Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics & Data Analysis 51 (2007) 6166–6179
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Classification by ensembles from random partitions of high-dimensional data

Hongshik Ahn[a,*], Hojin Moon[b], Melissa J. Fazzari[a], Noha Lim[a], James J. Chen[b], Ralph L. Kodell[c]

[a] *Department of Applied Mathematics and Statistics, Stony Brook University, Stony Brook, NY 11794-3600, USA*
[b] *Division of Biometry and Risk Assessment, National Center for Toxicological Research, FDA, Jefferson, AR 72079, USA*
[c] *Department of Biostatistics, University of Arkansas for Medical Sciences, 4301 West Markham Street, Slot 781, Little Rock, AR 72205, USA*

Received 20 September 2006; received in revised form 21 December 2006; accepted 22 December 2006
Available online 17 January 2007

Abstract

A robust classification procedure is developed based on ensembles of classifiers, with each classifier constructed from a different set of predictors determined by a random partition of the entire set of predictors. The proposed methods combine the results of multiple classifiers to achieve a substantially improved prediction compared to the optimal single classifier. This approach is designed specifically for high-dimensional data sets for which a classifier is sought. By combining classifiers built from each subspace of the predictors, the proposed methods achieve a computational advantage in tackling the growing problem of dimensionality. For each subspace of the predictors, we build a classification tree or logistic regression tree. Our study shows, using four real data sets from different areas, that our methods perform consistently well compared to widely used classification methods. For unbalanced data, our approach maintains the balance between sensitivity and specificity more adequately than many other classification methods considered in this study.

Keywords: Class prediction; Classification tree; Cross validation; Logistic regression; Majority voting; Risk profiling

1. Introduction

It is a well-understood phenomenon that a prediction model built from thousands of available predictor variables (m) and a relatively small sample size (N) can be quite unstable (Miller, 2002). Models that are developed using an intense selection process are highly prone to change with a new training sample. Furthermore, there is a multiplicity of good models when $N \ll m$, as observed by examination of model fit. In the gene expression literature, much has been written on the instability of gene expression "signatures". Numerous studies have illustrated the variability of the components of the signature and have cautioned that the final models are simply representative of many other potential profiles (Sparano et al., 2005; Michiels et al., 2005). Moreover, the model selected may generalize poorly to a new data set. Simple model averaging (Burnham and Anderson, 2002) has been utilized to address this issue by using a weighted average of many competing models, where the weight is based on an information theoretic statistic such as

* Corresponding author. Tel.: +1 631 632 8372; fax: +1 631 632 8490.
E-mail address: hahn@ams.sunysb.edu (H. Ahn).

0167-9473/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2006.12.043

AIC (Akaike, 1974). This allows the final prediction to be a function of many different models, a result that is inherently more stable than a single model.

Ensemble methodology is a natural next step to simple model averaging for class prediction. An ensemble uses the predictions of multiple base classifiers, typically through majority vote or averaged prediction, to produce a final ensemble-based decision (Breiman, 1996, 1998, 2001; Freund and Schapire, 1996). The ensemble-based prediction typically has lower generalization error rates than using a single model; the difference depending on the type of base classifier used, ensemble size and the diversity or correlation between base classifiers. We demonstrate in Section 2 that low or negatively correlated classifiers improve accuracy over positively correlated ones.

According to Duin and Tax (2000), ensemble methods for combining classifiers fall into three categories: (1) ensembles that combine classifiers of the same type trained on different types of features (parallel combining), (2) ensembles that combine classifiers of different types trained on the same set of features (stacked combining), and (3) ensembles that combine classifiers of the same type trained on the same set (or subsets of the same set) of features (weak combining). This paper is concerned with ensembles in category 3. Recently three ensemble voting approaches in this category, Boosting (Schapire, 1990; Freund and Schapire, 1996, 1997), bagging (Breiman, 1996) and random subspaces (RS: Ho, 1998) have received attention. Boosting changes adaptively the distribution of the training set based on the performance of previously created classifiers. For combining the classifiers, it takes a weighted majority vote of their predictions. The bagging algorithm uses bootstrap samples to build the base classifiers. The final classification produced by the ensemble of these base classifiers is obtained using equal weight voting. RS combines multiple classification trees constructed in randomly selected subspaces. The final classification is obtained by an equal weight voting of the base trees. Breiman (2001) developed random forest (RF) by combining classification trees such that each tree is generated by bagging and a random subspace of the predictors is used at each node.

We introduce an ensemble-based approach for classification called CERP (classification by ensembles from random partitions). This approach is designed specifically for high-dimensional data sets for which a classifier is sought. Variable pre-selection is not required. We are able to bypass the constraint of large m and small N by randomly partitioning the m variables into n mutually exclusive subspaces. A benefit of partitioning the input space is that each subspace may be treated separately until aggregation, a computational advantage that will be important as the dimension of data sets grows beyond that which may be easily handled as a whole. An optimal tree (Breiman et al., 1984) is built within each subspace using only the m/n-dimensional space of the variables in the partition. CERP uses two main methods for generating base classifiers: C-T CERP creates optimal classification trees and LR-T CERP creates logistic regression trees (see Ahn and Chen, 1997). Both methods are illustrated in this article. CERP combines the results of these multiple trees to achieve an improved accuracy of class prediction by a majority voting or by taking the average of the predicted values within an ensemble. In LR-T CERP, logistic regression model can be used without losing the ensemble accuracy for data with $N \ll m$ by a random partition without a variable selection.

Multiple ensembles are generated by randomly re-partitioning the feature space and building trees. While CERP captures most of the features contained in the data, only a few randomly selected variables are used by each tree classifier in C-T CERP or LR-T CERP. When we have multiple ensembles, fresh new information can be obtained by a different partition of the variables in each additional ensemble. The multiple ensembles contribute to a further gain of the overall accuracy. A major gain of CERP over a single optimal tree is achieved by the random partitioning in a single ensemble. The multiple ensemble allows overlap of feature spaces, and the gain by adding ensembles is moderate.

Although CERP is not fundamentally different from the other ensemble classification methods, the method for creation of diverse base classifiers in the CERP algorithm may allow for better accuracy with smaller ensembles. The feature spaces in different classifiers are mutually exclusive in CERP due to a random partition. Both RF and RS cause overlap of the predictor variables among base classifiers due to the random selection of features. Hansen and Salamon (1990), Ho (1998) and Kuncheva et al. (2003) noted that ensemble error rate is most reduced in ensembles whose members make individual errors in a less correlated manner. Therefore, we expect a rapid error reduction by the CERP approach.

Using four high-dimensional data sets from different areas of application, we compare CERP to the most commonly used and best performing classification methods: RF, support vector machines (SVM: Vapnik, 1995), Boosting, k-nearest neighbors (kNN), linear discriminant analysis (LDA), shrunken centroid (SC: Tibshirani et al., 2002) and single optimal trees (CART: Breiman et al., 1984; QUEST: Loh and Shih, 1997). The applications are (1) the detection of allelic expression of imprinted genes based on human and mouse genomic data (Reik and Walter, 2001), (2) the classification of human colon tissue samples for cancer status based on genomic profiles (Alon et al., 1999), (3) the

classification of chemicals with respect to estrogen binding activity based on structural descriptors and physical properties (Blair et al., 2000), and (4) the classification of acute leukemias into acute lymphoblastic leukemia (ALL) or acute myeloid leukemia (AML) based on each individual patient's gene-expression profile (Golub et al., 1999). We show that the accuracy of CERP is consistently one of the best among the classification methods considered and that CERP can improve the balance of sensitivity and specificity over certain other methods. It is anticipated that the proposed methods can be used to improve class prediction in many other areas of application involving high-dimensional prediction sets.

The proposed methods are implemented in C and R. All the parameter values are determined in the training phase. Therefore, no fine tuning is necessary in running CERP. A downloadable version of C-T CERP program is provided at http://www.ams.sunysb.edu/∼hahn/research/CERP.html.

2. Enhancement of class prediction by ensemble voting methods

A motivation for ensembles is that a combination of the outputs of many weak classifiers produces a powerful committee (Hastie et al., 2001). Assume independence among the n classifiers, where n is odd. We note that making n odd prevents ties. Let X_i denote a random variable indicating a correct classification by the ith classifier. If the prediction accuracy of each classifier is p, then $X_i \sim \text{Bernoulli}(p)$, and the number of accurate classifications by the ensemble majority voting method is $Y = \sum_{i=1}^{n} X_i \sim \text{binomial}(n, p)$. We let $n = 2k + 1$, where k is a nonnegative integer. Define $A_n = P(Y \geqslant k + 1)$. Then the prediction accuracy of the ensemble classification by a majority voting is

$$A_n = \sum_{i=k+1}^{n} \binom{n}{i} p^i (1-p)^{n-i}. \tag{1}$$

Lam and Suen (1997) showed that $A_{2k+1} = .5$ for $k = 0, 1, \ldots$ when $p = .5$; the sequence $\{A_{2k+1}\}$ is strictly increasing when $p > .5$; and $\{A_{2k+1}\}$ is strictly decreasing when $p < .5$. If n is large, then $Y \xrightarrow{d} N(np, np(1-p))$ by the central limit theorem. It is easy to show that $\lim_{k\to\infty} A_{2k+1} = 1$, and the prediction accuracy of the ensemble voting method converges to 1 when $p > .5$,

If the classifiers in the ensemble are correlated, then we can use the beta-binomial model (Williams, 1975). This model allows only positive correlation ρ in order to satisfy $\text{Var}(p) > 0$. Prentice (1986) showed that the beta-binomial model may be extended to cases where $\rho < 0$ for certain values. His extended beta-binomial model is valid when $\rho \geqslant \max\{-p(n-p-1)^{-1}, -(1-p)[n-(1-p)-1]^{-1}\}$.

Table 1 illustrates the prediction accuracy obtained by ensemble majority voting. When $\rho = 0$, the standard binomial probability in (1) is used for $n \leqslant 25$, and the normal approximation is used for a larger n. The beta-binomial model is used when the correlation is positive, and the extended beta-binomial model is used when the correlation is negative. The table illustrates that negatively correlated classifiers improve the prediction accuracy more rapidly than the independent classifiers, while the improvement slows down when the correlation increases.

The improvement of the ensemble accuracy illustrated above is valid under the assumption of equal accuracy of the base classifiers and equal correlation among the classifiers. Without these constraints, Breiman (2001) obtained the upper bound for the generalization error when the accuracy of each classifier is at least a half. According to this theorem, the ensemble accuracy converges to 1 when the classifiers are independent. When the classifiers are correlated, the accuracy will converge to a point within the range between $1 - \overline{\rho}(1-s^2)/s^2$ and 1, where s is the strength of the set of classifiers and $\overline{\rho}$ is the average correlation of the tree classifiers.

These results imply that the ensemble voting method by the CERP approach can be improved fast in terms of class prediction accuracy by reducing high correlation caused by the overlap of predictor variables. However, there is a limitation. Since the number of disjoint subsets for a fixed set of predictors is limited, convergence to the perfect accuracy cannot be achieved and there is a bound. Furthermore, p decreases as n increases because the number of disjoint subsets (n) in an ensemble and the number of predictors in a subset are inversely proportional given a fixed number of predictors in a data set. Thus the improvement of the ensemble accuracy is expected to be slower than the numbers shown in Table 1. However, for high-dimensional data with hundreds or thousands of predictor variables, a fast improvement of prediction accuracy can be achieved by tens or hundreds of classifiers generated in CERP.

Table 1
Enhancement of the prediction accuracy by ensemble majority voting

n	ρ	p (prediction accuracy of each base classifier)						
		.50	.55	.60	.70	.80	.90	.95
3	−.05	.5	.58	.66	.80	.91	.98	NA[a]
	0	.5	.57	.67	.78	.90	.97	.99
	.1	.5	.57	.64	.76	.87	.95	.98
	.3	.5	.56	.62	.73	.84	.93	.97
7	−.025	.5	.62	.73	.90	.98	NA	NA
	0	.5	.61	.71	.87	.97	1.00	1.00[b]
	.1	.5	.59	.67	.81	.92	.98	.99
	.3	.5	.57	.63	.75	.86	.94	.97
15	−.01	.5	.67	.81	.96	1.00	NA	NA
	0	.5	.65	.79	.95	1.00	1.00	1.00
	.1	.5	.60	.70	.85	.95	.99	1.00
	.3	.5	.57	.64	.76	.87	.95	.98
25	−.01	.5	.72	.88	.99	NA	NA	NA
	0	.5	.69	.85	.99	1.00	1.00	1.00
	.1	.5	.61	.71	.87	.96	.99	1.00
	.3	.5	.57	.64	.77	.87	.95	.98
101	0	.5	.84	.98	1.00	1.00	1.00	1.00
	.1	.5	.62	.73	.89	.97	1.00	1.00
	.3	.5	.57	.64	.77	.88	.95	.98

[a]Not available using the extended beta-binomial model by Prentice (1986).
[b]Greater than or equal to .995.

3. CERP: classification by ensembles from random partitions

We propose a method for constructing CERP. The goal of this study is, by combining a group of weaker C-T, to achieve a classifier with a higher prediction accuracy than a single optimal tree obtained from the same sample. The overall scheme of the technique is shown in Fig. 1. We let Θ be the space of the predictors. In order to minimize the correlation among the ensemble of trees, Θ is randomly partitioned into k subspaces $(\theta_1, \theta_2, \ldots, \theta_k)$ with roughly equal sizes. Since the subspaces are randomly chosen from the same distribution, we assume that there is no bias in selection of the predictors in each subspace. At each of these subspaces, we construct a single optimal tree classifier. Based on the randomness, we expect nearly equal probability of the classification error among the k classifiers, and improvement of the prediction accuracy can be achieved as demonstrated in Section 2. C-T CERP uses a classification tree and LR-T CERP uses a logistic regression tree as the base classifier. CERP combines the results of these multiple trees to achieve an improved accuracy of class prediction by a majority voting of the classifiers or by taking the average of the predicted values within an ensemble. For a further improvement of the performance of CERP, we investigated a majority voting among a set of ensembles.

The number of feature spaces in a random partition is determined in the training phase by a nested cross validation in each learning set. A 10-fold CV is used for the C-T CERP, but a 3-fold CV is used for LR-T CERP for computational efficiency. In each learning set of a 10-fold CV, we first partition the predictor space such that a subspace has around $N/2$ predictors, build a CERP model, and calculate the accuracy in each subspace. In the same way, we attempt $N/3$, $N/4, \ldots, N/10$ and $N/12$ for the size of each subspace. The partition size resulting in the highest overall accuracy is chosen among these. Thus N/i will be chosen for some integer i, $i = 2, \ldots, 10$ or 12. The second step is to search the optimal size of the subspaces by a dual bisection method between $N/(i-1)$ and N/i, and between N/i and $N/(i+1)$ based on the overall accuracy. From this, we have two candidates for the partition size. We take the one with higher overall accuracy. This adaptive bisection algorithm is faster than a grid search. Further, it can succeed to obtain the global maximum which may be missed by the conventional bisection method.

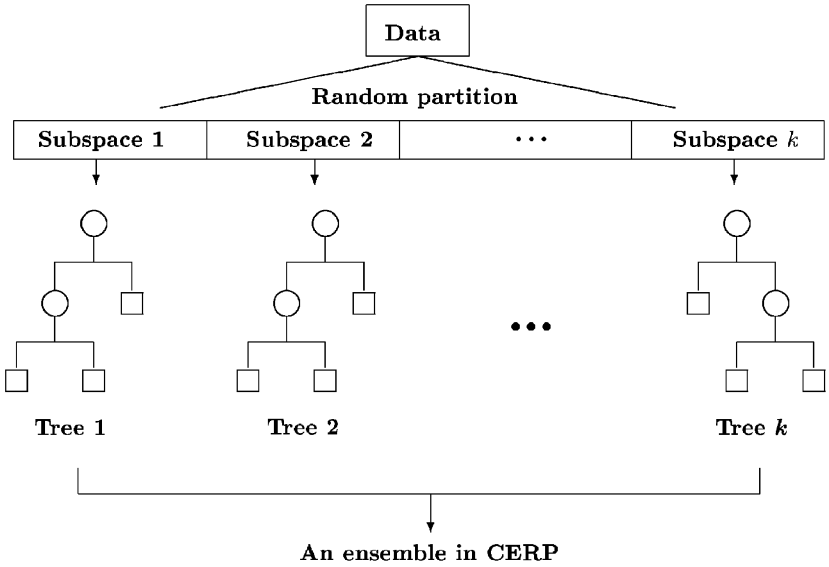

Fig. 1. The proposed CERP approach. The prediction is obtained by an ensemble of individual classification trees from each subspace of the predictors.

In a majority voting, the predicted values are classified as either 0 or 1 using a given threshold by each base classifier. For each observation, a majority voting is performed. As an alternative approach to the majority voting, we considered an averaging method. In this approach, the predicted values from all the base classifiers in an ensemble are averaged and classified as either 0 or 1 using a threshold on this average. Although the majority voting and averaging methods are similar in principle, the latter utilizes the fitted values of the logistic regression better in each tree. Thus the results by LR-T CERP given in this paper are obtained by the averaging method within an ensemble, while the results by C-T CERP are obtained by the majority voting method. The performance of CERP is evaluated using repeated CVs as described in Section 6.

3.1. C-T CERP: classification tree CERP

An optimal classification tree based on the CART algorithm (Breiman et al., 1984) is used as a base classifier in C-T CERP. In order to avoid overfitting data, the minimal cost-complexity pruning method is used. After a large tree is constructed, a nested sequence of subtrees is obtained by progressively deleting branches according to the pruning method. A 10-fold CV is used to obtain an optimal size of tree. In selecting the subtree with the smallest estimated CV error, we use the 1-SE rule (Breiman et al., 1984). In a base tree of C-T CERP, a node does not split and is declared a terminal node if the node contains a sample in only one class, or the split causes a child node having a sample size less than 5. The entire C-T CERP algorithm is implemented in C. The base tree is implemented in C, and the results match with *rpart* (Therneau and Atkinson, 1997) which is based on the CART methodology in the R package library.

3.2. LR-T CERP: logistic regression tree CERP

As an alternative approach to C-T CERP, we developed LR-T CERP. A logistic regression model requires a variable selection if the number of predictors exceeds the sample size. In LR-T CERP, however, variable selection is not required because each tree is constructed from a small subspace of the predictor variables unless the sample size in a terminal node is very small. For high-dimensional data, this is a huge advantage over other methods based on logistic regression models. Base trees of LR-T CERP are constructed using *rpart* in the R package library. At each terminal node of the pruned base trees in LR-T CERP, we fit the full logistic regression model with the given subset of predictors in each subset of the random partition. We implemented the entire algorithm of LR-T CERP in R. A node of a base tree does not split and is declared terminal if the sample size is less than 20. If the sample size (n_t) is smaller than or equal to

the number of predictors (m_t) in a terminal node t, then a univariate logistic regression model is fit with each predictor, and the $n_t - 2$ predictors with the smallest deviances plus the intercept term are included in the model. However, we observed that the sample size was larger than the number of predictors in the terminal node most of the time.

In order to improve a balance between sensitivity and specificity, two approaches have been attempted by researchers. Pazzani et al. (1994) and Domingos (1999) assigned a high cost to misclassification of the minority class, and Chen et al. (2005) proposed an ensemble classifier by building base classifiers with balanced samples using resampling techniques. In LR-T CERP, an optimal decision threshold for classification is searched in the base classifiers in the training phase. This approach shares the same principle as the methods by Pazzani et al. (1994) and Domingos (1999). Instead of a threshold of .5, a high misclassification cost may be assigned by using the rate of the positive responses in the data as a threshold. When r is the rate of the positive responses, we classify a sample as 1 if the fitted value is larger than r, and classify it as 0 otherwise. The rate of the positive responses is not necessarily the optimal choice of the threshold in terms of balancing sensitivity and specificity. The optimal threshold usually lies between .5 and the rate of the positive responses. In this study, we observed that the choice of a threshold did not affect the balance of C-T CERP significantly. However, we found a substantial improvement in balancing sensitivity and specificity for LR-T CERP using a different threshold from .5 for unbalanced data.

To search the optimal threshold of LR-T CERP, a nested 10-fold CV is performed in each learning set $L_i, i = 1, \ldots, 10$ as follows: within L_i, we use a finite grid with increment of .01 between .5 and r.

1. By applying each of the thresholds $ts_j, ts_j = .50, .51, \ldots, r$ (or $ts_j = r, r + .01, \ldots, .49, .50$), conduct the following 10-fold CV: construct an LR-T CERP classifier with one ensemble in each of the learning samples $L_{i1}, \ldots, L_{i,10}$ and evaluate the accuracy using the corresponding test samples with ts_j.
2. Choose the threshold with the highest prediction accuracy from part 1, say ts_i.
3. Apply ts_i to the test sample corresponding to L_i.

Only one ensemble is used in this nested CV because of the tendency that the optimal threshold for LR-T CERP is similar for one or multiple ensembles.

4. Existing classification and prediction methods

4.1. RF: random forest

RF is available as a package (*RandomForest*) in R. The number of trees generated may vary using the *ntree* option in R, but it has been shown to work well at the default of *ntree* = 500. Various values of *ntree* have been examined for all data sets we considered, but no improvement has been observed by increasing it beyond 500. The *ntree* value with the best accuracy is presented in this comparison. The number of features selected randomly at each node may also be varied; however, the default value of $m^{1/2}$ (or floor($m^{1/2}$) for a noninteger value) seems to have consistently good results across many examples according to our test with various choices. The RF program in the R package with the best options are used in our comparison.

4.2. SVM: support vector machines

The SVM program in R using the *e1071* package is used in the comparison. Some care needs to be taken with respect to the choice of kernel (we examined linear and radial basis using default parameters) as well as the parameters for these transformations such as kernel windows.

4.3. Boosting methods

The R Boosting package (*boost*) contains four different approaches: AdaBoost, LogitBoost, L2Boost, and BagBoost. LogitBoost performs slightly better than AdaBoost in general in a variety of traditional classification problems and microarray gene expression data (Dettling and Buhlman, 2003). Since BagBoost conducts repeated baggings in addition to Boosting, it is more computer-intensive than the other three methods. Since LogitBoost, L2Boost, and BagBoost gave similar results, we include AdaBoost and LogitBoost in the comparison. These R Boosting programs use the

classification tree (*rpart*) with a single split as the base classifier. The number of Boosting iterations used was 100 (*mfinal* = 100).

4.4. kNN: k-nearest neighbor classifiers

The R kNN package (*class*) is used in this comparison. Following the method of Dudoit et al. (2002), we used the ratio of between group to within group sums of squares (BW ratio) for each feature and retain those with the highest ratio. Optimal numbers of predictor variables (p) and the value of k in the nearest neighbor were searched in the training phase using nested CV. Pairs of (p, k) with the highest accuracy was chosen in each learning set. For p, we included the values of p with increment of 10, starting with 10. For the optimal p found, the best p variables were selected based on ranking of the BW ratio.

4.5. Linear discriminant analysis (LDA)

There are various forms of LDA. In our comparison, the R packages (*sma* for DLDA and *mass* for FLDA) are used, and the BW ratio is used in the variable selection in learning sets. As described in Section 4.4, optimal number of predictors in terms of accuracy is selected using the BW ratio in learning sets using nested CV. When the number of features exceeds the sample size, the covariance matrix in FLDA will not have full rank, and thus it cannot be inverted. A pseudo inverse is used instead of the usual matrix inverse in this case. The R function *lda* uses truncated singular value decomposition.

4.6. Shrunken centroid (SC)

The R software package (*pamr*) of SC (Tibshirani et al., 2002) with a soft thresholding option is used in our comparison. As described in Sections 4.4, and 4.5, we performed variable selection using BW ratio in learning sets for each data.

4.7. Single optimal trees

The optimal trees obtained from CART and QUEST algorithms are included in the comparison. A single CART tree is built as described in Section 3.1. We used the classification tree program included in the C-T CERP program to build an optimal CART tree. In contrast to the exhaustive search method of CART, QUEST uses a discriminant-based procedure for splitting. QUEST can use linear combination splits in order to achieve gains in classification accuracy over univariate splits. The binary executable is obtained from http://www.stat.wisc.edu/∼loh/quest.html. A shell program is developed to conduct 20 CVs using this executable program. The linear combination split option is used. It is known to perform better than the default option of univariate split in general. We used the defaults for the other options.

5. Data sets

5.1. Identification of imprinted genes

Imprinted genes give rise to numerous human diseases (Reik and Walter, 2001). These genes are unusually predisposed to causing a disease because of the silencing of expression of one of the two homologs at an imprinted locus, requiring only heterozygosity for a mutation affecting the active allele to cause complete loss of gene expression. As the pattern of silencing reflects the gametic origin of the gene, the gamete is said to imprint the locus with a memory of its origin. Greally (2002) described the first characteristic sequence parameter that discriminates imprinted regions—a paucity of short interspersed transposable elements (SINEs).

The genomic data collected to study imprinted genes were from the UCSC Genome Browser (http://genome.ucsc.edu/). Annotation data were downloaded for the human genome (hg16, July 2003 freeze). The data contain 131 samples and 1446 predictors. Among the 131 samples, 43 are imprinted and 88 are control genes (nonimprinted). The current data set has been made available by John Greally at http://greallylab.aecom.yu.edu/∼greally/imprinting_data.txt.

5.2. Classification of colon tissue samples

The DNA microarray technology has been increasingly used in cancer research, which enables classification of tissue samples based only on gene expression data, without prior and often subjective biological knowledge (Golub et al., 1999; Dudoit et al., 2002). Gene expression in 40 colon adenocarcinoma tissue samples and 22 normal colon tissue samples was analyzed with an Affymetrix oligonucleotide array complementary to more than 6500 human genes (Alon et al., 1999). The current data set contains the expression of the 2000 genes with highest minimal intensity across the 62 tissue samples. The goal here is to classify new unlabeled tissue samples as being cancerous or noncancerous. The data are available at http://microarray.princeton.edu/oncology/affydata/index.html.

5.3. Classification of chemicals for estrogen activity

A large number of environmental chemicals known as endocrine-disrupting chemicals (EDCs) are suspected of disrupting endocrine functions by mimicking or antagonizing natural hormones in animals and humans (Hileman, 1997). The NCTR (National Center for Toxicological Research) estrogen activity data set consists of 232 structurally diverse chemicals. Among these 232 samples, 131 chemicals exhibit estrogen receptor binding activity and 101 are inactive in a competitive estrogen receptor binding assay (Blair et al., 2000). These chemicals were selected a priori based on structural characteristics and tested in a well validated and standardized in vitro rat uterine cytosol ER competitive-binding assay (Blair et al., 2000; Branham et al., 2002). This structurally diverse data set has 312 predictors generated using the Molconn-Z software 4.07 (http://www.fda.gov/nctr/science/centers/toxicoinformatics/edkb/documents/erNCTR232.txt).

5.4. Classification of leukemia subtypes

The Golub leukemia data set is introduced in one of the seminal papers applying statistical classification techniques to microarray data. Golub et al. (1999) classified AML and ALL subtypes using a variant of LDA based on gene expression profiling. We have included it here again to illustrate CERP on a data set with known and validated performance using state-of-the-art classifiers such as those studied in Dudoit et al. (2002). The training set originally used in the analysis by Golub et al. included 38 samples of ALL and AML. An additional 34 samples were then used as the test set. As done in Dudoit et al., we combined the training and test sets for our analyses, therefore obtaining 47 ALL and 25 AML samples. The data were obtained through the website http://www.broad.mit.edu/cancer/software/genepattern/datasets/ and were pre-processed as described in Golub et al. such that 3571 genes were in the data set.

6. Results

We evaluated the prediction accuracy of the CERP approach as well as a balance between sensitivity and specificity using the four data sets discussed in Section 5. Before applying the methods, we removed the predictors that had identical values for more than 98% of the samples in order to reduce the possibility that a predictor in a learning set would not have distinct values in the CV for building an optimal tree. For the estrogen data, 250 out of 312 predictors were selected using this criterion, and for the gene imprinting data, 1248 out of 1446 predictors were selected for the analysis. For the colon and leukemia data sets, all the predictors were included by this criterion. The evaluation and comparison of CERP and other methods were conducted by averaging the results from 20 replications of 10-fold CV in order to achieve a stable result. Twenty CVs should be sufficient according to Molinaro et al. (2005) who recommended 10 runs of 10-fold CV to have low MSE and bias.

We compared the prediction accuracy of C-T CERP and LR-T CERP with other standard classification methods. Fig. 2 depicts the accuracy of each classification model with a 1-sd bar for the four data sets, based on the average of 20 runs of 10-fold CV. Tables 2 through 3 provide the accuracy of each method, with sensitivity and specificity for the first three data sets. For the leukemia data, the balance of sensitivity and specificities is not an issue due to almost perfect classification by most of the classification methods. Thus we do not provide a table. For the leukemia data, all the methods considered in this study except CART gave high accuracies ranging from 95% to 98%. The accuracy of CART was near 82%, while QUEST gave the accuracy of 96%. For the methods with variable selection, the average

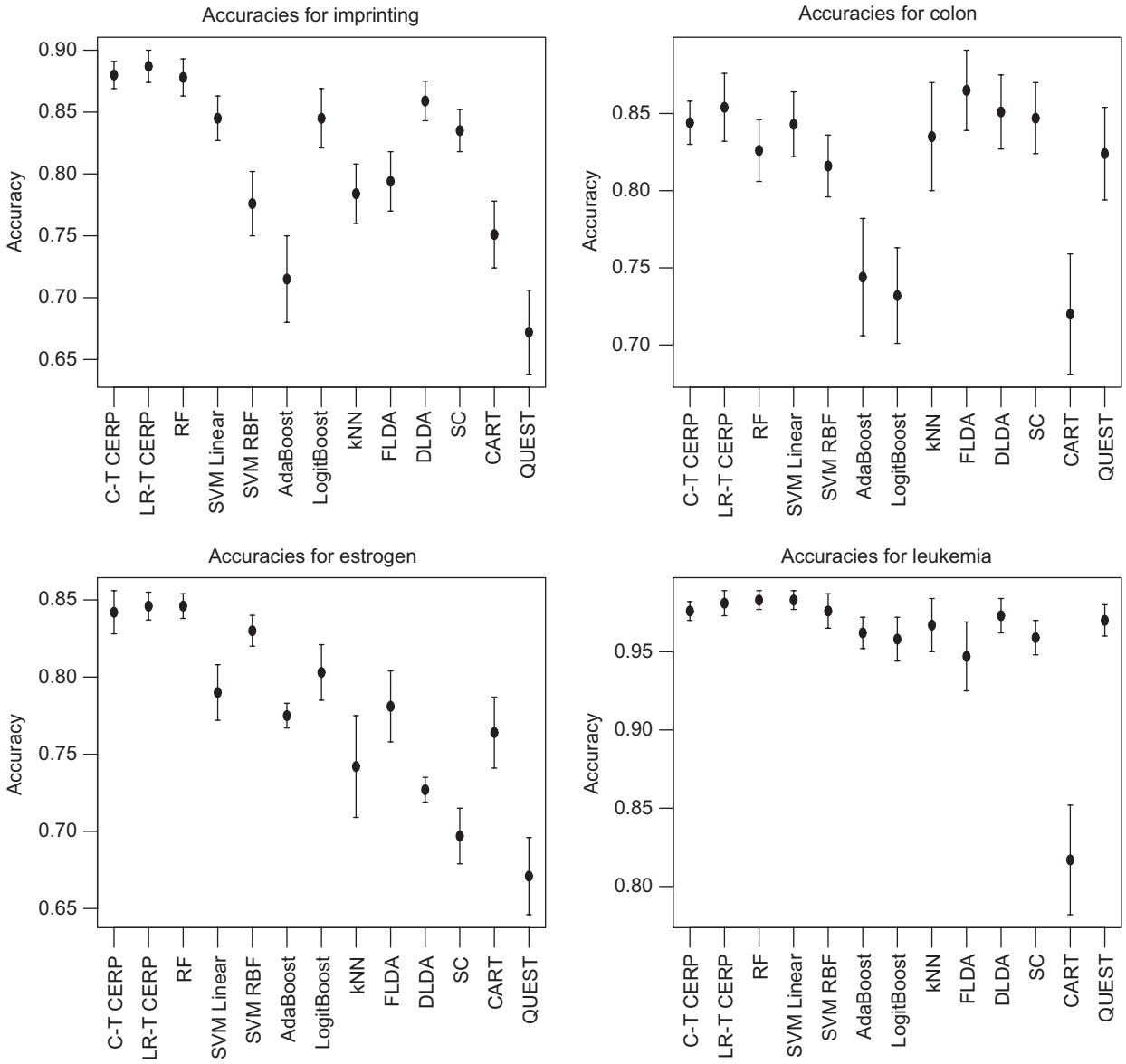

Fig. 2. Comparison of accuracies (with 1-sd bars) of classification methods for each data set based on 20 replications of 10-fold CV.

numbers (sd in parentheses) of predictor variables selected in the training phase were 197 (507) for kNN, 178 (151) for FLDA, 140 (367) for DLDA and 20 (16) for SC. For kNN, the average (sd in parentheses) value of k was 3.1 (3.3).

In the comparison of the methods, CERP did not require any fine tuning of parameters because they are determined in the training phase inside the program. For the most relevant comparison, we provide the best result we obtained for each data set for the other methods and specify the parameters used in the footnote. For the methods requiring variable pre-selection, an optimal number of variables is searched and the variables are selected by the BW ratio in the training phase. For kNN, the optimal value of k is also obtained in nested CV. Analysis of a previously published colon data set was similar to the results obtained by others using these and related methods (Alon et al., 1999; Tsai et al., 2004; Ambroise and McLachlan, 2002). Some of the standard methods included in our comparison gave higher accuracies than these previous results. This shows our effort to obtain the best possible results from these methods.

Table 2
Accuracy (sd in parentheses) of classification methods for the gene *imprinting* data with 43 cases and 88 controls

Method	Approach	#Predictors	Overall	Sensitivity	Specificity
CERP	C-T CERP[a]	All	.88 (.01)	.72 (.03)	.95 (.04)
	LR-T CERP[b]	All	.89 (.01)	.72 (.04)	.97 (.02)
RF[c]		All	.88 (.01)	.65 (.03)	.99 (.01)
SVM	Lin. kernel	All	.85 (.02)	.69 (.04)	.92 (.02)
	RBF[d]	All	.78 (.03)	.44 (.06)	.94 (.02)
Boosting	AdaBoost	All	.72 (.04)	.44 (.09)	.85 (.03)
	LogitBoost	All	.84 (.02)	.72 (.03)	.90 (.03)
kNN		261 (322)[e,f]	.78 (.02)	.67 (.05)	.84 (.03)
LDA	FLDA	40 (11)[e]	.79 (.02)	.63 (.06)	.88 (.03)
	DLDA	391 (107)[e]	.86 (.02)	.63 (.04)	.97 (.01)
SC		All	.84 (.01)	.70 (.02)	.91 (.01)
Single tree	CART	All	.75 (.03)	.77 (.06)	.74 (.04)
	QUEST	All	.67 (.03)	.37 (.11)	.82 (.04)

Twenty repetitions of 10-fold CV were performed for each method unless otherwise specified. When variable selection is done in the training phase, the mean value of the predictors (sd in parentheses) is given.

[a]Average partition size per ensemble: 77.5.

[b]Average partition size per ensemble: 71.3.

[c]Number of trees: 500; number of predictors selected in each node of a tree: default (floor[$m^{1/2}$]).

[d]Radial basis function (default option for the SVM function in the R package *e1071*).

[e]Mean (sd in parentheses) numbers of predictors selected in the training phase.

[f]Average (sd in parentheses) of k obtained in the training phase: 2.7 (1.7).

Table 2 shows that for the gene imprinting data, the sensitivity and specificity were 72% and 95%, respectively, by C-T CERP, while they were 65% and 99%, respectively, by RF. These results support the criticism about RF on the imbalance by Dudoit and Fridlyand (2003). For RF, we tried various choices for the number of variables to be selected in each node of a tree, and the number of trees including the default values. Moreover, we examined various cut-off (threshold) options in RF package of R, but the result did not substantially differ beyond the random error. Both CERP and RF performed better than the other methods in terms of accuracy for the data. For DLDA, the accuracy reached the highest when a large number of variables were pre-selected. SC obtained the best performance when all predictors were included.

For the colon data (see Table 3), LR-T CERP gave 85% of prediction accuracy. The Boosting methods did not give higher than 75% overall prediction rate for this particular data set, although they performed quite well on the other three data sets we examined. Ambroise and McLachlan (2002) compared the difference in error estimates between internal and external CVs using SVM with a combination of the linear kernel and a backward elimination gene selection procedure. They reported that all prediction accuracies were well below 85% based on external CV. Tsai et al. (2004) reported that their maximum prediction accuracy using kNN (with $k = 1$) and SVM was 84%. We obtained slightly better results than these previous work for the discriminant analysis or nearest neighbor (centroid) methods. FLDA showed a strong performance with 87% accuracy when variable selection was not performed. In our study, DLDA showed an improved accuracy when variables were pre-selected. SVM performed better when the linear kernel was used instead of the radial based function (RBF). DLDA, kNN, and SC gave relatively high accuracy when predictors were pre-selected. It is interesting that with all the variables in the model, FLDA obtained the highest accuracy, while DLDA performed poorly. It is notable that QUEST is comparable to RF for the colon and leukemia data sets, while it was considerably worse than RF for the other two data sets. For this data set, RF required fewer trees for the optimal performance compared to other data sets.

For the estrogen data (see Table 4), DLDA did not perform well even with a variable selection because it assumes that features are not correlated. FLDA performed better than DLDA with a variable selection, but it was still inferior to the top-performing methods. The performance of the aggregation methods was strong when the best options were used. Because the data are reasonably balanced (proportion of the positive responses in the data is .56), the balance of sensitivity and specificity was good in most of the methods. In SVM, RBF performed better than linear kernel unlike in the other data sets. It appears to be due to a nonlinear relationship that was not captured using a linear function of the predictors. The performance of kNN and SC was not comparable to many other methods even after a variable selection.

Table 3
Accuracy (sd in parentheses) of classification methods for the *colon* data with 22 cases and 40 controls

Method	Approach	#predictors	Overall	Sensitivity	Specificity
CERP	C-T CERP[a]	All	.84 (.01)	.87 (.02)	.80 (.03)
	LR-T CERP[b]	All	.85 (.02)	.87 (.01)	.83 (.05)
RF[c]		All	.83 (.02)	.89 (.01)	.72 (.05)
SVM	Lin. kernel	All	.84 (.02)	.87 (.02)	.80 (.05)
	RBF[d]	All	.82 (.02)	.94 (.01)	.59 (.05)
Boosting	AdaBoost	All	.74 (.04)	.83 (.04)	.59 (.08)
	LogitBoost	All	.73 (.03)	.83 (.04)	.56 (.06)
kNN		303 (397)[e,f]	.84 (.04)	.88 (.02)	.75 (.08)
LDA	FLDA	All	.87 (.02)	.89 (.02)	.85 (.05)
	DLDA	32 (25)[e]	.85 (.02)	.86 (.02)	.83 (.04)
SC		46 (117)[e]	.85 (.02)	.87 (.02)	.81 (.07)
Single tree	CART	All	.72 (.04)	.79 (.05)	.60 (.08)
	QUEST	All	.82 (.03)	.86 (.03)	.75 (.06)

Twenty repetitions of 10-fold CV were performed for each method. When variable selection is done in the training phase, the mean value of the predictors (sd in parentheses) is given.
[a]Average partition size per ensemble: 48.0.
[b]Average partition size per ensemble: 172.3.
[c]Number of trees: 200; number of predictors selected in each node of a tree: default (floor[$m^{1/2}$]).
[d]Radial basis function (default option for the SVM function in the R package *e1071*).
[e]Mean (sd in parentheses) numbers of predictors selected in the training phase.
[f]Average (sd in parentheses) of k obtained in the training phase: 5.1 (2.3).

Table 4
Accuracy (sd in parentheses) of classification methods for the *estrogen* data with 131 cases and 101 controls

Method	Approach	#Predictors	Overall	Sensitivity	Specificity
CERP	C-T CERP[a]	All	.84 (.01)	.88 (.02)	.80 (.02)
	LR-T CERP[b]	All	.85 (.01)	.89 (.01)	.79 (.02)
RF[c]		All	.85 (.01)	.88 (.01)	.80 (.01)
SVM	Lin. kernel	All	.79 (.02)	.83 (.02)	.74 (.03)
	RBF[d]	All	.83 (.01)	.89 (.01)	.75 (.02)
Boosting	AdaBoost	All	.78 (.01)	.89 (.01)	.63 (.02)
	LogitBoost	All	.80 (.02)	.84 (.02)	.76 (.03)
kNN		63 (63)[e,f]	.74 (.03)	.83 (.03)	.62 (.05)
LDA	FLDA	60 (25)[e]	.78 (.02)	.87 (.02)	.66 (.03)
	DLDA	49 (33)[e]	.73 (.01)	.76 (.02)	.69 (.02)
SC		51 (51)[e]	.70 (.02)	.75 (.03)	.63 (.02)
Single tree	CART	All	.76 (.02)	.79 (.03)	.74 (.03)
	QUEST	All	.67 (.03)	.73 (.03)	.60 (.05)

Twenty repetitions of 10-fold CV were performed for each method unless otherwise specified. When variable selection is done in the training phase, the mean value of the predictors (sd in parentheses) is given.
[a]Average partition size per ensemble: 15.5.
[b]Average partition size per ensemble: 11.1.
[c]Number of trees: 500; number of predictors selected in each node of a tree: default (floor[$m^{1/2}$]).
[d]Radial basis function (default option for the SVM function in the R package *e1071*).
[e]Mean (sd in parentheses) numbers of predictors selected in the training phase.
[f]Average (sd in parentheses) of k obtained in the training phase: 3.1 (2.2).

From these results, it is clear that CERP performs consistently well. In addition, the balance between sensitivity and specificity has not been lost, even on unbalanced data like the imprinting data set (rate of positive responses: 33%) and the colon cancer data (positive rate: 65%). The unbalanced data present a difficulty to some of the other methods under consideration, including RF.

In general, results from the Boosting methods were not as good as those from the other methods except for the single trees. We also tried a regularized version of AdaBoost such as Epsilon-Boosting, but it was computationally unstable and did not alter findings meaningfully. We observed that DLDA was comparable to all of the other standard methods considered in this paper except for the estrogen data. DLDA and kNN often require a reduced set of features for a better performance, though it can be applied to data when $N \ll m$. Although not reported, the performance of kNN, LDA and SC was poor in general when all the predictors were included in the model. A simple screening tool such as the BW ratio is unable to capture interactions in the data and variable selection via ranking may result in a set of highly correlated classifiers. Therefore, the optimal number of variables must be determined. For the gene imprinting data, many features are highly correlated, which contributes to the poor relative performance after variable screening for lower values of m. One of the strengths of CERP is that no variable screening is necessary; we may consider all of the features simultaneously in using multiple ensembles.

The performance of SVM was highly dependent on the choice of a kernel. The default option of RBF in the R SVM package gave higher accuracy than linear kernel for the estrogen data, while it gave lower accuracy for the other three data sets.

It is clear that CERP and RF show a significant improvement over the single optimal CART. The overall accuracy of QUEST is comparable to other aggregation methods for the colon and leukemia data, while it is poor for the estrogen and gene imprinting data.

When we performed CV, the run time of CERP was reasonable compared to that of RF and Boosting. For the gene imprinting data, for example, it took approximately 7 min to finish a 10-fold CV for C-T CERP with 11 ensembles, and it took approximately 4 min for RF on a Windows XP 2.8 GHz machine.

7. Discussion

We have introduced a new ensemble-based classifier called CERP. CERP is built using ensembles of classification trees (C-T CERP) and logistic regression trees (LR-T CERP) as base classifiers. In this sense, the method is not fundamentally different from RF, which is tree-based also. However, one important distinction between CERP and RF is the way the high-dimensional data are handled. CERP uses a partitioning scheme to create mutually exclusive subsets of the features, while RF randomly samples from the entire pool of features at each node. Both introduce diversity, which is necessary to produce gains by taking a consensus of many classifiers. An advantage of CERP over RF is that we can achieve a rapid improvement of the prediction accuracy by ensemble voting due to a small correlation among the trees by avoiding the overlap of each subset where the individual tree is constructed. We have shown empirically that huge data sets need not be handled as a whole; the subspaces of the feature space created through partitioning may be treated independently and separately until after the classifiers are developed. This gives CERP a huge computational advantage of tackling the growing problem of dimensionality. Like RF, CERP does not require variable pre-selection, thus it is straightforward and easy to implement the algorithm.

We also showed that CERP is comparable to conventional classification methods. The classification methods we selected for comparison performed well with no one method outperforming the others consistently. Using a CV estimate, we achieved a consistently high accuracy using CERP. This high accuracy was achieved due to the diversity created between classifiers.

According to our study, RF is comparable to other existing classification methods in view of accuracy. However, it has been shown to perform poorly in instances of class imbalance (Segal, 2004). Dudoit and Fridlyand (2003) noted that RF had difficulties with unbalanced class sizes and almost always predicted the majority class. Chen et al. (2004) stated that most commonly used classification algorithms do not work well for the imbalance of sensitivity and specificity because they aim to minimize the overall error rate, rather than paying special attention to the minority class. We also found in the study on the gene imprinting data that RF gives a poor sensitivity, while it gives almost perfect specificity. This imbalance is not desirable because a goal here is to identify more positives (imprinted genes). CERP appears to perform well in balancing specificity and sensitivity in unbalanced data sets. The main reason is that the base trees of CERP are optimal trees, while those of RF are fully grown trees without pruning. In LR-T CERP, the optimal threshold choice helps improve the balance. We are mainly interested in showing the enhanced accuracy in this paper, but the better balance is a strong attribute of CERP as well. Further exploration of the performance of CERP with respect to imbalance will be done in future work. Among the classification and discrimination methods considered in this study, LDA, kNN, and SC showed a robust balance of the sensitivity and specificity with a reasonable accuracy. However, a

drawback of these methods is that they often require a pre-selection of predictors for a good performance. A variable selection can be complex and time consuming.

Since all the parameters are determined in the training phase of the program, CERP did not require any fine tuning for specific data sets in the comparison. Thus it can be used for any type of high-dimensional data set. Although RF tends to perform well with the default parameter values, the performance may depend on the number of classifiers or number of randomly selected predictors in each node of a tree. When using the RBF kernel for SVMs, a fine tuning of the relevant parameters such as kernel width is needed. Often a default does not work very well for a large number of attributes.

An advantage to combining decision trees is that we are able to use highly flexible models. After averaging the predictions across base classifiers, we are able to retain the interactions and nonlinear components found in each of the trees. However, CERP need not necessarily be tree-based. Conceptually, any type of base classifier can be used. This will be a topic of future research.

Both CERP and RF show significant improvement over CART. Although CERP appears to perform consistently well on data sets from different areas, a few issues remain to be investigated. An obvious cost of this or any ensemble approach is the "black box" lack of interpretability of the resulting classifier. However, as the microarray literature has shown, the final model is hardly "final" and variable importance must be treated cautiously given the computational burden of weighing each variable's contribution among all combinations of other variables. For predictive purposes, it is most important to generate an accurate and robust model, with descriptive aspects left as a separate exercise and not generated as a by-product of one representative model. We will explore variable importance in future studies.

Acknowledgment

Hongshik Ahn's research was partially supported by the Faculty Research Participation Program at the NCTR administered by the Oak Ridge Institute for Science and Education through an interagency agreement between USDOE and USFDA.

References

Ahn, H., Chen, J.J., 1997. Tree-structured logistic model for over-dispersed binomial data with application to modeling developmental effects. Biometrics 53, 435–455.

Akaike, H., 1974. A new look at the statistical identification model. IEEE Trans. Automat. Control 19, 716–723.

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

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

Blair, R., Fang, H., Branham, W.S., Hass, B., Dial, S.L., Moland, C.L., Tong, W., Shi, L., Perkins, R., Sheehan, D.M., 2000. Estrogen receptor relative binding affinities of 188 natural and xenochemicals: structural diversity of ligands. Toxicological Sci. 54, 138–153.

Branham, W.S., Dial, S.L., Moland, C.L., Hass, B.S., Blair, R.M., Fang, H., Shi, L., Tong, W., Perkins, R., Sheehan, D.M., 2002. Phytoestrogens and mycoestrogens bind to the rat uterine estrogen receptor. J. Nutrition 132, 658–664.

Breiman, L., 1996. Bagging predictors. Mach. Learning 24, 123–140.

Breiman, L., 1998. Arcing classifiers. Ann. Statist. 26, 801–849.

Breiman, L., 2001. Random forest. Mach. Learning 45, 5–32.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and Regression Trees. Wadsworth, Belmont, CA.

Burnham, K.P., Anderson, D.R., 2002. Model Selection and Multinomial Inference: A Practical Information-Theoretic Approach. second ed. Springer, New York.

Chen, C., Liaw, A., Breiman, L., 2004. Using random forest to learn imbalanced data. Technical Report #666, Department of Statistics, University of California, Berkeley.

Chen, J.J., Tsai, C.A., Young, J.F., Kodell, R.L., 2005. Classification ensembles for unbalanced class sizes in predictive toxicology. SAR and QSAR Environmental Res. 16, 517–529.

Dettling, M., Buhlman, P., 2003. Boosting for tumor classification with gene expression data. Bioinformatics 19, 1061–1069.

Domingos, P., 1999. Metacost: a general method for making classifiers cost-sensitive. In: Proceedings of the 5th SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM Press, San Diego, CA, pp. 155–164.

Dudoit, S., Fridlyand, J., 2003. Classification in microarray experiments. In: Speed, T.P. (Ed.), Statistical Analysis of Gene Expression Microarray Data. Chapman & Hall, CRC Press, Boca Raton, FL. (Chapter 3).

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

Duin, R.P.W., Tax, D.M.J., 2000. Experiments with classifier combining rules. In: Kittler, J., Roli, F. (Eds.), Lecture Notes in Computer Science, vol. 1857. Springer, Berlin, pp. 16–29.

Freund, Y., Schapire, R., 1996. Experiments with a new boosting algorithm. In: Machine Learning: Proceedings of the Thirteenth International Conference.pp. 148–156.

Freund, Y., Schapire, R., 1997. A decision-theoretic generalization of online learning and an application to boosting. J. Comput. System Sci. 55, 119–139.

Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D., Lander, E.S., 1999. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286 (5439), 531–537.

Greally, J.M., 2002. Short interspersed transposable elements (SINEs) are excluded from imprinted regions in the human genome. Proc. Nat. Acad. Sci. 99, 327–332.

Hansen, L.K., Salamon, P., 1990. Neural network ensembles. IEEE Trans. Pattern Anal. Mach. Intelligence 12, 993–1001.

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

Hileman, B., 1997. Hormone disrupter research expands. Chem. Eng. News 75, 24–25.

Ho, T.K., 1998. The random subspace method for constructing decision forests. IEEE Trans. Pattern Anal. Mach. Intelligence 20, 832–844.

Kuncheva, L.I., Whitaker, C.J., Shipp, C.A., Duin, R.P.W., 2003. Limits on the majority vote accuracy in classifier fusion. Pattern Anal. Appl. 6, 22–31.

Lam, L., Suen, C.Y., 1997. Application of majority voting to pattern recognition: an analysis of its behavior and performance. IEEE Trans. Systems Man Cybernet. 27, 553–568.

Loh, W.-Y., Shih, Y.S., 1997. Split selection methods for classification trees. Statist. Sinica 7, 815–840.

Michiels, S., Koscielny, S., Hill, C., 2005. Prediction of cancer outcome with microarrays: a multiple random validation strategy. Lancet 365 (9458), 488–492.

Miller, A., 2002. Subset Selection in Regression. second ed. Chapman & Hall, CRC, Los Angeles, CA.

Molinaro, A.M., Simon, R., Pfeiffer, R.M., 2005. Prediction error estimation: a comparison of resampling methods. Bioinformatics 21, 3301–3307.

Pazzani, M., Merz, C., Murphy, P., Ali, K., Hume, T., Brunk, C., 1994. Reducing misclassification costs: knowledge-intensive approaches to learning from noisy data. In: Proceedings of the 11th International Conference on Machine Learning, New Brunswick, NJ, ML-94, pp. 217–225.

Prentice, R.L., 1986. Binary regression using an extended beta-binomial distribution, with discussion of correlation induced by covariate measurement errors. J. Amer. Statist. Assoc. 81, 321–327.

Reik, W., Walter, J., 2001. Genomic imprinting: parental influence on the genome. Natur. Rev. Genetics 2 (1), 21–32.

Schapire, R.E., 1990. The strength of weak learnability. Mach. Learning 5, 197–227.

Segal, M.R., 2004. Center for bioinformatics and molecular biostatistics. Technical Report, University of California, San Francisco.

Sparano, J.A., Fazzari, M.J., Childs, G., 2005. Clinical application of molecular profiling in breast cancer. Future Oncology 1, 485–496.

Therneau, T.M., Atkinson, E.J., 1997. An introduction to recursive partitioning using the RPART routines. Technical Report, Mayo Foundation.

Tibshirani, R., Hastie, T., Narasimhan, B., Chu, G., 2002. Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc. Nat. Acad. Sci. 99, 6567–6572.

Tsai, C.A., Chen, C.H., Lee, T.C., Ho, I.C., Yang, U.C., Chen, J.J., 2004. Gene selection for sample classifications in microarray experiments. DNA Cell Biology 23, 607–614.

Vapnik, V., 1995. The Nature of Statistical Learning Theory. Springer, New York.

Williams, D.A., 1975. The analysis of binary responses from toxicological experiments involving reproduction and teratogenicity. Biometrics 31, 949–952.