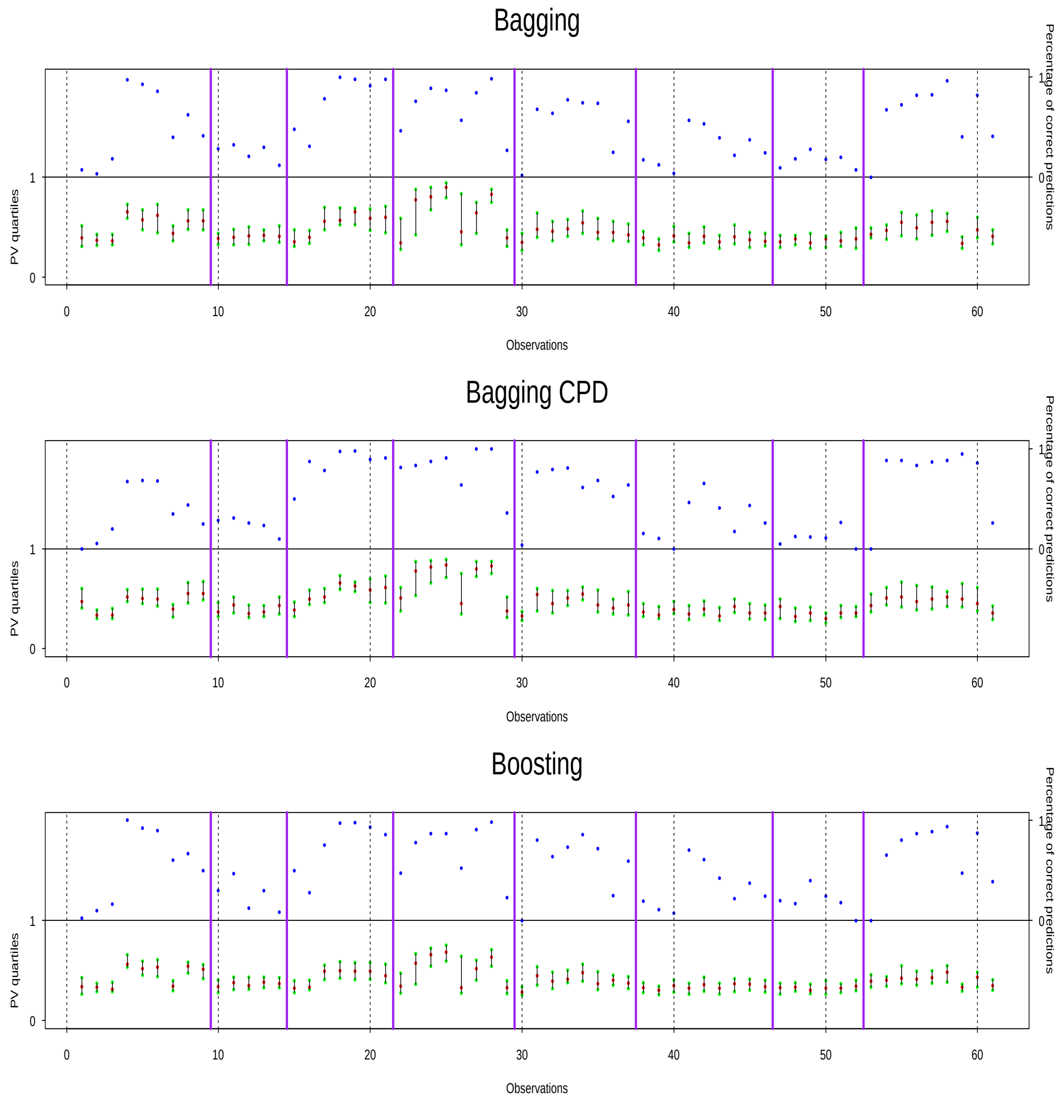
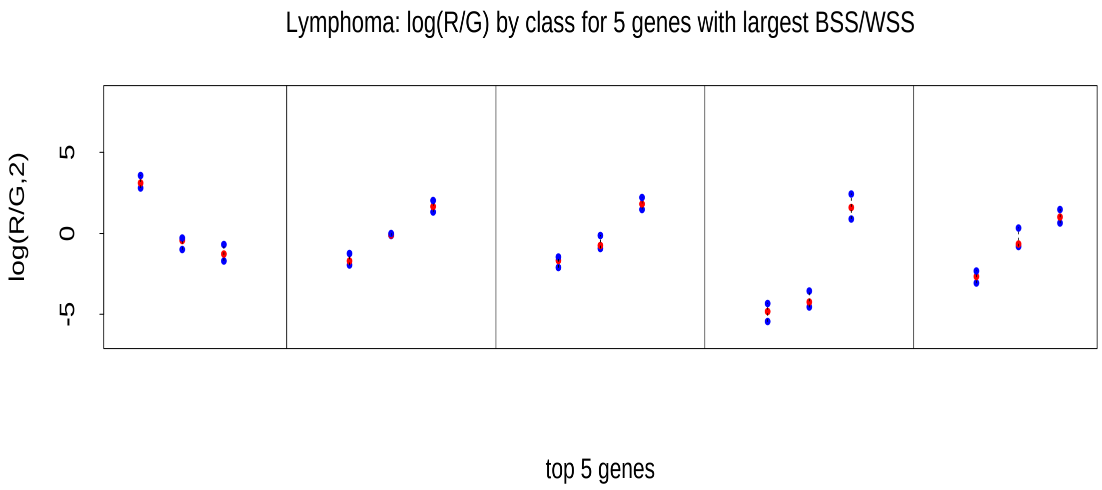


Figure 17: *NCI 60 data - Prediction votes.* Plots of percentage of correct predictions (top panel) and three number summaries (median, lower and upper quartile) of prediction votes (lower panel) for each observation. The observations are ordered by class: 7+2 breast, 5 CNS, 7 colon, 6+2 leukemia, 8 melanoma, 9 NSCLC, 6 ovarian, 9 renal. Classifiers based on $p = 30$ genes; $N = 150$ LS/TS runs for $2:1$ scheme.



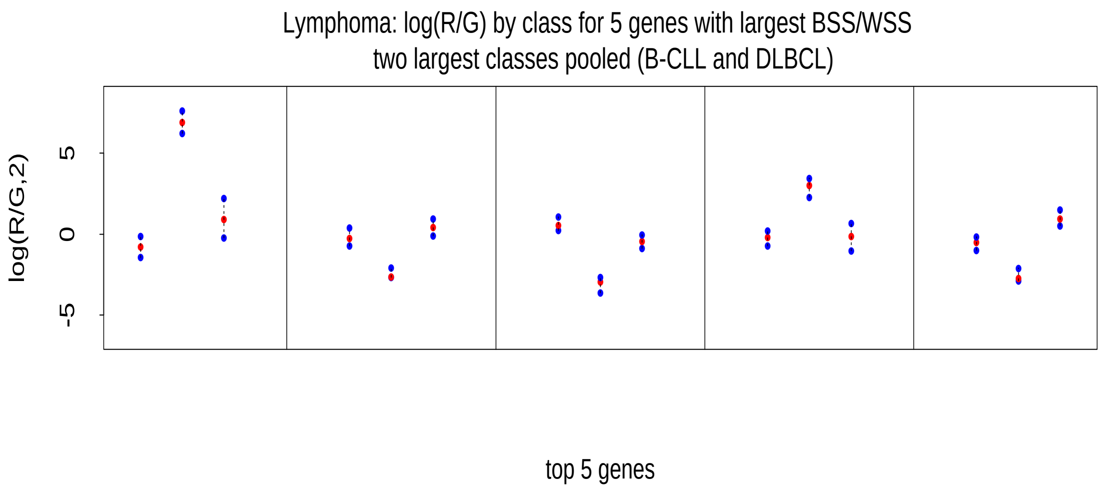


Figure 18: *Lymphoma data.* Median and upper and lower quartiles of gene expression levels (base 2 logarithm of the Cy5/Cy3 fluorescence ratio) within each of the three lymphoma classes. In the top figure, the genes are ordered according to their BSS/WSS ratio. These 5 genes successfully discriminate between the two largest classes (class 1 or B-CLL and class 3 or DLBCL) but not between the smallest class (class 2 or FL) and the two largest. In the bottom figure, the genes are ordered according to their BSS/WSS ratio with the largest two classes pooled. These 5 genes successfully differentiate between class 2 and the other two classes.