Hindawi Publishing Corporation
EURASIP Journal on Bioinformatics and Systems Biology
Volume 2009, Article ID 158368, 10 pages
doi:10.1155/2009/158368

Research Article

Is Bagging Effective in the Classification of Small-Sample Genomic and Proteomic Data?

T. T. Vu and U. M. Braga-Neto

Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77843-3128, USA

Correspondence should be addressed to U. M. Braga-Neto, ulisses@ece.tamu.edu

Received 1 August 2008; Revised 4 December 2008; Accepted 19 January 2009

Recommended by Yufei Huang

There has been considerable interest recently in the application of bagging in the classification of both gene-expression data and protein-abundance mass spectrometry data. The approach is often justified by the improvement it produces on the performance of unstable, overfitting classification rules under small-sample situations. However, the question of real practical interest is whether the ensemble scheme will improve performance of those classifiers sufficiently to beat the performance of single stable, nonoverfitting classifiers, in the case of small-sample genomic and proteomic data sets. To investigate that question, we conducted a detailed empirical study, using publicly-available data sets from published genomic and proteomic studies. We observed that, under t-test and RELIEF filter-based feature selection, bagging generally does a good job of improving the performance of unstable, overfitting classifiers, such as CART decision trees and neural networks, but that improvement was not sufficient to beat the performance of single stable, nonoverfitting classifiers, such as diagonal and plain linear discriminant analysis, or 3-nearest neighbors. Furthermore, as expected, the ensemble method did not improve the performance of these classifiers significantly. Representative experimental results are presented and discussed in this work.

1. Introduction

Randomized ensemble methods for classifier design combine the decision of an ensemble of classifiers designed on randomly perturbed versions of the available data [1–5]. The combination is often done by means of majority voting among the individual classifier decisions [4–6], whereas the data perturbation usually employs the bootstrap resampling approach, which corresponds to sampling uniformly with replacement from the original data [7, 8]. The combination of bootstrap resampling and majority voting is known as bootstrap aggregation or *bagging* [4, 5].

There has been considerable interest recently in the application of bagging in the classification of both gene-expression data [9–12] and protein-abundance mass spectrometry data [13–18]. However, there is scant theoretical justification for the use of this heuristic, other than the expectation that combining the decision of several classifiers will regularize and improve the performance of unstable overfitting classification rules, such as unpruned decision trees, provided one uses a large enough number of classifiers in the ensemble [4, 5]. It is also claimed that ensemble rules "do not overfit," meaning that classification error converges as the number of component classifiers tends to infinity [5].

However, the main performance issue is not whether the ensemble scheme improves the classification error of a single unstable overfitting classifier, or whether its classification error converges to a fixed limit; these are important questions, which have been studied in the literature (in particular when the component classifiers are decision trees) [5, 19–23], but the question of main practical interest is whether the ensemble scheme will improve the performance of unstable overfitting classifiers *sufficiently* to beat the performance of single stable, nonoverfitting classifiers, particularly in small-sample settings. Therefore, there is a pressing need to examine rigorously the suitability and validity of the ensemble approach in the classification of small-sample genomic and proteomic data. In this paper, we present results from a comprehensive empirical study concerning the effect of bagging on the performance of several classification rules, including diagonal and plain linear discriminant analysis, 3-nearest neighbors, CART decision trees, and neural networks, using real data from published microarray and mass spectrometry studies. Here we are concerned exclusively with the performance in terms of the true classification error, and therefore we employ filter-based feature selection and holdout estimation based on large samples in order to allow accurate classification error estimation. Similar studies recently published [11, 12] rely on small-sample wrapper feature selection and small-sample error estimation methods, which will obscure the issue of how bagging really affects the true classification error. In particular, there is evidence that filter-based feature selection outperforms wrapper feature selection in small-sample settings [24]. In our experiments, we employ the one-tailed paired t-test to assess whether the expected true classification error is significantly smaller for the bagged classifier as opposed to the original base classifier, under different number of samples, dimensionality, and number of classifiers in the ensemble. Clearly, the heuristic is beneficial for the particular classification rule if and only there is a significant decrease in expected classification error, otherwise the procedure is to be avoided; however the magnitude of improvement is also a factor—a small improvement in performance may not be worth the extra computation required (which is roughly m times larger for the bagging classifier, where m is the number of classifiers in the ensemble). The full results of the empirical study are available on a companion website http://www.ece.tamu.edu/~ulisses/bagging/index.html.

2. Randomized Ensemble Classification Rules

Classification involves a *feature vector* X in a feature space V, a *label* $Y \in \{0,1\}$, and a *classifier* $\psi : V \to \{0,1\}$, such that $\psi(x)$ attempts to predict the value of Y for a given observation $X = x$. The joint *feature-label distribution* F of the pair (X, Y) completely characterizes the stochastic properties of the classification problem. In practice, a classification rule is used to design a classifier based on sample training data. Working formally, a *classification rule* is a mapping $\Psi_n : [V \times \{0,1\}]^n \to \{0,1\}^V$, which takes an i.i.d. sample $S_n = \{(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\}$ of feature-label pairs drawn from the feature-label distribution to a *designed classifier* $\psi_n = \Psi_n(S_n)$. The *classification error* is the probability that classification is erroneous given the sample data, that is, $\varepsilon_n = P(\psi_n(X) \neq Y \mid S_n)$. Note that the classification error is random only through the training data S_n. The *expected classification error* $E[\varepsilon_n]$ is the average classification error over all possible sample data sets; it is a fixed parameter of the classification rule and feature-label distribution, and used as the measure of performance of the former given the latter.

Randomization approaches based on resampling can be seen as drawing i.i.d. samples $S_k^* = \{(X_1^*, Y_1^*), (X_2^*, Y_2^*), \ldots, (X_k^*, Y_k^*)\}$ from a surrogate joint-feature label distribution F^*, which is a function of the original training data S_n. In the bootstrap resampling approach, one has $k = n$, and the randomized sample S_n^* corresponds to sampling uniformly n training points from S_n *with* replacement. This corresponds to using the *empirical distribution* of the data S_n as the surrogate joint-feature label distribution F^*; the empirical distribution assigns discrete probability mass $1/n$ at each observed data point in S_n. Some of the original training points may appear multiple times, whereas others may not appear at all in the *bootstrap sample* S_n^*. Note that, given S_n, the bootstrap sample S_n^* is conditionally independent from the original feature-label distribution F.

In aggregation by majority voting, a classifier is obtained based on majority voting among individual classifiers designed on the randomized samples S_k^* using the original classification rule Ψ_n. This leads to an *ensemble classification rule* Ψ_n^R, such that

$$\psi_n^R(x) = \Psi_n^R(S_n)(x) = \begin{cases} 1, & E[\Psi_n(S_k^*)(x) \mid S_n] > \dfrac{1}{2}, \\ 0, & \text{otherwise}, \end{cases} \qquad (1)$$

for $x \in V$, where expectation is with respect to the random mechanism F^*, fixed at the observed value of S_n. For bootstrap majority voting, or bagging, the expectation in (1) usually has to be approximated by Monte Carlo sampling, which leads to the "bagged" classifier:

$$\psi_{n,m}^B(x) = \begin{cases} 1, & \dfrac{1}{m}\displaystyle\sum_{j=1}^{m} \psi_n^{*(j)}(x) > \dfrac{1}{2}, \\ 0, & \text{otherwise}, \end{cases} \qquad (2)$$

where the classifiers $\psi_n^{*(j)}$ are designed by the original classification rule Ψ_n on bootstrap samples $S_n^{*(j)}$, for $j = 1, \ldots, m$, for large enough m (notice the parallel with the development in [25], particulary equations (2.8)–(2.10), and accompanying discussion).

The issue of how large m has to be so that (2) is a good Monte Carlo approximation is a critical issue in the application of bagging. Note that m represents the number of classifiers that must be designed to be part of the ensemble, so that a computational problem may emerge if m is made too large. In addition, even if a suitable m is found, the performance of the ensemble must be compared to that of the base classification rule, to see if there is significant improvement. Even more importantly, the performance of the ensemble has to be compared to that of other classification rules; that the ensemble improves the performance of an unstable overfitting classifier is of small value if it can be bested by a single stable, nonoverfitting classifier. In the next section, we present a comprehensive empirical study that addresses these questions.

3. Experimental Study

In this section, we report the results obtained from a large simulation study based on publicly-available patient data from genomic and proteomic studies, which measured the performance of the bagging heuristic through the expected classification error, for varying number of component classifiers, sample size, and dimensionality.

3.1. Methods. We considered in our experiment several classification rules, listed here in order of complexity: diagonal linear discriminant analysis (DLDA), linear discriminant analysis (LDA), 3-nearest neighbors (3NN), decision trees (CART), and neural networks (NNET) [26, 27]. DLDA is an extension of LDA where only the diagonal elements (the variances) of the covariance matrix are estimated, while the off-diagonal elements (the covariances) are assumed to be zero. Bagging is applied to each of these base classification rules and its performance recorded for varying number of individual classifiers. The neural network consists of a one-hidden layer with 4 nodes and standard sigmoids as nonlinearities. The network is trained by Levenberg-Marquardt optimization with a maximum of 30 iterations. CART is applied with a stopping criterion. Splitting is stopped when there are fewer than 3 points in a given node. This is distinct from the approach advocated in [5] for random forests, where unpruned, fully grown trees are used instead; the reason for this is that we did not attempt to implement the approach in [5] (which involves concepts as random node splitting and is thus specific to decision trees), but rather to study the behavior of bagging, which is the centerpiece of such ensemble methods, across different classification rules. Resampling is done by means of *balanced* bootstrapping, where all samples are made to appear exactly the same number of times in the computation [28].

We selected data sets with large number N of samples (see below) in order to be able to estimate the true error accurately using held out testing data. In each case, 1000 training data sets of size $n = 20$, 40, and 60 were drawn uniformly and independently from the total pool of N samples. The training data are drawn in a stratified fashion, following the approximate proportion of each class in the original data. Based on the training data, a filter-based gene selection step is employed to select the top p discriminating genes; we considered in this study $p = 2, 3, 5, 8$. The univariate feature selection methods used in the filter step are the Welch two-sample t-test [29] and the RELIEF method [30]—in the latter case, we employ the 1-nearest neighbor method when searching for hits and misses. After classifier design, the true classification error for each data set of size n is approximated by a holdout estimator, whereby the $N - n$ sample points not drawn are used as the test set (a good approximation to the classification error, given that $N \gg n$). The expected classification error is then estimated as the sample mean of classification error over the 1000 training data sets. The sample size n is kept small, as we are interested in the small-sample properties of bagging. Note also that we also must have $N \gg n$ in order to provide for large enough testing sets, as well as to make sure that consecutive training sets do not significantly overlap, so that the expected classification error can be accurately approximated. As can be easily verified, the expected ratio of overlapping sample points between two samples of size n from a population of size N is given simply by n/N. In all cases considered here the expected overlap is around 20% less, which we consider to be acceptable, except in the case of the lung cancer data set with $n = 60$. This latter case is therefore not included in our results. The one-tailed paired t-test is employed to assess whether the ensemble classifier has an expected error that is significantly smaller than that of the corresponding individual classifier.

3.2. Data Sets. We utilized the following publicly-available data sets from published studies in order to study the performance of bagging in the context of genomics and proteomics applications.

3.2.1. Breast Cancer Gene Expression Data. These data come from the breast cancer classification study in [31], which analyzed $N = 295$ gene-expression microarrays containing a total of 25760 transcripts each. Filter-based feature selection was performed on a 70-gene prognosis profile, previously published by the same authors in [32]. Classification is between the good-prognosis class (115 samples), and the poor-prognosis class (180 samples), where prognosis is determined retrospectively in terms of survivability [31].

3.2.2. Lung Cancer Gene Expression Data. We employed here the data set "A" from the study in [33] on nonsmall-cell lung carcinomas (NSCLC), which analyzed $N = 186$ gene-expression microarrays containing a total of 12600 transcripts each. NSCLC is subclassified as adenocarcinomas, squamous cell carcinomas and large-cell carcinomas, of which adenocarcinomas are the most common subtypes and of interest to classify from other subtypes of NSCLC. Classification is thus between adenocarcinomas (139 samples) and non-adenocarcinomas (47 samples).

3.2.3. Prostate Cancer Protein Abundance Data. Given the recent keen interest on deriving serum-based proteomic biomarkers for the diagnosis of cancer [34], we also included in this study data from a proteomic study of prostate cancer reported in [35]. It consists of SELDI-TOF mass spectrometry of $N = 326$ samples, which yield mass spectra for 45000 m/z (mass over charge) values. Filter-based feature selection is employed to find the top discriminatory m/z values to be used in the experiment. Classification is between prostate cancer patients (167 samples) and noncancer patients, including benign prostatic hyperplasia and healthy patients (159 samples). We use the raw spectra values, without baseline subtraction, as we found that this leads to better classification rates.

3.3. Results and Discussion. We present results for sample sizes $n = 20$ and $n = 40$ and dimensionality $p = 2$ and $p = 5$, which are representative of the full set of results, available on the companion website http://www.ece.tamu.edu/~ulisses/bagging/index.html. The case $p = 2$ is displayed in Tables 1, 2, and 3, each of which corresponds to a different data set. Each table displays the expected classification error as a function of the number m of classifiers used in the ensemble, for different base classification rules, feature selection methods, and sample sizes. We used in all cases an odd number m of classifiers in the ensembles, to avoid tie-breaking issues. Errors that are smaller for the ensemble classifier as

Table 1: Expected classification error of selected experiments for breast cancer gene-expression data under two different features selection methods (t-test and RELIEF) for $p = 2$. Bold-face type indicates the values that are smaller for the ensemble classifier as compared to a single classifier at a 99% significance level, according to the one-tailed paired t-test.

Rule	FS	n	Single	$m = 5$	$m = 11$	$m = 15$	$m = 21$	$m = 25$	$m = 31$	$m = 35$	$m = 41$	$m = 45$	$m = 51$
DLDA	t-test	20	0.202	0.215	0.208	0.206	0.204	0.204	0.204	0.204	0.203	0.204	0.203
DLDA	t-test	40	0.198	0.205	0.201	0.200	0.200	0.199	0.199	0.199	0.199	0.199	0.199
DLDA	RELIEF	20	0.202	0.215	0.207	0.206	0.204	0.204	0.204	0.203	0.203	0.203	0.203
DLDA	RELIEF	40	0.198	0.206	0.201	0.201	0.200	0.200	0.199	0.199	0.199	0.199	0.199
LDA	t-test	20	0.212	0.237	0.224	0.220	0.217	0.217	0.216	0.216	0.215	0.215	0.214
LDA	t-test	40	0.204	0.217	0.209	0.208	0.207	0.206	0.206	0.206	0.205	0.205	0.205
LDA	RELIEF	20	0.213	0.239	0.225	0.222	0.219	0.218	0.218	0.217	0.216	0.216	0.216
LDA	RELIEF	40	0.203	0.218	0.210	0.207	0.206	0.206	0.205	0.205	0.205	0.205	0.205
3NN	t-test	20	0.230	0.281	0.246	0.241	0.235	0.234	0.231	0.231	0.230	0.229	**0.229**
3NN	t-test	40	0.228	0.274	0.241	0.235	0.231	0.229	0.228	0.227	**0.226**	**0.226**	**0.225**
3NN	RELIEF	20	0.234	0.282	0.248	0.242	0.238	0.236	0.234	0.234	0.233	**0.233**	**0.232**
3NN	RELIEF	40	0.227	0.271	0.241	0.235	0.231	0.229	0.227	0.227	**0.226**	**0.225**	**0.225**
CART	t-test	20	0.259	0.297	0.263	0.256	**0.250**	**0.247**	**0.246**	**0.244**	**0.243**	**0.242**	**0.242**
CART	t-test	40	0.257	0.294	0.258	**0.252**	**0.245**	**0.244**	**0.242**	**0.240**	**0.239**	**0.239**	**0.237**
CART	RELIEF	20	0.263	0.299	0.265	**0.258**	**0.253**	**0.250**	**0.247**	**0.247**	**0.245**	**0.245**	**0.244**
CART	RELIEF	40	0.256	0.293	0.260	**0.253**	**0.245**	**0.244**	**0.241**	**0.240**	**0.239**	**0.239**	**0.238**
NNET	t-test	20	0.252	0.293	**0.246**	**0.240**	**0.230**	**0.230**	**0.225**	**0.224**	**0.223**	**0.222**	**0.221**
NNET	t-test	40	0.226	0.256	0.225	**0.219**	**0.215**	**0.213**	**0.212**	**0.210**	**0.210**	**0.209**	**0.209**
NNET	RELIEF	20	0.255	0.298	**0.248**	**0.240**	**0.233**	**0.232**	**0.229**	**0.228**	**0.226**	**0.225**	**0.224**
NNET	RELIEF	40	0.230	0.260	0.227	**0.220**	**0.216**	**0.213**	**0.213**	**0.212**	**0.211**	**0.210**	**0.209**

Table 2: Expected classification error of selected experiments for lung cancer gene-expression data under two different features selection methods (t-test and RELIEF) for $p = 2$. Bold-face type indicates the values that are smaller for the ensemble classifier as compared to a single classifier at a 99% significance level, according to the one-tailed paired t-test.

Rule	FS	n	Single	$m = 5$	$m = 11$	$m = 15$	$m = 21$	$m = 25$	$m = 31$	$m = 35$	$m = 41$	$m = 45$	$m = 51$
DLDA	t-test	20	0.190	0.191	0.190	0.190	**0.189**	**0.190**	**0.189**	**0.189**	0.190	**0.190**	**0.190**
DLDA	t-test	40	0.186	0.187	0.186	0.186	**0.186**	0.186	0.186	**0.186**	**0.186**	0.186	0.186
DLDA	RELIEF	20	0.235	0.253	0.238	0.239	0.235	0.236	**0.233**	0.233	0.234	**0.232**	**0.233**
DLDA	RELIEF	40	0.207	0.212	0.209	0.208	0.207	0.207	0.207	0.207	0.207	0.207	0.206
LDA	t-test	20	0.201	0.206	0.203	0.203	0.202	0.202	0.203	0.202	0.202	0.202	0.203
LDA	t-test	40	0.192	0.194	0.193	0.193	0.193	0.193	0.192	0.192	0.193	0.192	0.192
LDA	RELIEF	20	0.262	0.295	0.274	0.271	0.265	0.265	0.263	0.263	0.260	0.261	0.261
LDA	RELIEF	40	0.208	0.223	0.214	0.213	0.212	0.212	0.210	0.211	0.210	0.210	0.208
3NN	t-test	20	0.122	0.151	0.130	0.126	0.124	0.123	0.122	0.121	0.121	0.121	**0.120**
3NN	t-test	40	0.123	0.147	0.129	0.127	0.125	0.124	0.123	0.123	**0.122**	**0.122**	**0.121**
3NN	RELIEF	20	0.247	0.334	0.265	0.258	0.249	0.248	0.246	0.247	**0.244**	**0.244**	**0.243**
3NN	RELIEF	40	0.232	0.317	0.252	0.243	0.238	0.235	0.234	0.233	0.232	**0.231**	**0.230**
CART	t-test	20	0.160	0.182	0.161	**0.155**	**0.152**	**0.151**	**0.150**	**0.149**	**0.148**	**0.148**	**0.147**
CART	t-test	40	0.156	0.177	0.155	**0.150**	**0.146**	**0.145**	**0.144**	**0.143**	**0.142**	**0.142**	**0.142**
CART	RELIEF	20	0.297	0.302	**0.280**	**0.274**	**0.269**	**0.267**	**0.266**	**0.264**	**0.263**	**0.262**	**0.263**
CART	RELIEF	40	0.297	0.297	**0.273**	**0.268**	**0.263**	**0.261**	**0.260**	**0.258**	**0.257**	**0.257**	**0.256**
NNET	t-test	20	0.216	0.244	0.235	0.232	0.231	0.229	0.228	0.228	0.227	0.227	0.226
NNET	t-test	40	0.195	0.232	0.215	0.212	0.208	0.207	0.205	0.204	0.203	0.202	0.202
NNET	RELIEF	20	0.239	0.257	0.247	0.247	0.244	0.242	0.242	0.241	0.242	0.242	0.241
NNET	RELIEF	40	0.231	0.252	0.242	0.241	0.238	0.236	0.235	0.234	0.234	0.235	0.233

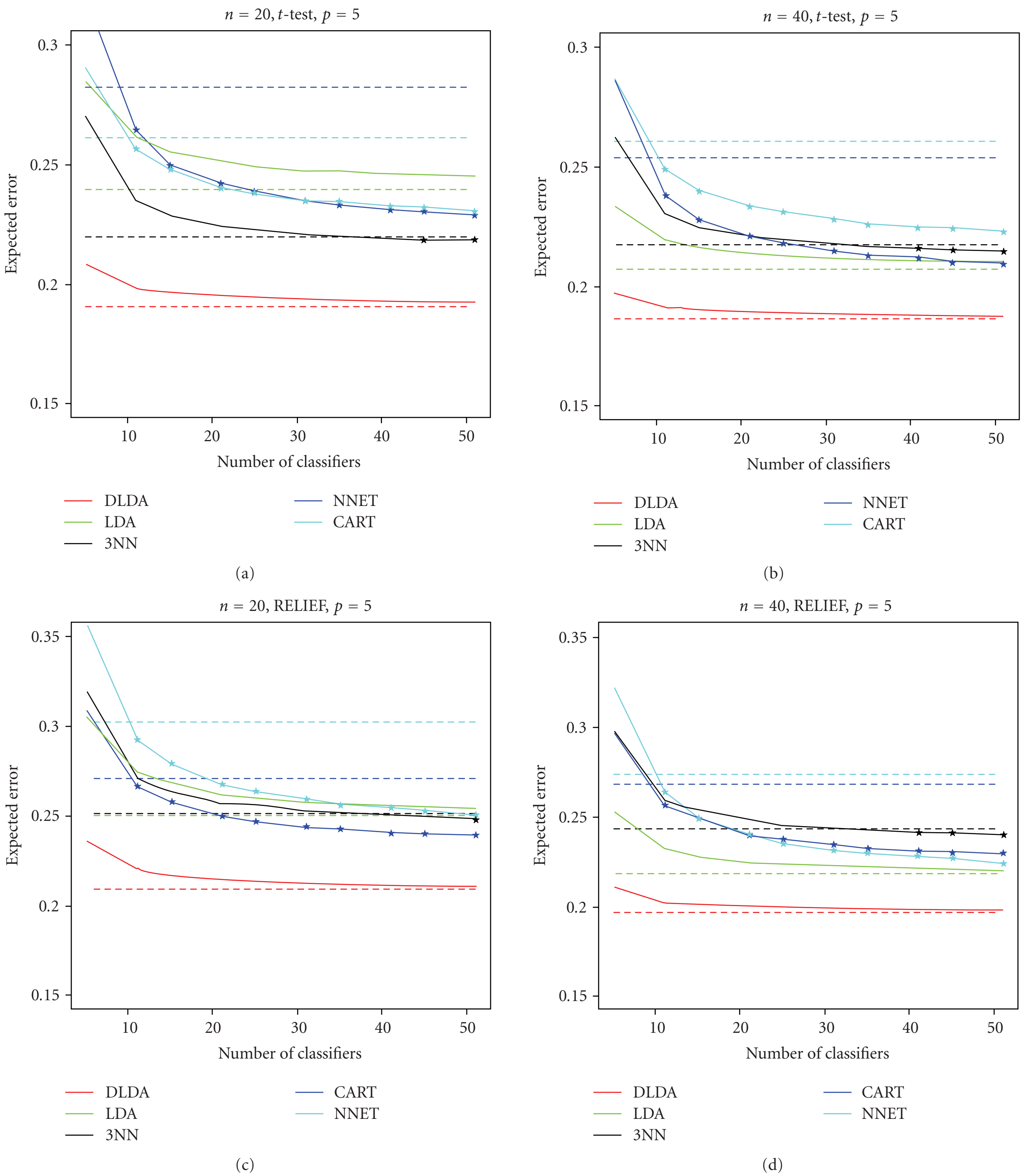
Table 3: Expected classification error of selected experiments for prostate cancer protein-abundance data under two different features selection methods (t-test and RELIEF) for $p = 2$. Bold-face type indicates the values that are smaller for the ensemble classifier as compared to a single classifier at a 99% significance level, according to the one-tailed paired t-test.

Rule	FS	n	Single	$m = 5$	$m = 11$	$m = 15$	$m = 21$	$m = 25$	$m = 31$	$m = 35$	$m = 41$	$m = 45$	$m = 51$
DLDA	t-test	20	0.188	0.211	0.199	0.196	0.194	0.194	0.193	0.192	0.191	0.191	0.191
DLDA	t-test	40	0.187	0.207	0.196	0.194	0.192	0.191	0.191	0.191	0.190	0.190	0.189
DLDA	RELIEF	20	0.468	0.523	0.492	0.484	0.477	0.475	0.471	0.472	0.469	0.467	0.466
DLDA	RELIEF	40	0.458	0.502	0.477	0.474	0.465	0.469	0.465	0.463	0.462	0.463	0.460
LDA	t-test	20	0.212	0.241	0.225	0.222	0.219	0.218	0.216	0.216	0.215	0.215	0.215
LDA	t-test	40	0.198	0.224	0.210	0.208	0.205	0.204	0.203	0.202	0.202	0.202	0.201
LDA	RELIEF	20	0.422	0.492	0.449	0.435	0.426	0.426	0.422	0.419	0.417	**0.415**	**0.410**
LDA	RELIEF	40	0.416	0.479	0.440	0.433	0.426	0.421	0.420	0.418	0.416	0.415	0.413
3NN	t-test	20	0.187	0.251	0.203	0.195	0.192	0.189	0.187	0.187	0.186	**0.185**	**0.185**
3NN	t-test	40	0.153	0.208	0.168	0.162	0.158	0.156	0.154	0.153	**0.152**	**0.152**	**0.151**
3NN	RELIEF	20	0.268	0.355	0.307	0.299	0.287	0.284	0.280	0.278	0.277	0.276	0.275
3NN	RELIEF	40	0.222	0.283	0.248	0.239	0.233	0.231	0.229	0.228	0.226	0.226	0.224
CART	t-test	20	0.232	0.247	**0.223**	**0.218**	**0.213**	**0.210**	**0.209**	**0.209**	**0.208**	**0.209**	**0.208**
CART	t-test	40	0.213	0.219	**0.198**	**0.194**	**0.189**	**0.189**	**0.187**	**0.185**	**0.185**	**0.185**	**0.184**
CART	RELIEF	20	0.244	0.284	0.259	0.256	0.251	0.249	0.247	0.245	0.244	0.244	0.243
CART	RELIEF	40	0.222	0.250	0.233	0.229	0.226	0.225	0.224	0.223	0.223	0.223	0.221
NNET	t-test	20	0.297	0.300	**0.271**	**0.266**	**0.260**	**0.259**	**0.256**	**0.256**	**0.254**	**0.254**	**0.253**
NNET	t-test	40	0.277	0.274	**0.254**	**0.248**	**0.244**	**0.244**	**0.240**	**0.241**	**0.239**	**0.239**	**0.239**
NNET	RELIEF	20	0.345	0.382	**0.337**	**0.324**	**0.318**	**0.314**	**0.313**	**0.313**	**0.312**	**0.309**	**0.307**
NNET	RELIEF	40	0.329	0.348	**0.312**	**0.303**	**0.295**	**0.294**	**0.290**	**0.290**	**0.290**	**0.288**	**0.289**

compared to a single classifier at a 99% significance level, according to the one-tailed paired t-test, are indicated by bold-face type. This allows one to immediately observe that bagging is able to improve the performance of the unstable overfitting CART and NNET classifiers; in most cases, a small ensemble is required, and the improvement in performance is substantial. In contrast, bagging does not improve the performance of the stable, nonoverfitting DLDA, LDA, and 3NN classifiers, except via a large ensemble; and even so the improvement in magnitude is quite small, and certainly does not justify the extra computational cost (note that in the case of the simplest classification rule, DLDA, there is no improvement at all). This is in agreement with what is known about the ensemble approach (e.g., see [5]).

However, of larger interest here is the performance of the ensemble against a single instance of the stable, nonoverfitting classifiers. This can be better visualized in the plots of Figures 1, 2, and 3, which display the expected classification errors as a function of number of component classifiers in the ensemble, for the case $p = 5$. The error of a single classifier is indicated by a horizontal dashed line. Marks indicate the values that are smaller for the ensemble classifier as compared to a single component classifier at a 99% significance level, according to the one-tailed paired t-test. One observes that as ensemble size increases, classification error decreases and tends to converge to a fixed value (in agreement with [5]), but we can also see that the error is usually larger at very small ensemble sizes, as compared to the error of the individual classifier. We can again observe that, in most cases, bagging is able to improve the performance of CART and NNET, but that is not significantly so, or at all, for DLDA, LDA, and 3NN. More importantly, we can see that the improvement on the performance of CART and NNET is not sufficient to beat the performance of single DLDA, LDA, or 3NN classifiers (with the exception of the prostate cancer data with RELIEF feature selection, which we comment on below).

As we can see in Figures 1–3, the breast cancer gene-expression data produces linear features that favor single DLDA and LDA classifiers (the latter do not perform so well at $n = 20$, due to the difficulty of estimating the entire covariance matrix at this sample size, which affects DLDA less), while the lung cancer gene-expression data produce nonlinear features, in which case, according to the results, the best option overall is to use a single 3NN classifier, followed closely by a bagged NNET in t-test feature selection and a bagged CART in RELIEF feature selection. The case of the prostate cancer proteomic data is peculiar in that it presents the only case where the best option was not a DLDA, LDA, or 3NN classifier, but in fact a single CART classifier, namely, the case $n = 20$ (with either $p = 2$ or $p = 5$) for RELIEF feature selection (the results for t-test feature selection, on the other hand, are very similar to the ones obtained for the lung cancer data set). Note that, in this case, the best performance is achieved by a single CART classifier, rather than the ensemble CART scheme. We also point out that the classification errors obtained with t-test feature selection are smaller than the ones obtained with RELIEF feature



(a)

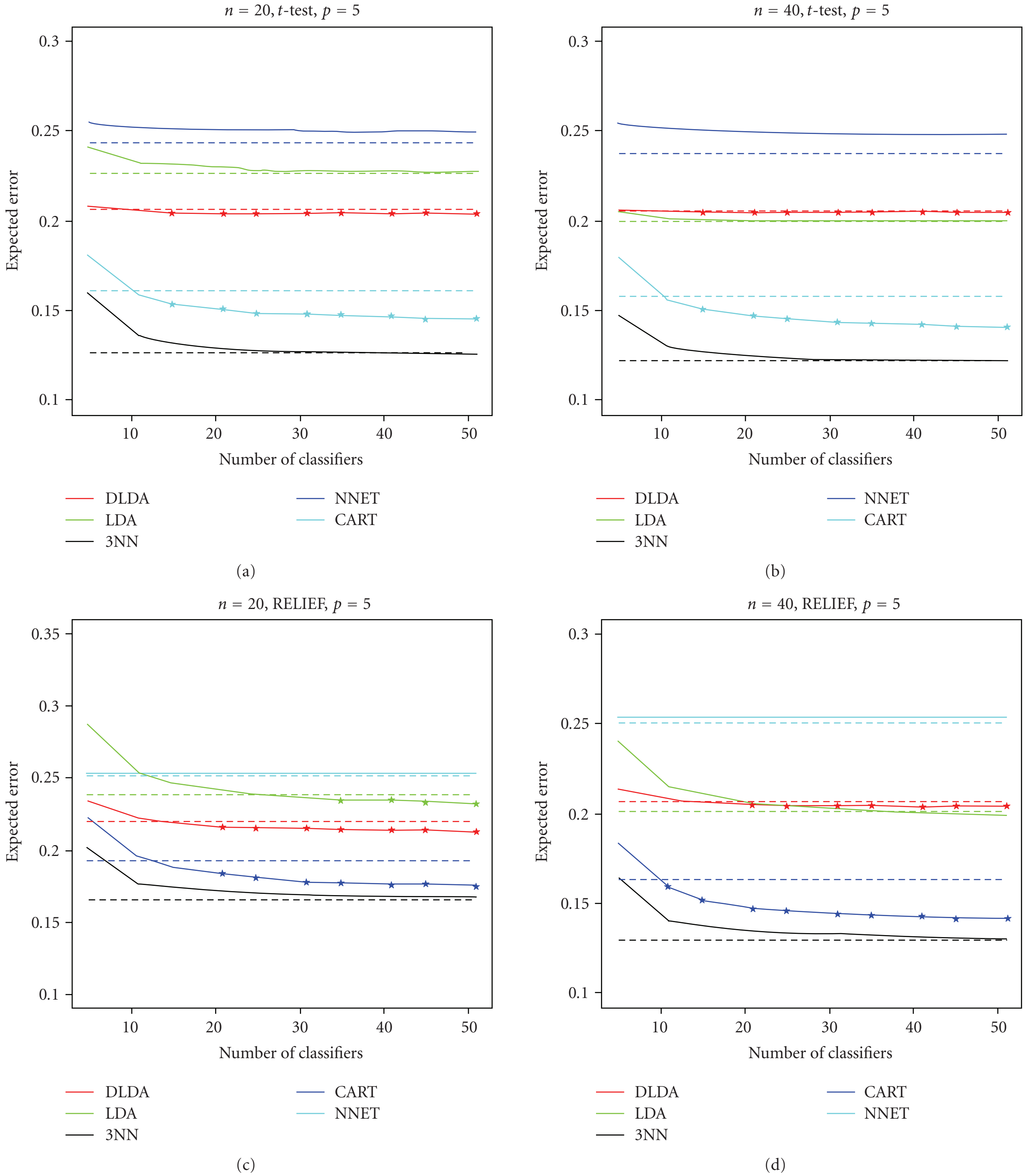
(b)

(c)

(d)

Figure 1: Expected classification error as a function of number of classifiers in the ensemble for selected experiments with the breast cancer gene expression data (full results available on the companion website). The error of a single classifier is indicated by a horizontal dashed line. Marks indicate the values that are smaller for the ensemble classifier as compared to a single classifier at a 99% significance level, according to the one-tailed paired t-test.

selection, indicating that RELIEF is not a good option in this case due to the very small-sample size (in fact, there is evidence that t-test filter-based feature selection may be the method of choice in small-sample cases [24]), in the case $n = 40$, the difference between 3NN and CART essentially disappears. It is also interesting that in the case $n = 20$ and $p = 5$, for RELIEF feature selection, bagging is able to improve the performance of LDA by a good margin in the



(a)

(b)

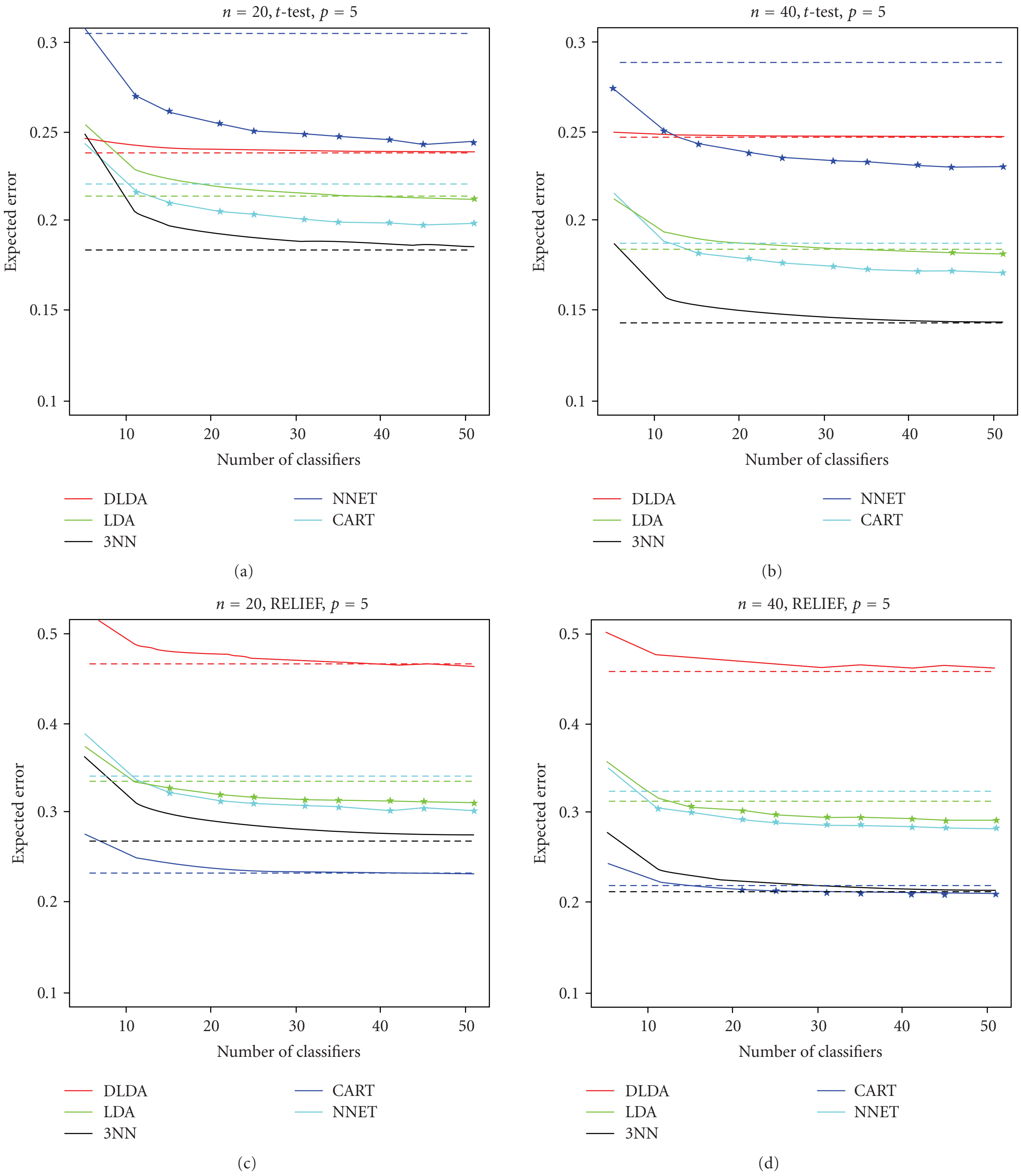
(c)

(d)

Figure 2: Expected classification error as a function of number of classifiers in the ensemble for selected experiments with the lung cancer gene expression data (full results available on the companion website). The error of a single classifier is indicated by a horizontal dashed line. Marks indicate the values that are smaller for the ensemble classifier as compared to a single classifier at a 99% significance level, according to the one-tailed paired t-test.

case of the prostate cancer data. This is due to the fact that the combination of LDA and RELIEF feature selection produce an unstable overfitting classification rule at this acute small-sample scenario.

The results obtained with t-test feature selection are consistent across all data sets. When using RELIEF feature selection, there is a degree of contrast between the results for the prostate cancer protein-abundance data set and the ones



(a)

(b)

(c)

(d)

Figure 3: Expected classification error as a function of number of classifiers in the ensemble for selected experiments with the prostate cancer protein abundance data (full results available on the companion website). The error of a single classifier is indicated by a horizontal dashed line. Marks indicate the values that are smaller for the ensemble classifier as compared to a single classifier at a 99% significance level, according to the one-tailed paired t-test.

for the gene-expression data sets, which may be attributed to the differences in technology as well as the fact that we do not employ baseline subtraction for the proteomics data in order to achieve better classification rates.

We remark that results are not expected to change much if ensemble sizes are increased further (beyond $m = 51$), as can be seen from convergence of the expected classification error curves in Figures 1–3.

4. Conclusion

In this paper we conducted a detailed empirical study of the ensemble approach to classification of small-sample genomic and proteomic data. The main performance issue is not whether the ensemble scheme improves the classification error of an unstable overfitting classifier (e.g., CART, NNET), or whether its classification error converges to a fixed limit; but rather whether the ensemble scheme will improve performance of the unstable overfitting classifier *sufficiently* to beat the performance of single stable, nonoverfitting classifiers (e.g., DLDA, LDA, and 3NN). We observed that this never was the case for any of the data sets and experimental conditions considered here, except in the case of the proteomics data set with RELIEF feature selection in acute small-sample cases, when nevertheless the performance of a single unstable overfitting classifier (in this case, CART) was better or comparable to the corresponding ensemble classifier. We observed that in most cases bagging does a good (sometimes, admirable) job of improving the performance of unstable overfitting classifiers, but that improvement was not enough to beat the performance of single stable nonoverfitting classifiers.

The main message to be gleaned from this study by practitioners is that the use of bagging in classification of small-sample genomics and proteomics data increases computational cost, but is not likely to improve overall classification accuracy over other, more simple, approaches. The solution we recommend is to use simple classification rules and avoid bagging in these scenarios. It is important to stress that we do not give a definitive recommendation on the use of the random forest method for small-sample genomics and proteomics data; however, we do think that this study does provide a step in that direction, since the random forest method depends partly, if not significantly, for its success on the effectiveness of bagging. Further research is needed to investigate this question.

References

[1] R. E. Schapire, "The strength of weak learnability," *Machine Learning*, vol. 5, no. 2, pp. 197–227, 1990.

[2] Y. Freund, "Boosting a weak learning algorithm by majority," in *Proceedings of the 3rd Annual Workshop on Computational Learning Theory (COLT '90)*, pp. 202–216, Rochester, NY, USA, August 1990.

[3] L. Xu, A. Krzyzak, and C. Y. Suen, "Methods of combining multiple classifiers and their applications to handwriting recognition," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 22, no. 3, pp. 418–435, 1992.

[4] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[5] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[6] L. Lam and C. Y. Suen, "Application of majority voting to pattern recognition: an analysis of its behavior and performance," *IEEE Transactions on Systems, Man and Cybernetics Part A*, vol. 27, no. 5, pp. 553–568, 1997.

[7] B. Efron, "Bootstrap methods: another look at the jacknife," *Annals of Statistics*, vol. 7, pp. 1–26, 1979.

[8] B. Efron, *The Jackknife, the Bootstrap and Other Resampling Plans*, vol. 38 of *CBMS-NSF Regional Conference Series in Applied Mathematics*, SIAM, Philadelphia, Pa, USA, 1982.

[9] S. Alvarez, R. Diaz-Uriarte, A. Osorio, et al., "A predictor based on the somatic genomic changes of the BRCA1/BRCA2 breast cancer tumors identifies the non-BRCA1/BRCA2 tumors with BRCA1 promoter hypermethylation," *Clinical Cancer Research*, vol. 11, no. 3, pp. 1146–1153, 2005.

[10] E. C. Gunther, D. J. Stone, R. W. Gerwien, P. Bento, and M. P. Heyes, "Prediction of clinical drug efficacy by classification of drug-induced genomic expression profiles in vitro," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 16, pp. 9608–9613, 2003.

[11] R. Díaz-Uríarte and S. Alvarez de Andrés, "Gene selection and classification of microarray data using random forest," *BMC Bioinformatics*, vol. 7, article 3, pp. 1–13, 2006.

[12] A. Statnikov, L. Wang, and C. F. Aliferis, "A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification," *BMC Bioinformatics*, vol. 9, article 319, pp. 1–10, 2008.

[13] G. Izmirlian, "Application of the random forest classification algorithm to a SELDI-TOF proteomics study in the setting of a cancer prevention trial," *Annals of the New York Academy of Sciences*, vol. 1020, pp. 154–174, 2004.

[14] B. Wu, T. Abbott, D. Fishman, et al., "Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data," *Bioinformatics*, vol. 19, no. 13, pp. 1636–1643, 2003.

[15] P. Geurts, M. Fillet, D. de Seny, et al., "Proteomic mass spectra classification using decision tree based ensemble methods," *Bioinformatics*, vol. 21, no. 14, pp. 3138–3145, 2005.

[16] B. Zhang, T. D. Pham, and Y. Zhang, "Bagging support vector machine for classification of SELDI-TOF mass spectra of ovarian cancer serum samples," in *Proceedings of the 20th Australian Joint Conference on Artificial Intelligence (AI '07)*, vol. 4830 of *Lecture Notes in Computer Science*, pp. 820–826, Gold Coast, Australia, December 2007.

[17] A. Assareh, M. Moradi, and V. Esmaeili, "A novel ensemble strategy for classification of prostate cancer protein mass spectra," in *Proceedings of the 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS '07)*, pp. 5987–5990, Lyon, France, August 2007.

[18] W. Tong, Q. Xie, H. Hong, et al., "Using decision forest to classify prostate cancer samples on the basis of SELDI-TOF MS data: assessing chance correlation and prediction confidence," *Environmental Health Perspectives*, vol. 112, no. 16, pp. 1622–1627, 2004.

[19] T. G. Dietterich, "An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization," *Machine Learning*, vol. 40, no. 2, pp. 139–157, 2000.

[20] S. Dudoit, J. Fridlyand, and T. P. Speed, "Comparison of discrimination methods for the classification of tumors using gene expression data," *Journal of the American Statistical Association*, vol. 97, no. 457, pp. 77–87, 2002.

[21] L. Hansen and P. Salamon, "Neural network ensembles," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 10, pp. 993–1001, 1990.

[22] E. Bauer and R. Kohavi, "An empirical comparison of voting classification algorithms: bagging, boosting, and variants," *Machine Learning*, vol. 36, no. 1, pp. 105–139, 1999.

[23] S. Y. Sohn and H. W. Shin, "Experimental study for the comparison of classifier combination methods," *Pattern Recognition*, vol. 40, no. 1, pp. 33–40, 2007.

[24] J. Hua, W. D. Tembe, and E. R. Dougherty, "Performance of feature-selection methods in the classification of high-dimension data," *Pattern Recognition*, vol. 42, no. 3, pp. 409–424, 2009.

[25] B. Efron, "Estimating the error rate of a prediction rule: improvement on cross-validation," *Journal of the American Statistical Association*, vol. 78, no. 382, pp. 316–331, 1983.

[26] L. Devroye, L. Gyorfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*, Springer, New York, NY, USA, 1996.

[27] U. M. Braga-Neto and E. Dougherty, "Classification," in *Genomic Signal Processing and Statistics*, E. Dougherty, I. Shmulevich, J. Chen, and Z. J. Wang, Eds., EURASIP Book Series on Signal Processing and Communication, pp. 93–128, Hindawi, New York, NY, USA, 2005.

[28] M. Chernick, *Bootstrap Methods: A Practitioner's Guide*, John Wiley & Sons, New York, NY, USA, 1999.

[29] E. Lehmann and J. Romano, *Testing Statistical Hypotheses*, Springer, New York, NY, USA, 2005.

[30] K. Kira and L. A. Rendell, "The feature selection problem: traditional methods and a new algorithm," in *Proceedings of the 10th National Conference on Artificial Intelligence (AAAI '92)*, pp. 129–134, San Jose, Calif, USA, July 1992.

[31] M. J. van de Vijver, Y. D. He, L. J. van't Veer, et al., "A gene-expression signature as a predictor of survival in breast cancer," *The New England Journal of Medicine*, vol. 347, no. 25, pp. 1999–2009, 2002.

[32] L. J. van't Veer, H. Dai, M. J. van de Vijver, et al., "Gene expression profiling predicts clinical outcome of breast cancer," *Nature*, vol. 415, no. 6871, pp. 530–536, 2002.

[33] A. Bhattacharjee, W. G. Richards, J. Staunton, et al., "Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 24, pp. 13790–13795, 2001.

[34] H. J. Issaq, T. D. Veenstra, T. P. Conrads, and D. Felschow, "The SELDI-TOF MS approach to proteomics: protein profiling and biomarker identification," *Biochemical and Biophysical Research Communications*, vol. 292, no. 3, pp. 587–592, 2002.

[35] B.-L. Adam, Y. Qu, J. W. Davis, et al., "Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men," *Cancer Research*, vol. 62, no. 13, pp. 3609–3614, 2002.