IMS Collections
Contemporary Developments in Bayesian Analysis and Statistical Decision Theory: A Festschrift for William E. Strawderman
Vol. 8 (2012) 235–246

DOI: 10.1214/11-IMSCOLL816

Ensemble classifiers

Dhammika Amaratunga, Javier Cabrera*, Yauheniya Cherckas and Yung-Seop Lee

Johnson & Johnson Pharma, Rutgers University, Johnson & Johnson Pharma and Dongguk University

Abstract: Ensemble classification methods like Random Forest are powerful and versatile classifiers. We explore variations in the ensemble approach and demonstrate the strong performance of ensemble versions of Linear Discriminant Analysis (LDA) variants such as LDA-PCA (LDA after a Principal Components Analysis step to reduce dimensionality) and LASSO in situations characterized by a huge number of features and a small number of samples such as DNA microarray data. We also demonstrate the value of enriching the ensembles with features that are most likely to be informative in situations where only a very small percentage of the features actually carries classification information. Notably, in the case studies we analyzed, the enriched ensemble procedure with LDA-PCA as base classifier had a misclassification rate that was essentially half that observed with Random Forest.

Contents

1. Introduction

Ensemble classification methods like Random Forest (2001a, 2001a) that operate by aggregating predictions from multiple classifiers are among the most powerful classification methods available today for a wide variety of problems (Amit and Geman (1997), Dietterich (2000), Lin and Jeon (2006), Meinshausen (2006), Biau et al. (2008)). This is certainly the case for datasets with huge numbers of features such as gene expression data from DNA microarray experiments. It is often quite difficult to attain high generalized classification accuracy with such datasets in large

e-mail: damaratu@its.jnj.com; cabrera@rutgers.edu; ycherka2@its.jnj.com; yung@dongguk.edu

*Corresponding Author: Prof. Javier Cabrera, Department of Statistics and Biostatistics, Rutgers University, Piscataway, NJ08854, USA

AMS 2000 subject classifications: Primary 62P10, 68T10, 68T05

Keywords and phrases: classification, ensemble, lasso, linear discriminant analysis, microarray, random forest

part because of the limited number of samples usually available in such experiments and the huge excess of features, which leaves many classification methods prone to overfitting. Nevertheless, ensemble methods offer increased accuracy, reliability and protection against overfitting when faced with such problems (Dudoit et al. (2002), Breiman and Cutler (2003), Lee et al. (2005), Amaratunga et al. (2008), Statnikov et al. (2008)).

In developing Random Forest, Breiman argued the importance of two factors in driving the performance of the procedure. First, the base classifiers that constitute the ensemble must be accurate. Second, the base classifiers must be diverse with greater diversity tending to reduce generalization error. Breiman sought to maximize the diversity in Random Forest by not only bagging cases (Breiman, 1996) but also by using unpruned classification trees as the base classifiers and choosing the optimum split at each node of each tree from among a simple random sample of features. Node-wise random feature selection attenuates correlation across features, reduces competition among features, and gives specialized features a chance to contribute. Since trees grown to maximum depth are good classifiers, bias is kept low and the first condition is satisfied as well. Assigning test cases to classes by majority vote over a set of bootstrapped classification trees compensates for overfitting in any one tree. Breiman (2001) invoked the Law of Large Numbers to argue that this procedure converged and would not overfit.

However, when the data contains only a few samples but a huge number of features of which only a small percentage actually carries classification information, the issue of diversity becomes less of a concern and the issue of accuracy becomes increasingly crucial. In such situations, if simple random sampling is used to select features at each node, it is possible that most of the base classifiers would lose accuracy as they could be comprised largely of uninformative features; as a result, the performance of the ensemble suffers. In this case, Amaratunga et al. (2008) demonstrate that enriching the base classifiers with features that exhibit better apparent separability information improves the overall classification accuracy of Random Forest despite the modest reduction in diversity.

However Enriched Forest inherits some of the drawbacks or pitfalls of Random Forest, that are well known. For example random forest will likely miss a structure that is hidden in the dimensionality and that is oblique to a coordinate-wise grid. We explore this concept further in this paper. We consider situations in which the data contain a huge number of features, particularly situations in which only a small fraction of the features carries classification information; these latter we refer to henceforth as "low-signal" situations. In such situations we will replace the node-wise simple random sampling of features with base-classifier-wise weighted random sampling of features and then also replace classification trees with more stable classifiers such as Linear Discriminant Analysis (Fisher, 1936) and LASSO (Tibshirani, 1996) and assess the impact on performance. We demonstrate that these new ensembles attain superior performance in certain low-signal situations. In fact, in all the examples studied, the ensemble version of LDA-PCA (Linear Discriminant Analysis after a Principal Components Analysis step to reduce dimensionality) significantly reduced the misclassification rate of Random Forest.

2. Methods

A classifier is constructed via a training set of the form $X_T = \{(x_i, y_i)i = 1, \ldots, N\}$, where x_i is a G-vector of data for the ith case and yi is a binary response variable

indicating the class to which the ith case belongs. In the type of problems of interest here, the number of features, G, tends to be very large, perhaps orders of magnitude larger than the number of cases, N. As such, the potential for overfitting is considerable. Therefore, it is crucial to derive a classifier that is not prone to overfitting and also, since there is rarely an independent test set available to assess the generalizability of the classification, one whose performance could be judged via X_T itself. Ensemble classifiers are able to support this dual objective. Running an ensemble classifier involves generating a bundle of base classifiers and then aggregating the predictions from these base classifiers so as to derive overall predictions and also a measure of classification performance (essentially following Breiman (1999)'s proposal of building many classifiers from small pieces or "bites" of data and pasting votes to derive predictions to in data mining situations where N is very large). The base classifiers could be classification trees as in Random Forest or other methods like Linear Discriminant Analysis or Support Vector Machines. As we shall now discuss, such ensembles can be constructed in various ways. A general algorithm that encompasses the following methods is given in the Appendix.

2.1. Bagging-only ensembles (BagE)

Breiman (1996) developed bagging as a way of improving unstable single-run classification procedures. The basis to bagging is the bootstrap step of selecting N cases at random with replacement from X_T. These "in-bag" cases, say there are N^* of them after discarding replicates, are used for deriving a base classifier. The remaining $(N - N^*)$ cases are "out-of-bag" cases and, since they are not used for deriving the base classifier, they can be used to assess how well the base classifier is performing at discriminating these cases; thus they can be used to assess the performance of the classifier. This process is repeated a large number (say R) of times and a bundle of R base classifiers is produced. The prediction for any case, whether it is a training case or a test case, is based on generating a prediction from each classifier for this case and assigning it an overall prediction via majority vote. The confusion matrix of predicted class versus actual class can be summarized to give an out-of-bag error rate which can be used as a performance measure for the ensemble as a whole. The out-of-bag error rate is calculated by estimated the error rate of each classifier in the out-of bag sample that was not used to construct the classifier. This is done to avoid or reduce the amount of overfitting.

2.2. Simple ensembles with simple random filtering (SimE)

Random case selection is carried out as above. Following that, G^* features (where $G^* = \sqrt{G}$) are selected using simple random sampling. A base classifier is determined using these N^* cases and G^* features. Predictions and performance assessments are done as above.

2.3. Enriched ensembles (EnrE)

In low-signal situations, the use of simple random sampling as in simple ensembles above would lead to sub-optimal classifications as most of the base classifiers would be comprised largely of uninformative features. Following Amaratunga et al. (2008), we explore the feasibility of tilting the feature selection towards informative features. Enrichment is done as follows. Once the random case selection is

done as for bagging-only ensembles, each feature is assigned a score depending on how well it separates the classes in the training set; the greater the separability the lower the score. The scores are then converted to weights inversely proportional to the scores so that only high separability features receive high weights. Then G^* features are selected using weighted random sampling with these weights instead of using simple random sampling, thereby increasing the likelihood that the features selected contain informative features. Other than this, the method is the same as simple ensembles. One way to generate scores is to use t tests. For each feature, perform a t-test with the N^* in-bag cases to test for a class difference. Calculate the p-values and convert them to FDR corrected p-values namely p-values (Benjamini and Hochberg (1996), Storey and Tibshirani, 2003). The feature with the i-th highest score (i.e., i-th smallest p-value) would then be assigned weight $w_i : w_i = \text{median}(a_{min}, w_i', a_{max})$, where $w_i' = (1/q_i) - 0.99$, q_i is its q-value, a_{min}=0.01 and a_max=999, which gives a wide range of variation (of order $G \approx 105$) across weights. Select G^* features using weighted random sampling without replacement with these weights w_i.

Remark. In spirit, Random Forest is a simple ensemble with classification tree as the classification method; however, there is a slight difference in that in Random Forest the feature selection is done at the node level rather than at the tree level.

Remark. q-values are FDR corrected p-values (Benjamini and Hochberg (1996)). In order to calculate q-values we assume that under the Null hypothesis p-values are uniformly distributed. q-values are obtained by dividing the ordered p-values $p_{(i)}$ by the α-quantile of the corresponding order statistic $p^0_{(i)}(\alpha)$. Then $q^*_{(i)} = p_{(i)}/p^0_{(i)}(\alpha)$. The sequence $q_{(i)}$ is forced to be monotonic on i by defining $q(\alpha) = q^*_{(\alpha)}$ and $q_{(i)} = max(q^*_{(i)}, q_{(i-1)})$.

Remark. The use of q-values rather than p-values for generating weights reduces the risk of overfitting and produces weights that are more representative of separation (see Amaratunga et al. (2008)).

Remark. When the sample size is very small and the features are multitudinous, a method such as Conditional t (Amaratunga and Cabrera, 2008), in which strength is borrowed across features, may improve the performance of enriched feature-select ensembles as suggested by Amaratunga et al. (2008) in their work on Enriched Random Forest.

Remark. The computational cost of the "Enrichment" step of the general algorithm described in the Appendix is of order N^* times G^*. Since this step is performed R times, once for each sample the cost is of order R times N^* times G^*. R and N^* is at least one order of magnitude smaller than G^* so the computation is reasonable when G^* is below 106.

The base classifiers that form the foundation of these ensembles can be constructed using almost any standard classification method. In Random Forest, the classification algorithm used is a classification tree, as its instability promotes diversity. However, in low-signal situations, the individual trees lack accuracy and the performance of the ensemble is weak.

A remedy is to instead use a more stable classification procedure such as Fisher's Linear Discriminant Analysis (LDA) (Fisher, 1936), one of the oldest and most widely used classification methods. However, one immediately encounters a difficulty in that even $\sqrt{G}$ features are many more features than there are cases so that LDA cannot be applied directly. There are however several ways to resolve this.

One is to first use Principal Components Analysis (PCA) and keep only the first $N-1$ principal components thereby reducing the dimensionality of the data from $\sqrt{G}$ to $N-1$. We will refer to this method as LDA-PCA (Belhumeur et al. (1997); used by Amaratunga and Cabrera (2004) for microarray data; a bagged version of LDA-PCA was proposed by Liu and Chen (2005)). Another is to ignore all correlations between features; this is Diagonal Linear Discriminant Analysis (DLDA), a method that seems to perform well with microarray data (Dudoit et al., 2002). Yet another is to use a penalized version of logistic regression which would be akin to penalized LDA; one such version is LASSO (originally developed by Tibshirani (1996), adapted for logistic regression by Lokhorst (1999) and used for gene selection by Shevade and Keerthi (2003)). Other classification methods such as Partial Least Squares (PLS), k Nearest Neighbors (kNN), and support vector machines (SVM) could be used (see Hastie et al. (2001) for a review of classification methods) and were included in our performance assessment in the next Section. All these methods were run using standard R-packages, and with the default settings that are recommended by the package developers.

3. Results

The performance of the various ensemble classification methods was evaluated using several datasets for which the classes were known. All the datasets are related to DNA microarrays capable of measuring the expression levels of genes, tens of thousands of genes at a time. These datasets are representative of the type of problem we consider here as typically gene expression measurements are only taken on a few samples and often only a few genes carry classification information. In these experiments, the samples are the cases and the genes are the features.

The datasets are listed in Table 1; all datasets, other than the Slc17A5 datasets, were downloaded from ArrayExpress. The extent to which the groups in these

TABLE 1

The datasets used in the evaluation. The statistics, S_q and S_h, are rough measures of group separation: S_q is 100 times the percentage of features with q-values less than 0.10 and S_h is $\Phi(p_H)$ where p_H is the p-value of the Hotelling's test statistic in the space spanned by the first $(N/4)$ principal components.

Dataset Name	No. of genes	Samples	S_q	S_h	Reference
Slc17A5 Day 0	45101	wild type (6) vs knockout (6)	0.016	1.73	Raghavan et al. (2007)
Slc17A5 Day 18	45101	wild type (6) vs knockout (6)	2.660	4.14	Raghavan et al. (2007)
Slc17A5 Day 0 (scrambled)	45101	wild type (6) vs knockout (6)	0.000	0.37	-
Slc17A5 Day 18 (scrambled)	45101	wild type (6) vs knockout (6)	0.000	0.74	-
Astrocytoma	12625	low grade (8) vs high grade (6)	2.503	3.35	MacDonald (2001)
Breast cancer	15926	normal (11) vs patients (24)	84.111	∞	Chan et al. (2005)
Epilepsy	31099	control (6) vs phenytoin (7)	12.586	3.26	Salomon (2005)
HIV Encephalitis	12625	reference (12) vs encephalitis (16)	0.000	2.92	Masiliah et al. (2004)
Human Lymph Node Sinus	22283	tonsils (10) vs lymph node (10	42.248	8.21	Martens et al. (2006)
Macular degeneration	12625	healthy (18) vs diseased (18)	7.810	4.62	Strunnikova et al. (2005)

datasets separate was assessed roughly via two measures:

- The percentage S_q (multiplied by 100) of features that have q-values less than 0.1.
- A standardized measure of the extent S_h by which the feature group means separate in the first $(N/4)$ principal components as measured by $\Phi(p_H)$, where p_H is the p-value of the Hotelling's test statistic in the space spanned by these coordinates.

Large values of S_q and S_h would indicate better separation. These values are reported in Table 1.

The datasets range from those for which there is a clear separation between groups such as the Slc17A5 Day 18 data and those for which there is a weak separation between groups such as the Slc17A5 Day 0 data (the Slc17A5 data are discussed by Moechars et al. (2005) and Raghavan et al. (2007) and used by Amaratunga et al. (2008) for assessing the performance of Enriched Random Forest). These are both comparisons between wildtype mice and mice whose Slc17A5 gene had been knocked out. The biology implies that there would be gene expression differences, subtle ones at the neo-natal (Day 0) stage and clear ones at later stages (such as on Day 18). The former is an instance of a low-signal situation. In addition, two "Scrambled" datasets were created by permuting the samples of the two Slc17A5 datasets, thereby ensuring that they exhibit no signal; these datasets will be used to verify that the methods are not overfitting, an aspect of the evaluation that is particularly important for the enriched ensembles as, if the weighting is not done carefully, it is possible to "find" spurious classifications in datasets that have no true separation.

These datasets were analyzed using the ensemble methods described above, including Random Forest (RF), Enriched Random Forest (ERF) and BagE, SimE and EnrE ensembles with PCA-LDA, DLDA, LASSO, PLS, kNN, and SVM in turn as base classifiers. For ERF, both t-based weights and Ct-based weights were used.

The results of the evaluation are shown in Table 2. They show that:

- In datasets exhibiting a clear separation between classes, such as the Slc17A5 Day 18 and Human Lymph Node Sinus datasets, all the ensemble methods are able to accurately detect the separation.
- In datasets where the separation between classes is subtle (i.e., in low-signal situations), such as the Slc17A5 Day 0 and HIV Encephalitis datasets, only certain enriched ensemble procedures, particularly the one driven by LDA-PCA, are able to pick up the separation.
- In datasets where there is no separation between classes, such as the Scrambled datasets, none of the ensemble methods report finding a separation. Thus, none of the methods tend to overfit.
- The enriched ensemble procedure driven by LDA-PCA was overall the best performer and more or less halved the error rate of the basic Random Forest procedure in every single dataset studied.

For all the methods, enrichment reduced the error rate of the procedure substantially (with a few exceptions). Therefore, in addition to the above performance assessment, we carried out a few simple simulations ("simple" in the sense that we did not attempt to simulate all the complexities of microarray data, but rather incorporated some of the characteristics of such data) to illustrate the performance of an enriched ensemble in comparison to a simple ensemble and in comparison to Random Forest and Enriched Random Forest.

TABLE 2

Results of running the various ensemble procedures through the datasets in Table 1. The numbers shown are the out-of-bag error rates with R=1000 runs.

Dataset Name	LDA PCA (BagE)	LDA PCA (SimE)	LDA PCA (EnrE)	DLDA DLDA (BagE)	DLDA DLDA (SimE)	DLDA DLDA (EnrE)	LASSO (BagE)	LASSO (SimE)	LASSO (EnrE)
Slc17A5 Day 0	0.417	0.583	0.000	0.500	0.583	0.250	0.000	0.500	0.000
Slc17A5 Day 18	0.000	0.000	0.000	0.083	0.083	0.000	0.083	0.000	0.083
Slc17A5 Day 0 (scrambled)	0.833	0.750	0.833	0.833	0.667	0.667	0.583	0.833	0.833
Slc17A5 Day 18 (scrambled)	0.667	0.583	0.583	0.583	0.583	0.583	0.750	0.750	0.750
Astrocytoma	0.143	0.143	0.071	0.429	0.429	0.214	0.214	0.143	0.214
Breast cancer	0.000	0.000	0.000	0.314	0.029	0.029	0.000	0.029	0.029
Epilepsy	0.077	0.154	0.077	0.538	0.077	0.077	0.077	0.077	0.077
HIV Encephalitis	0.143	0.179	0.179	0.286	0.286	0.286	0.179	0.214	0.179
Human Lymph Node Sinus	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Macular degeneration	0.083	0.111	0.056	0.139	0.139	0.111	0.028	0.083	0.028

Dataset Name	KNN (BagE)	KNN (SimE)	KNN (EnrE)	PLS (BagE)	PLS (SimE)	PLS (EnrE)	SVM (BagE)	SVM (SimE)	SVM (EnrE)
Slc17A5 Day 0	0.333	0.833	0.167	0.333	0.417	0.250	0.500	0.583	0.500
Slc17A5 Day 18	0.083	0.000	0.000	0.000	0.000	0.000	0.167	0.250	0.000
Slc17A5 Day 0 (scrambled)	0.833	0.667	0.667	0.583	0.667	0.667	0.833	0.750	0.833
Slc17A5 Day 18 (scrambled)	0.583	0.583	0.667	0.583	0.500	0.667	0.583	0.583	0.583
Astrocytoma	0.214	0.286	0.071	0.143	0.214	0.071	0.214	0.214	0.071
Breast cancer	0.029	0.029	0.029	0.000	0.000	0.000	0.000	0.029	0.029
Epilepsy	0.077	0.154	0.077	0.077	0.077	0.077	0.077	0.077	0.077
HIV Encephalitis	0.321	0.250	0.286	0.143	0.179	0.107	0.107	0.107	0.107
Human Lymph Node Sinus	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Macular degeneration	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Dataset Name	RF	ERF	ERF(Ct)
Slc17A5 Day 0	0.583	0.167	0.000
Slc17A5 Day 18	0.083	0.000	0.000
Slc17A5 Day 0 (scrambled)	0.750	0.833	0.667
Slc17A5 Day 18 (scrambled)	0.583	0.667	0.667
Astrocytoma	0.214	0.000	0.071
Breast cancer	0.029	0.029	0.029
Epilepsy	0.154	0.154	0.154
HIV Encephalitis	0.357	0.250	0.250
Human Lymph Node Sinus	0.000	0.000	0.000
Macular Degeneration	0.111	0.083	0.083

In the first simulation, we generated a dataset consisting of $N = 30$ samples, which fall into two classes with even-numbered samples in one class and odd-numbered samples in the other, and $G = k_U + k_I$ features, of which k_U were random and therefore uninformative and k_I were informative, representing a structure associated with the true classification and defined via a latent vari-

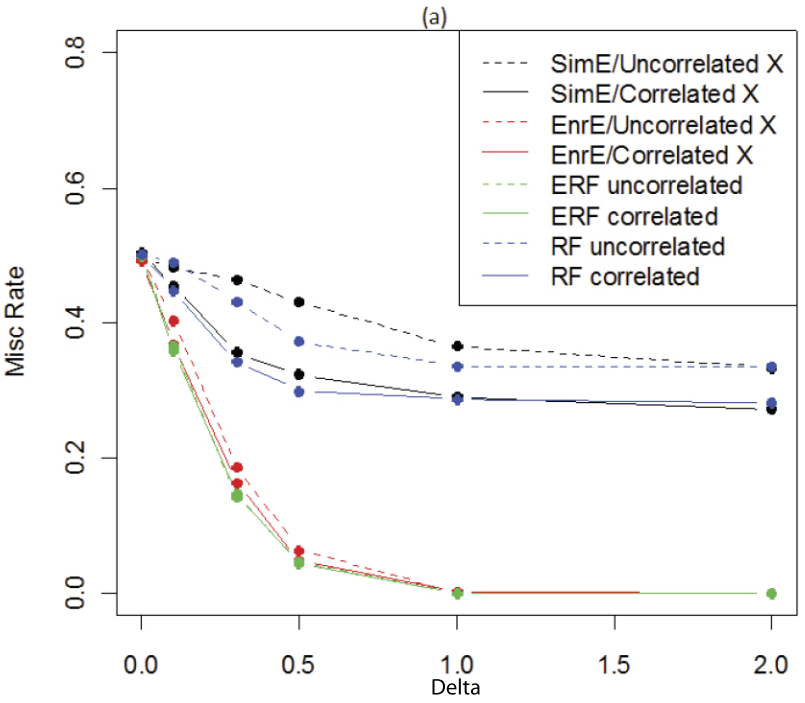

FIG 1. *Simulation results: Misclassification rates are shown as a function of* Δ *with* k_I *fixed at* $k_I = 2$. *Here* k_I *is the number of informative features and* Δ *is the degree of separation exhibited by the informative features. "Uncorrelated X" corresponds to the cases where there is no secondary structure and "Correlated X" corresponds to the cases where there is.*

able $Z_i = (-1)^i\Delta + \gamma_i, i = 1, \ldots, N$, with $\gamma_i \sim N(0, \tau^2)$. X was generated as: $X_{ij} = \epsilon_{ij}$ for $i = 1, \ldots, N$; $j = 1, \ldots, k_U$ and $X_{ij} = Z_i + \epsilon_{ij}$ for $i = 1, \ldots, N$; $j = k_U + 1, \ldots, k_U + k_I$, with $\epsilon_{ij} \sim N(0, \sigma^2)$. The response variable was generated from a binomial distribution: $Y \sim Binomial(logit^{-1}(Z_i), 1)$. For all the simulations, we set $\tau = \sigma = 0.2$ and all features were normalized to unit variance. The value of Δ represents the strength of the signal of the k_I informative features and by varying Δ we are able to evaluate the performance of a method under different signal strengths. For each Δ value of $0, 0.1, 0.3, 0.5, 1$ and 2, $k_U = 500$ and $k_I = 2$, we generated 300 datasets and, since the true classes are known, we used the average misclassification rate as a performance assessment measure to compare EnrE LDA-PCA to SimE LDA-PCA. A graph of average misclassification rate versus Δ is shown in Figure 1.

The second simulation mimics a situation in which there is an unobserved covariate V that is involved in the differential expression of some of the genes but not in the processes that we are studying. Because of the small sample size this covariate V might be unbalanced with respect to the response and show a mild correlation with Y but will imply a different grouping than the response. For example, suppose that of the two classes in the response one has 70% males while the other has 30% males and suppose that there is a set of k'_U genes that is more highly expressed in females than in males. Those genes will be mildly correlated with Y even though they are unrelated to the process which generated the Y grouping. This set of k'_U genes constitutes a secondary structure that is often present in microarray data but that may be unknown or even if suspected may be difficult to incorporate into

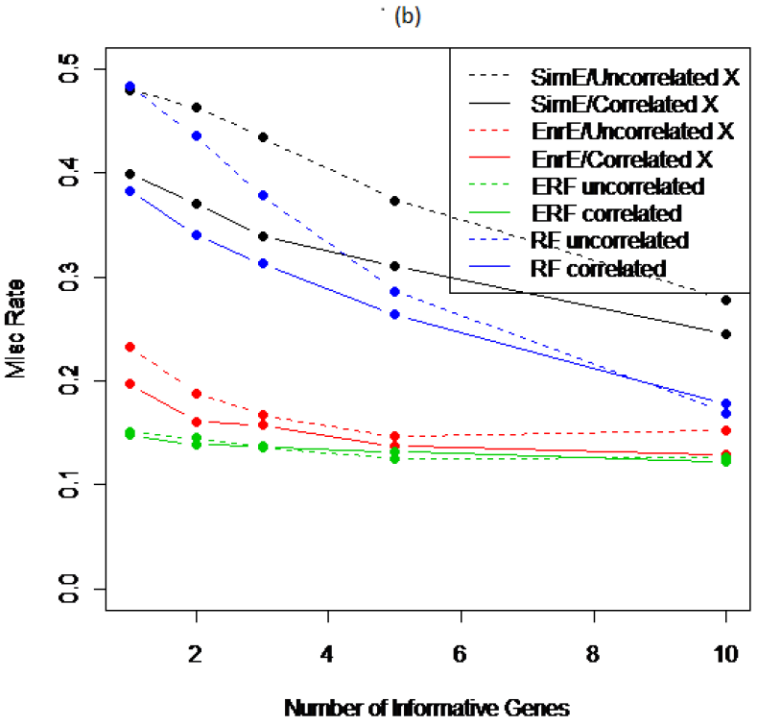

FIG 2. *Simulation Results. Misclassification rates are shown as a function of k_I with Δ fixed at $\Delta = 0.3$. Here k_I is the number of informative features and Δ is the degree of separation exhibited by the informative features. "Uncorrelated X" corresponds to the cases where there is no secondary structure and "Correlated X" corresponds to the cases where there is.*

the analysis via covariates due to the limited sample size. In order to study the effects of such a structure, a simulation was generated in the same way as above, except that the first $k'_U = 200$ of the $k_U = 500$ random features were modified by adding another latent variable V that was mildly correlated to the response Y in the same way as in the example. V_i was assigned the values $\{-1, 1\}$ with probabilities $\{0.7, 0.3\}$ for i even and with probabilities $\{0.3, 0.7\}$ for i odd. The correlation between V and Y is approximately 0.6. Then the first k'_U features are generated as $X_{ij} = V_i + \epsilon_{ij}, i = 1, \ldots, N$, and $j = 1, \ldots, k'_U$. The results of this simulation are also shown in Figure 1.

The third and fourth simulations are the same as the previous two simulations except that instead of varying δ we fixed it at $\delta = 0.3$ and changed the value of k_I over the set $1, 2, 3, 5, 10$. The results of these simulations are shown in Figure 2.

In all four simulations, the enriched procedures (ERF and EnrE) performed substantially better than the non-enriched procedures. For very small δ, the separation between classes was clearly elusive to all methods, but, as δ increased, EnrE and ERF were able to detect the separation much more rapidly than SimE and RF. Thus, even for moderate values of δ, there was a clear advantage to EnrE versus SimE. For large values of δ, EnrE and ERF reached perfect classification whereas SimE and RF appeared to plateau at about 30% misclassification. The presence of a moderate secondary structure altered the likelihood of misclassification for SimE and RF but not for EnrE or ERF. Regardless of the value of δ, there was a clear advantage to EnrE and ERF over SimE and RF when the number of separating

genes was small. As the number of features carrying signal increased, the performance of all the methods improved. However, EnrE and ERF was already almost at peak performance for that value of δ with a small k_I, whereas for SimE, the peak performance only occurred when k_I was large, in which case both SimE and EnrE performed similarly. Thus this simulation study implies that:

- For SimE and RF (the non-enriched procedure) to have good performance, both the size of the signal and the number of features carrying the signal must be large. There was no difference in performance between SimE and RF.
- For EnrE or ERF (the enriched procedures) to have good performance, the size of the signal can be moderate and the number of features carrying the signal could be very small. There was no difference in performance between EnrE and ERF.

4. Discussion

Over the years, ensemble methods have been shown to be effective classifiers in situations prone to overfitting. In this paper, we have presented a novel class of ensemble classifiers and demonstrated, via several real data examples and simulations, the superior performance of these new methods. In cases where the classification is driven by only a small percentage of the recorded features, the value of enriching the ensembles by incorporating a q-value based weighting scheme was also shown.

We used microarray data here because it is the type of data that motivated this work and because it is a good and popular example of the high dimensional plus low sample size structure we emphasize in this paper. We expect that, with the great strides being made in data collection and data management technologies, such data are likely to be more common in the future and we conjecture that the methodology we have presented here would be applicable generally.

An R library, Eclass, of the ensemble classification methods described in this paper is available at the authors' websites:

http://www.rci.rutgers.edu/~cabrera/DNAMR/

http://www.geocities.com/damaratung/

Appendix A: The general algorithm

Run the following steps R times.

1. Select N cases with replacement, discarding any replicates. These are the N^* in-bag cases that will be used for developing a base classifier. The remaining $(N - N^*)$ cases are the out-of-bag cases.
2. Skip this step for BagE and SimE. For EnrE only, do the following. For each feature, perform a t-test with the N^* in-bag cases to test for a class difference. Calculate the p-values and convert them to q-values.
3. Skip this step for BagE. For SimE, set $w_i = 1$ for all i. For EnrE, calculate a weight wi for each feature i: $w_i = median(a_{min}, w'\_i, a_{max})$, where $w'_i = (1/q_i) - 0.99$. $a_{min} = 0.01$ and $a_{max} = 999$.
4. For BagE, retain all G features ($G^* = G$). For SimE and EnrE, select G^* features (where $G^* = \sqrt{G}$) using random sampling without replacement with weights w_i.
5. Determine a base classifier using these N^* cases and G^* features. Finally, use the ensemble of base classifiers to predict the class of the $(N - N^*)$ out-of-bag cases.

References

Amaratunga, D. and Cabrera, J. (2004). *Exploration and Analysis of DNA Microarray and Protein Array Data*. Wiley, New York.

Amaratunga, D. and Cabrera, J. (2009). A conditional t suite of tests for identifying differentially expressed genes in a DNA microarray experiment with little replication. *Statistics in Biopharmaceutical Research* **1** 26–38.

Amaratunga, D., Cabrera, J. and Lee, Y.S. (2008). Enriched random forest. *Bioinformatics* **24** 2010–2014.

Amit, Y. and Geman, D. (1997). Shape quantization and recognition with randomized trees. *Neural Computation* **9** 1545–1588.

Belhumeur, P.N., Hespanha, J.P. and Kriegman, D.J. (1997). Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection, *IEEE Trans. PAMI*, Special Issue on Face Recognition, **19** 711–20.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B* **57** 289–300.

Biau, G., Devroye, L. and Lugosi, G. (2008). Consistency of random forests and other averaging classifiers. *Journal of Machine Learning Research* **9** 2015–2033.

Breiman, L. (1996). Bagging Predictors. *Machine Learning* **26** 123–140.

Breiman, L. (1999). Pasting small votes for classification in large databases and on-line. *Machine Learning* **36** 85–103.

Breiman, L. (2001a). Random Forests. *Machine Learning* **45** 5–32.

Breiman, L. (2001b). Weld Lecture II: Looking Inside the Black Box.

Breiman, L. and Cutler, A. (2003). Random Forests Manual (version 4.0), Technical Report of the University of California, Berkeley, Department of Statistics.

Dietterich, T.G. (2000). An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization. *Machine Learning* **40** 139–157.

Dudoit, S., Freeland, J. and Speed, T.P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association* **97** 77–87.

Fisher, R.A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics* **7** 179–188.

Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer-Velag, New York.

Lee, J.W., Lee, J.B., Park, M. and Song, S.H. (2005). An extensive evaluation of recent classification tools applied to microarray data. *Computational Statistics and Data Analysis* **48** 869–885.

Lin, Y. and Jeon, Y. (2006). Random forests and adaptive nearest neighbors. *Journal of the American Statistical Association* **101** 578–590.

Liu, J. and Chen, S. (2005). Resampling LDA/QR and PCA+LDA for face recognition, Australian Conference on Artificial Intelligence, 1221–1224.

Lokhorst, J. (1999). The lasso and generalized linear models. Honors Project. University of Adelaide, Adelaide.

Meinshausen, N. (2006). Quantile regression forests. *Journal of Machine Learning Research* **7** 983–999.

Moechars, D., Vanacker, N., Cryns, K., Andries, L., Mancini, G. and Verheijen, F. (2005). Sialin-deficient mice: a novel animal model for infantile free sialic acid storage disease, ISSD: Society for Neuroscience 35th Annual Meeting. Washington, USA.

Raghavan, N., De Bondt, A., Talloen, W., Moechars, D., Göhlmann, H. and Amaratunga, D. (2007). The high-level similarity of some disparate gene expression measures. *Bioinformatics* **23** 3032–3038.

Shevade, S.K. and Keerthi, S.S. (2003). A simple and efficient algorithm for gene selection using sparse logistic regression. *Bioinformatics* **19** 2246–2253.

Statnikov, A., Wang, L. and Aliferis, C.F. (2008). A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. *BMC Bioinformatics* **9** 319.

Storey, J.D. and Tibshirani, R. (2003). Statistical significance for genome-wide studies. *Proceedings of the National Academy of Sciences* **100** 9440–9445.

Strunnikova, N., Hilmer, S., Flippin, J., Robinson, M., Hoffman, E. and Csaky, K. (2005). Differences in gene expression profiles in dermal fibroblasts from control and patients with age-related macular degeneration elicited by oxidative injury. *Free radical biology and medicine* **39** 781–96.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Royal. Statist. Soc B* **58** 267–288.