

1 **Ensemble methods for classification of patients for personalized**
2 **medicine with high-dimensional data**

3
4 Hojin Moon¹, Hongshik Ahn², Ralph L. Kodell¹, Songjoon Baek¹, Chien-Ju Lin¹,
5 Taewon Lee¹ and James J. Chen¹

6
7 ¹Division of Biometry and Risk Assessment
8 National Center for Toxicological Research, FDA, Jefferson, AR 72079

9
10 ²Department of Applied Mathematics and Statistics
11 Stony Brook University, Stony Brook, NY 11794-3600
12
13

14
15
16
17
18 **Corresponding author:**

19
20 **Hojin Moon, Ph.D.**
21 Mathematical Statistician
22 Division of Biometry and Risk Assessment
23 National Center for Toxicological Research
24 U. S. Food and Drug Administration
25 3900 NCTR Road, HFT-20
26 Jefferson, AR 72079
27 Tel: 870-543-7931
28 Fax: 870-543-7662
29 Email: hojin.moon@fda.hhs.gov

1 **Abstract**

2 **Motivation:** Personalized medicine is defined by the use of genomic signatures of
3 patients in a target population for assignment of more effective therapies as well as better
4 diagnosis and earlier interventions that might prevent or delay disease. Classification
5 algorithms can be used for prediction of response to therapy to help individualize clinical
6 assignment of treatment. The algorithms are required to be highly accurate for optimal
7 treatment on each patient. Typically, there are numerous genomic and clinical variables
8 over a relatively small number of patients, which presents challenges for most traditional
9 classification algorithms to avoid over-fitting the data. We developed a robust
10 classification algorithm for high-dimensional data based on ensembles of classifiers built
11 from the optimal number of random partitions of the feature space. The software is
12 available on request from the authors.

13 **Results:** The proposed algorithm is applied to genomic data sets on lymphoma patients
14 and lung cancer patients to distinguish disease subtypes for optimal treatment and to
15 genomic data on breast cancer patients to identify patients most likely to benefit from
16 adjuvant chemotherapy after surgery. The performance of the proposed algorithm is
17 consistently good compared to the other classification algorithms. The predictive
18 accuracy can be improved by adding some relevant demographic, clinical and/or
19 histopathological measurements to the genomic data.

20

21 **Key words:** Class prediction; Cross-validation; Ensembles; Majority voting; Risk
22 profiling

23

1 **1 Introduction**

2 Providing guidance on specific therapies for pathologically distinct tumor types to
3 maximize efficacy and minimize toxicity is important for cancer treatment [1,2]. For
4 clinically heterogeneous diffuse large B-cell lymphoma (DLBCL), there exist two
5 molecularly distinct forms of DLBCL: germinal centre B-like DLBCL and activated B-
6 like DLBCL. Patients with germinal center B-like DLBCL have significantly better
7 overall survival than those with activated B-like DLBCL [3]. Consequently, they may
8 require less aggressive chemotherapy. For tumors of the lung, the pathological distinction
9 between malignant pleural mesothelioma (MPM) and adenocarcinoma (ADCA) may be
10 troublesome. Early MPM is best treated with extrapleural pneumonectomy followed by
11 chemoradiation, whereas ADCA is treated with chemotherapy alone [4]. Thus, accurate
12 classification of tumor samples and right treatment for distinct tumor types are essential
13 for efficient cancer treatment and prolonged survival on a target population of patients.

14 Microarray technology has been increasingly used in cancer research, because of
15 its potential for classification of tissue samples based only on gene expression data,
16 without acquiring prior and often subjective biological knowledge [1,5,6]. Microarrays
17 are simply ordered sets of DNA molecules of known sequence. With DNA microarray
18 technology, one can simultaneously measure expression profiles for thousands of genes
19 in tissue samples. Much research involving microarray data analysis is focused on
20 distinguishing between different cancer types using gene expression profiles from disease
21 samples, thereby allowing more accurate diagnosis and effective treatment of each patient.

22 Gene expression data might also be used to improve disease prognosis in order to
23 prevent some patients from having to undergo painful unsuccessful therapies and

1 unnecessary toxicity. For example, adjuvant chemotherapy for breast cancer after surgery
2 could reduce the risk of distant metastases; however, seventy to eighty percent of patients
3 receiving this treatment would be expected to survive metastasis-free without it [7]. Gene
4 expression profiles of sporadic breast cancers could be used to predict metastases better
5 than clinical and histopathological prognostic factors including tumor grade, tumor size,
6 angioinvasion, age and estrogen receptor. The strongest predictors for metastases such as
7 lymph node status and histological grade fail to classify accurately breast tumors
8 according to their clinical behavior [7,8].

9 Classification algorithms can be used to process high-dimensional genomic data
10 to distinguish disease subtypes and to predict response to therapy in order to help
11 individualize clinical assignment of treatment. Class prediction is a supervised learning
12 method where the algorithm learns from a training set (known samples) and establishes a
13 prediction rule to classify a test set (new samples). Development of a class prediction
14 algorithm generally consists of selection of predictors and fitting a prediction model to
15 develop the classification rule using training samples. Some classification algorithms
16 such as the classification tree or stepwise logistic regression perform these
17 simultaneously. Sensitivity (SN), specificity (SP) and accuracy, as well as positive
18 predictive value (PPV) and negative predictive value (NPV) are primary criteria used in
19 the evaluation of the performance of a classification algorithm. The SN is the proportion
20 of correct positive classifications out of the number of true positives. The SP is the
21 proportion of correct negative classifications out of the number of true negatives. The
22 accuracy is the total number of correct classifications out of the total number of samples.
23 The PPV is the probability that a patient is positive given a positive prediction. Its

1 complement, 1-PPV, is the false discovery rate (FDR). The NPV is the probability that a
2 patient is negative given a negative prediction. Algorithms with high SN and high SP as
3 well as high PPV and high NPV, which will have high accuracy, are obviously desirable.

4 Recently an ensemble-based classification algorithm, Classification by Ensembles
5 from Random Partitions (CERP), has been developed for high-dimensional data [9]. An
6 ensemble of classifiers can form a superior classifier even though individual classifiers
7 might be somewhat weak and error-prone in making decisions [10]. Moreover, an
8 ensemble of ensembles can further enhance class prediction [9]. In this paper, we propose
9 Classification-Tree CERP (C-T CERP), an ensemble of ensembles of optimal numbers of
10 pruned classification trees based on the Classification and Regression Trees (CART) [11]
11 algorithm. We relax the constraint of a fixed number of classifiers in an ensemble [9] and
12 derive the optimal number of classifiers from an adaptive bisection algorithm using
13 nested cross-validation. Individual classifiers in an ensemble are constructed from
14 randomly partitioned mutually exclusive subsets of the entire predictor space. Our
15 adaptive bisection method can be generalized to any classification algorithm to find an
16 optimal number of classifiers in an ensemble.

17 The performance of C-T CERP is compared to other well-known classification
18 algorithms: Random Forest (RF) [12], Boosting [13,14,15], Decision Forest (DF) [16],
19 Support Vector Machine (SVM) [17], Diagonal Linear Discriminant Analysis (DLDA)
20 [6], Shrunken Centroids (SC) [18], CART, Classification Rule with Unbiased Interaction
21 Selection and Estimation (CRUISE) [19], and Quick, Unbiased and Efficient Statistical
22 Tree (QUEST) [20].

1 C-T CERP utilizes a group of optimal trees from totally randomized parameter
2 spaces based on mutually exclusive subsets of the predictors. On the other hand, RF takes
3 bootstrap samples for each tree and randomly selects predictors from the entire set of
4 predictors at each node. Boosting is a general method for reducing the error of any
5 learning algorithm by a weighted majority vote of the outputs of the weak classifiers.
6 AdaBoost [13] fits an additive model in a base learner by optimizing an exponential loss
7 function. Similarly, LogitBoost [14] fits additive logistic regression models by taking the
8 binomial log-likelihood as a loss function. DF uses an averaging scheme to build an
9 ensemble of trees (not necessarily optimal) from mutually exclusive subsets of the
10 available entire parameter space in a sequential manner. C-T CERP, RF, Boosting and
11 DF are ensemble classifiers. SVM is a kernel-based machine learning approach, which
12 exploits information about the inner products in some feature space. DLDA is a
13 classification rule based on a linear discriminant function. DLDA is sometimes called
14 naïve Bayes because it is originated from a Bayesian setting, where the predicted class is
15 the one with maximum posterior probability [6]. SC is based on an enhancement of the
16 simple nearest centroid classifier. CART, CRUISE and QUEST are single optimal trees.
17 Among these single-tree algorithms, CART and QUEST yield binary trees whereas
18 CRUISE yields multiway splits.

19 The proposed algorithm is applied to three published data sets relevant to
20 personalized medicine. The algorithm is first used for the prediction of lymphoma
21 subtypes based on gene-expression in B-cell malignancies among DLBCL patients [3].
22 Similarly, it is employed on gene-expression data to distinguish MPM from ADCA of the
23 lung in order to identify the treatment that would result in the best possible outcome [4].

1 Our algorithm is then used to predict which breast cancer patients would benefit from
2 adjuvant chemotherapy after surgery based on gene-expression data [7]. We also
3 investigate if addition of seven more demographic, clinical and/or histopathological
4 variables (e.g., age, tumor size, tumor grade, angioinvasion, estrogen receptor,
5 progesterone receptor and lymphocytic infiltrate) to the high-dimensional genomic data
6 on breast cancer patients enhances classification accuracy. The performance of the
7 classification algorithm is assessed by twenty replications of 10-fold cross-validation
8 (CV).

9

10 **2 Methods**

11 The classification problem is to predict the class label Y , based on the gene expression
12 profile X , by constructing a classifier

$$13 \quad C : X \mapsto C(X),$$

14 using a training set such that the misclassification risk $P(C(X) \neq Y)$ is as small as
15 possible. When the dimension of gene expression profile m is much smaller than the
16 sample size n , a Bayes classifier or logistic regression can be employed for such a
17 problem. However, it is a well-understood phenomenon that a prediction model built
18 from thousands of available predictor variables and a relatively small sample size can be
19 quite unstable [21]. We propose a promising classification algorithm based on ensembles
20 of classifiers from random partitions of the feature space.

21

22 **2.1 Ensemble method to enhance prediction accuracy**

1 Let C_i be a random variable indicating a classification by the i -th independent classifier,
 2 where $C_i = 1$ if the classification is correct and $C_i = 0$ if not. We let p be the prediction
 3 accuracy of each classifier. Then the C_i are Bernoulli(p), and the number of accurate
 4 classifications by the ensemble majority vote is $Y = \sum_{i=1}^r C_i$, which is Binomial(r, p). Let
 5 $r = 2j + 1$, where j is a nonnegative integer. We define the prediction accuracy of the
 6 ensemble by majority voting as $A_r = P(Y \geq j+1)$. Then the prediction accuracy of the
 7 ensemble can be obtained using the standard binomial probability,
 8 $A_r = \sum_{i=j+1}^r \binom{r}{i} p^i (1-p)^{r-i}$. Lam and Suen [22] showed that the majority vote is
 9 guaranteed to give a higher accuracy than an individual classifier when the individual
 10 classifiers have an accuracy greater than 0.5. Recently, Ahn *et al.* [9] showed that the
 11 prediction accuracy of the ensemble voting method converges fast to 1 when given
 12 prediction accuracy of each individual classifier is close to 1, while it converges slowly to
 13 1 if the accuracy of individual classifier is slightly larger than 0.5.

14 In practice, the classifiers may be correlated to a certain degree. When classifiers
 15 are positively correlated, they tend to produce the same prediction outcomes. Kuncheva
 16 *et al.* [23] relaxed the restriction that the classifiers be independent. When the classifiers
 17 in the ensemble are positively correlated, we use the beta-binomial model [24,25] in
 18 order to illustrate correlation effects on the theoretical prediction accuracy by the
 19 majority vote. The beta-binomial is commonly used to model positively correlated binary
 20 variables. Given individual prediction accuracy μ and correlation ρ , the prediction
 21 accuracy of the ensemble A_r can be obtained using the probability mass function of the
 22 beta-binomial given as

1
$$P(Y = y) = \binom{r}{y} \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(y + \alpha)\Gamma(r - y + \beta)}{\Gamma(r + \alpha + \beta)},$$

2 where $\alpha = \mu(1 - \rho) / \rho$; $\beta = \alpha(1 - \mu) / \mu$.

3 Figure 1 illustrates the theoretical prediction accuracy obtained by majority voting
 4 at a given $\mu = .7$. The figure explains that independent classifiers improve the prediction
 5 accuracy more rapidly than correlated classifiers in an ensemble. For example, when the
 6 prediction accuracy of each base classifier is 70%, the class prediction accuracy by the
 7 majority vote in an ensemble reaches nearly 100% with $r = 101$ independent classifiers.
 8 On the other hand, the accuracy of the majority vote reaches only 89.1% and 77.1% with
 9 $r = 101$ positively correlated classifiers when the correlation ρ is equal to .1 and .3,
 10 respectively. These theoretical results imply that the prediction accuracy obtained by the
 11 majority vote will increase by adding more classifiers. However, if the classifiers are
 12 highly positively correlated, the addition will not help much to increase the prediction
 13 accuracy.

14 As Breiman [12] stated, when the classifiers are positively correlated, the
 15 generalization error is loosely upper bounded by $\bar{\rho}(1 - \bar{\mu}^2) / \bar{\mu}^2$, which is a function of
 16 overall correlation among trees in an ensemble ($\bar{\rho}$) and overall accuracy of trees in an
 17 ensemble ($\bar{\mu}$). CERP uses random partitioning to create mutually exclusive subsets of
 18 the features to induce diversity. If the number of partitions is larger, the prediction
 19 accuracy of the individual classifier would be lower. To compensate this loss, new
 20 ensembles are added. On the other hand, when the classifiers are negatively correlated,
 21 the prediction accuracy improves more rapidly than with independent classifiers. Ahn *et*

1 *al.* [9] reported enhancement of the prediction accuracy by ensemble majority voting of
2 negatively correlated classifiers.

3

4 **2.2 C-T CERP**

5 A schematic diagram of an ensemble of C-T CERP is shown in Figure 2. The
6 optimal number of partitions (r) in an ensemble for a given training set is obtained such
7 that the highest accuracy is achieved under a random partition of the feature space. This
8 is accomplished via a nested 10-fold cross-validation with our adaptive bisection
9 algorithm to reduce heavy computing time. The adaptive bisection method is illustrated
10 in Figure 3. This figure shows an example of how the prediction accuracy is improving
11 only to a certain point with a limited number of partitions (trees) in the feature space. It is
12 mainly because smaller feature subspaces along with excessive partitions would result in
13 lower prediction accuracy of each individual classifier. This illustrates the importance of
14 finding the optimal number of partitions in an ensemble. This figure demonstrates a case
15 where a conventional bisection method fails, shown with a dashed line, while our
16 adaptive bisection method can succeed to obtain the maximum (\blacklozenge), shown with a solid
17 line. Algorithms of C-T CERP and the adaptive bisection method are described in Tables
18 1 and 2.

19 Predictor variables in the data set are then randomly subdivided into r mutually
20 exclusive subsets. Using the i -th subset of predictors, a tree is constructed under the Gini
21 diversity index measure [11] defined by $i(s) = 1 - \sum_k p^2(k|s)$, where $p(k|s)$ is a
22 probability that a sample is in class k given that it falls into a node s . We let u and v be
23 the left and right subnodes, respectively, of a parent node s , in which a split $\theta^* \in \Theta$ of

1 node s maximizes a goodness-of-split function. A goodness-of-split criterion in this study
 2 is chosen such that the split θ^* at node s maximizes the reduction of the impurity,
 3 $\Delta i(\theta, s) = i(s) - [p_u i(u) + p_v i(v)]$, where a split θ of node s sends a proportion p_u of data
 4 cases in node s to u and a proportion p_v to v . This tree construction process for growing a
 5 large initial tree continues splitting the samples until either each terminal node is pure
 6 (i.e., the node cases are all in one class) or the total number of samples in a node is less
 7 than or equal to 5. To avoid over-fitting, the optimal trees in C-T CERP are obtained by
 8 employing the minimal cost-complexity pruning algorithm used in CART. In the pruning
 9 process, a nested sequence of subtrees is obtained by progressively deleting branches.
 10 This results in a decreasing sequence of subtrees in terms of tree complexity. One of
 11 these subtrees is selected as an optimal tree if a subtree produces a minimal internal
 12 cross-validated misclassification error within 1-SE in the pruning. We define a subtree T_{j_0}
 13 that has the minimum misclassification rate via cross-validation in pruning by
 14 $\hat{R}(T_{j_0}) = \min_j R^{\text{CV}}(T_j)$, where nested trees $T_j < T_{\text{max}}$. The final tree selected is then T_{j^*} ,
 15 where the index j^* is the maximum j satisfying $\hat{R}(T_{j^*}) \leq \hat{R}(T_{j_0}) + \text{SE}(\hat{R}(T_{j_0}))$. The
 16 standard error is calculated as

$$17 \quad \text{SE}(\hat{R}(T_{j_0})) = \sqrt{s^2 / N}; \text{ and } s^2 = \frac{1}{N} \sum_{i \in \Gamma} [C(i | k) - \hat{R}(T_{j_0})]^2,$$

18 where $C(i | k)$ is the cost of classifying a case in class k into class i and Γ is an index set
 19 of the cases from the entire sample with size N .

20 In C-T CERP, we employ majority voting among trees within individual
 21 ensembles and then among ensembles. In an ensemble, optimal trees are generated using
 22 a random partition of the feature space according to the optimal number of partitions of

1 the training set. New ensembles are then created by randomly re-partitioning the feature
2 space. In our preliminary study [9], most of the improvement in adding ensembles was
3 achieved by the first few ensembles, and then the improvement was slowed down as
4 more ensembles were added. Hence, we fix the default number of ensembles as 15 in this
5 study. For each observation, final ensemble prediction is based on the majority vote
6 across these ensembles. In a majority voting, the predicted values are classified as either
7 0 or 1 using a natural threshold 0.5 by each base classifier. C-T CERP is implemented in
8 C/C++.

9 A package (RandomForest) in R is used for the RF algorithm. The number of
10 trees is generated using the default of $n\text{tree} = 500$. The number of features selected at
11 each node in a tree is chosen using the default value of $\text{floor}(m^{1/2})$, where m is the total
12 number of features. Among many boosting methods, AdaBoost [13] and LogitBoost [14]
13 are adopted using a package (boost) in R with a default option. On the other hand, an
14 executable program is downloaded from [www.fda.gov/nctr/science/centers/
15 toxicoinformatics/DecisionForest/forest.zip](http://www.fda.gov/nctr/science/centers/toxicoinformatics/DecisionForest/forest.zip) for DF. For SVM, a package (e1071) in R is
16 applied with linear kernel. According to our preliminary study [9] SVM with linear
17 kernel is generally better than SVM with radial basis kernel for high-dimensional data. A
18 package (sma) in R is employed for DLDA with a default option. SC is implemented with
19 a package (pamr) in R with a soft thresholding option as a default. For single optimal
20 trees, CART is implemented with a package (rpart) in R with a default option. On the
21 other hand, executable files are downloaded from www.stat.wisc.edu/~loh/, and
22 implemented in R for CRUISE and QUEST. These choices of parameters were the best
23 choices for applications illustrated in this study.

1 We evaluated the prediction accuracy, the balance between SN and SP, and the
2 balance between PPV and NPV of the classification algorithms using 10-fold cross-
3 validation (CV). In many cases, the number of features (m) is much greater than the
4 number of patients (n). In such a case, a method for validly estimating prediction
5 accuracy is using CV. CV utilizes resampling without replacement of the entire data to
6 repeatedly develop classifiers on a training set and evaluates classifiers on a separate test
7 set, and then averages the procedure over the resamplings. We averaged the results from
8 20 replications of 10-fold CV in order to achieve a stable result. Twenty CVs should be
9 sufficient according to Molinaro *et al.* [26] who recommended 10 trials of 10-fold CV to
10 have low mean squared error and bias.

11

12 **3 Results**

13 This section presents the performance (accuracy, SN, SP, PPV, NPV) of C-T CERP
14 along with other various well-known classification algorithms using three published high-
15 dimensional microarray data sets for personalized medicine.

16

17 **3.1 Classification of lymphoma subtypes**

18 A target of cancer treatment is to diagnose accurately and to assign individualized
19 therapy for distinct tumor types. Despite the variety of clinical, morphological and
20 molecular parameters used to classify human malignancies, patients receiving the same
21 diagnosis can have markedly different clinical courses and treatment responses. DLBCL
22 is an aggressive malignancy of mature B lymphocytes with 25,000 cases of annual
23 incidence. Patients with DLBCL have highly variable clinical courses. Although most

1 patients respond initially to chemotherapy, fewer than half of the patients achieve a
2 durable remission [3,27]. There are two distinct types of DLBCL on the bases of
3 differentially expressed genes within the B-cell lineage: germinal centre B-like DLBCL
4 and activated B-like DLBCL. It is important to distinguish these two lymphoma subtypes
5 to maximize efficacy and minimize toxicity on chemotherapy responsiveness.

6 The data set consists of 47 samples, 24 of them are from germinal centre B-like
7 DLBCL, while 23 are activated B-like DLBCL. The positive rate of this data set is 48.9%
8 by considering activated B-like DLBCL as a positive. Each sample is described by 4,026
9 genes. We note that there exist missing values in this data set. A 10-nearest neighbor
10 averaging method [28] was employed to impute the missing data. The data set with 4,026
11 genes and 47 samples is publicly available at <http://llmpp.nih.gov/lymphoma/>.

12 Table 3 shows performance of classification algorithms for the lymphoma data.
13 C-T CERP, RF and SC algorithms gave less than 5% error rate (1-accuracy). Among
14 them, C-T CERP showed the lowest cross-validated error rate of 3.2% (RF: 4.3%; SC:
15 4.4%). The balance between Sensitivity and specificity of C-T CERP, RF, AdaBoost,
16 SVM, DLDA and SC was excellent; all sensitivities and specificities were above 90%.
17 The PPV and NPV of C-T CERP, RF, AdaBoost, SVM, DLDA and SC were also all
18 higher than 90%. The false discovery rates (1-PPV) of C-T CERP, RF, AdaBoost,
19 LogitBoost, SVM, DLDA, SC and QUEST were all less than 10%. Among them, the
20 FDR of C-T CERP was the lowest as 5.7% (RF: 6.5%; AdaBoost: 9.0%; LogitBoost:
21 7.1%; SVM: 8.0%; DLDA: 9.6%; SC: 7.1%; QUEST: 9.6%). C-T CERP performs
22 slightly better than RF and SC, and better than the other classification algorithms. Among

1 single optimal trees, CRUISE and QUEST performed better than CART. QUEST gave
2 the lowest FDR among single optimal trees.

3

4 **3.2 Classification of two tumor classes of the lung**

5 Accurate diagnosis of lung disease is critical in choosing individual treatment of a group
6 of patients. The pathological distinction between MPM and ADCA of the lung is
7 generally onerous from both clinical and pathological perspectives [4]. Upon accurate
8 distinction of highly lethal MPM from ADCA, treatment therapy, either extrapleural
9 pneumonectomy followed by chemoradiation or chemotherapy alone, can be correctly
10 assigned to a patient to achieve the best possible outcome.

11 The data set consists of 181 tissue samples (31 MPM and 150 ADCA) and each
12 sample is described by 12,533 genes. The positive rate of this data set is 17.1% by
13 considering MPM as a positive [29]. In order to filter out noise from unexpressed or
14 lowly expressed genes, the ratio of between-group to within-group sums of squares (BW
15 ratio [6]) is used to reduce the set of 12,533 genes to 5,000 genes having the largest BW
16 ratio. We let N be the number of samples and N_k be the samples in class k . The indicator
17 function $I(y_i = k)$ is equal to 1 if a class label y_i is equal to a class k . Otherwise, it is equal
18 to 0. Then the BW ration, $BW(j)$, is defined as

19
$$BW(j) = \frac{\sum_i \sum_k I(y_i = k)(\bar{x}_{kj} - \bar{x}_{\cdot j})^2}{\sum_i \sum_k I(y_i = k)(x_{ij} - \bar{x}_{kj})^2}, \text{ where } \bar{x}_{\cdot j} = \frac{\sum_i x_{ij}}{N}; \bar{x}_{kj} = \frac{\sum I(y_i = k) x_{ij}}{N_k},$$

20 for sample i and gene j . Ambroise and McLachlan [30] warned of selection bias when the
21 cross-validation is performed incorrectly based on a small pre-selected set (e.g., <400) of
22 genes. Our pre-filtering of unexpressed and lowly expressed genes to reduce noise is

1 unlikely to result in appreciable bias, as we retain 5,000 genes. In our external 10-fold
2 cross-validations, we allow all classification algorithms to select from this same set of
3 5,000 genes so that our comparisons are unbiased. This data set is available at
4 www.chestsurg.org/publications/2002-microarray.aspx.

5 Figure 4 illustrates performance of classification algorithms for the two tumor
6 classes on the lung. All algorithms gave less than 5% error rate (1-accuracy) except
7 CART. Among them, C-T CERP, RF, AdaBoost, LogitBoost, SVM, DLDA and SC
8 achieved less than 1% error rate (Accuracy: 99.5% for C-T CERP; 99.4% for RF,
9 AdaBoost, LogitBoost, SVM and DLDA; 99.8% for SC). This level of accuracy shows
10 the real potential for confident clinical assignment of therapies on an individual patient
11 basis. Although the data set is lopsided, the balance between SN and SP of all algorithms
12 was excellent except DF and single optimal trees. The PPV and NPV (not shown) of all
13 classification algorithms were higher than 90% except CART. The false discovery rates
14 of all classification methods were less than 1% except DF, SVM and single optimal trees.
15 Among single optimal trees, CRUISE and QUEST performed better than CART. QUEST
16 showed similar performance to DF which is more complex ensemble averaging method,
17 and maintained the lowest FDR (5%) among single optimal trees.

18

19 **3.3 Breast cancer classification**

20 The objective of this study was to use gene expression data to identify breast cancer
21 patients who might benefit from adjuvant chemotherapy according to classification of
22 prognostication with distant metastases. Patients who developed distant metastases within
23 5 years were categorized as ‘poor prognosis’, whereas patients who continued to be

1 disease-free after a period of at least 5 years were categorized as ‘good prognosis’. The
2 data contains 78 primary breast cancers (34 from patients in poor prognosis and 44 from
3 patients in good prognosis). These samples have been selected from patients who were
4 lymph node negative and under 55 years of age at diagnosis. The positive rate of this data
5 set is 43.6% by considering poor prognosis as a positive. Out of approximately 25,000
6 genes interrogated for expression levels, about 5,000 significantly regulated genes were
7 pre-filtered in the same fashion illustrated in van ‘t Veer *et al.* [7]. In addition, seven
8 relevant demographic/clinical/histopathological predictors were added to this gene
9 expression feature space to investigate if the addition of these variables improves the
10 prediction accuracy compared to genomic data only. These data sets were used for the
11 evaluation of all classification algorithms via the external 10-fold cross-validation. The
12 genomic data sets for breast cancer are publicly available at
13 www.rii.com/publications/2002/vantveer.html.

14 Tables 4 and 5 show performance of classification algorithms for the breast
15 cancer genomic data and genomic plus clinical, demographical and histopathological data,
16 respectively, based on 20 repetitions of 10-fold CV. When 7 more clinical variables are
17 added to the gene expression data, overall prediction accuracies appear to be slightly
18 improved compared to accuracies from genomic data only in almost all classification
19 algorithms considered. This is mainly due to an improvement in sensitivity. Still, the
20 overall accuracy is somewhat low for all the classifiers. The highest accuracies are
21 achieved by CT-CERP and LogitBoost (about 65%). The balance between SN and SP is
22 reasonably good for CT-CERP, LogitBoost, DLDA and SC and their sensitivities are
23 higher (> 50%) than the rest (< 50%). The highest positive predictive values (the lowest

1 FDRs) are achieved by CT-CERP (62%) and LogitBoost (61%). Among single optimal
2 trees, accuracies of CRUISE and QUEST are slightly higher than CART ($> 55\%$).
3 However, the balance between SN and SP in these single trees is unsatisfactory.

4 The basic idea in this application is that a person with a predicted good prognosis
5 based on a gene-expression profile might not want to be subjected to adjuvant
6 chemotherapy and its deleterious side effects. Current rule-based decision rules would
7 place almost all patients on adjuvant chemotherapy [7]. Our classification algorithm
8 could provide a choice to patients before taking chemotherapy so that most of the good-
9 prognosis patients who would not necessarily benefit from chemotherapy could elect out.
10 The positive and negative predictivities are not necessarily high enough for those data to
11 make it to clinical practice, but our algorithm and LogitBoost show slightly better
12 performance compared to the best known other classification algorithms.

13

14 **4 CONCLUSION AND DISCUSSION**

15 Recent advancements in biotechnology have accelerated research on the development of
16 molecular biomarkers for the diagnosis and treatment of disease. The Food and Drug
17 Administration envisions clinical tests to identify patients most likely to benefit from
18 particular drugs and patients most likely to experience adverse reactions [31]. Such
19 patient profiling will enable assignment of drug therapies on a scientifically sound
20 predictive basis rather than on an empirical trial-and-error basis. Because medicine is
21 always practiced on individuals not populations, the goal is to individualize therapies
22 from a population-based approach to an individualized approach.

1 We have presented statistical classification algorithms to accurately classify
2 patients into risk/benefit categories using high-dimensional genomic and other data.
3 Classification algorithms were illustrated by three published data sets and the new C-T
4 CERP was compared to the best known published classification procedures. The C-T
5 CERP algorithm classifies subjects based on a majority vote of individual members of
6 each ensemble and then a majority vote among ensembles. Because of its partitioning of
7 the predictor space, CERP can overcome the problem of having fewer samples than
8 predictors. Based on cross-validated results for several high-dimensional data sets, C-T
9 CERP is a consistently one of the best classification algorithms and maintains a good
10 balance between sensitivity and specificity with lower false discovery rate even when
11 sample sizes between classes are unbalanced.

12 In one application, lymphoma patients were classified as having either germinal
13 center diffuse large B-cell lymphoma (GC) or activated diffuse large B-cell lymphoma
14 (ACT) based on each individual patient's gene-expression profile. The distinction is
15 important because the chemotherapy regimens for two subtypes are very different, and
16 incorrect treatment assignment has both efficacy and toxicity consequences. Although
17 remissions can be achieved using GC therapy for ACT (and vice versa), cure rates are
18 markedly diminished, and unwarranted toxicities are encountered. Classification
19 algorithms are essential for the realization of personalized medicine in this application,
20 because distinguishing GC and ACT based on only clinical and morphological
21 parameters is difficult. Our algorithm correctly classified patients with the highest cross-
22 validated accuracy (96.8%) among the classification procedures we considered. This

1 level of accuracy shows the real potential for confident clinical assignment of therapies
2 on an individual patient basis.

3 Classification algorithms were also used to distinguish between malignant pleural
4 mesothelioma (MPM) and adenocarcinoma (ADCA) of the lung using high-dimensional
5 gene expression data, based on 20 repetitions of 10-fold CV. Highly accurate
6 classification between MPM and ADCA would be critical for choosing between
7 extrapleural pneumonectomy followed by chemoradiation and chemotherapy alone to
8 obtain the best possible outcome. The cross-validated results for this data set indicated
9 that highly accurate classifications were made using CT-CERP, RF, AdaBoost,
10 LogitBoost, DLDA and SC with an error rate less than 1%. Such accuracy would be good
11 enough for clinical practice.

12 In the other application, gene-expression profiles of post-surgery breast cancer
13 patients were used to classify these patients as having either a high or low likelihood of
14 developing distant metastasis within five years. If this were brought into clinical
15 application, a patient with a confidently predicted good prognosis might want to elect out
16 of adjuvant chemotherapy and its associated debilitating side effects. With current rule-
17 based decisions, almost all patients are subject to chemotherapy. When just a few
18 demographic, clinical and histopathological measures traditionally used for treatment
19 assignment were added to the numerous genomic predictors, the prediction accuracy
20 appeared to be enhanced further. According to the theory underlying our algorithm, the
21 more individual patient information that is used, whatever the source or type, the greater
22 is the likelihood that the prediction accuracy will increase. Thus, it is anticipated that the
23 combined use of multiple biomarkers on individual patients, including high-dimensional

1 proteomic and metabolomic profiles, could improve the prediction accuracy of data like
2 the present genomic data to a level suitable for clinical practice.

3 The C-T CERP algorithm appears to have good potential for biomedical decision
4 making in assignment of patients to treatment therapies based on individual risk factors
5 and disease characteristics (personalized medicine). Other biomedical applications can be
6 early detection and prediction of disease, evaluation and validation of biomarkers for
7 efficacy and toxicity, and identification of individuals with risk factors for specific health
8 consequences related to diet (personalized nutrition).

9

10 **ACKNOWLEDGEMENTS**

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

14

15 **REFERENCES**

- 16 [1] Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, et al.
17 Molecular classification of cancer: class discovery and class prediction by gene
18 expression monitoring. *Science* 286 (1999) 531-537.
- 19 [2] Zhang H, Yu C-Y, Singer B, Xiong M. Recursive partitioning for tumor classification
20 with gene expression microarray data. *Proc. Natl Acad. Sci. USA* 98 (2001) 6730-
21 6735.

- 1 [3] Alizadeh AA, Elsen MB, Davis ER, Ma C, Lossos IS, Rosenwald A, et al. Distinct
2 types of diffuse large B-cell lymphoma identified by gene expression profiling.
3 *Nature* 403 (2000) 503-511.
- 4 [4] Gordon GJ, Jensen RV, Hsiao L-L, Gullans SR, Blumenstock JE, Ramaswamy S, et
5 al. Translation of microarray data into clinically relevant cancer diagnostic tests using
6 gene expression ratios in lung cancer and mesothelioma. *Cancer Res.* 62 (2002) 4963-
7 4967.
- 8 [5] Alexandridis R, Lin S, Irwin M. Class discovery and classification of tumor samples
9 using mixture modeling of gene expression data – a unified approach. *Bioinformatics*
10 20 (2004) 2545-2552.
- 11 [6] Dudoit S, Fridlyand J, Speed TP. Comparison of discrimination methods for the
12 classification of tumors using gene expression data. *J. Am. Stat. Assoc.* 97 (2002) 77-
13 87.
- 14 [7] van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AAM, Mao M., et al. Gene
15 expression profiling predicts clinical outcome of breast cancer. *Nature* 415 (2002)
16 530-536.
- 17 [8] McGuire WL. Breast cancer prognostic factors: evaluation guidelines. *J. Natl Cancer*
18 *I.* 83 (1991) 154-155.
- 19 [9] Ahn H, Moon H, Fazzari MJ, Lim N, Chen JJ, Kodell RL. Classification by
20 ensembles from random partitions. Technical Report SUNYSB-AMS-06-03, Stony
21 Brook University, NY, 2006.
- 22 [10] Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data*
23 *Mining, Inference, and Prediction* (Springer, New York, 2001).

- 1 [11] Breiman L, Friedman JH, Olshen RA, Stone CJ. *Classification and Regression*
2 *Trees* (Wadsworth, California, 1984).
- 3 [12] Breiman L. Random forest. *Mach. Learn.* 45 (2001) 5-32.
- 4 [13] Freund Y, Schapire R. A decision-theoretic generalization of online learning and an
5 application to boosting. *J. Comput. Syst. Sci.* 55 (1997) 119-139.
- 6 [14] Friedman J, Hastie T, Tibshirani R. Additive logistic regression: a statistical view of
7 boosting. *Ann. Stat.* 28 (2000) 337-374.
- 8 [15] Schapire R. The strength of weak learnability. *Mach. Learn.* 5 (1990) 197-227.
- 9 [16] Tong W, Hong H, Fang H, Xie Q, Perkins R. Decision forest: combining the
10 predictions of multiple independent decision tree models. *J. Chem. Inf. Comp. Sci.*
11 43 (2003) 525-531.
- 12 [17] Vapnik V. *The Nature of Statistical Learning Theory* (Springer, New York, 1995).
- 13 [18] Tibshirani R, Hastie T, Narasimhan B, Chu G. Diagnosis of multiple cancer types
14 by shrunken centroids of gene expression. *Proc. Natl Acad. Sci. USA* 99 (2002)
15 6567-6572.
- 16 [19] Kim H, Loh W-Y. Classification trees with unbiased multiway splits. *J. Am. Stat.*
17 *Assoc.* 96 (2001) 589-604.
- 18 [20] Loh W-Y, Shih Y-S. Split selection methods for classification trees. *Stat. Sinica* 7
19 (1997) 815-840.
- 20 [21] Miller A. *Subset selection in regression- 2nd ed.* (Chapman and Hall/CRC, Los
21 Angeles, CA, 2002).
- 22 [22] Lam L, Suen CY. Application of majority voting to pattern recognition: An analysis
23 of its behavior and performance. *IEEE T. Syst. Man Cy. A* 27 (1997) 553-568.

- 1 [23] Kuncheva LI, Whitaker CJ, Shipp CA, Duin RPW. Limits on the majority vote
2 accuracy in classifier fusion. *Pattern Anal. Appl.* 6 (2003) 22-31.
- 3 [24] Ahn H, Chen JJ. Tree-structured logistic regression model for over-dispersed
4 binomial data with application to modeling developmental effects. *Biometrics* 53
5 (1997) 435-455.
- 6 [25] Williams DA. The analysis of binary responses from toxicological experiments
7 involving reproduction and teratogenicity. *Biometrics* 31 (1975) 949-952.
- 8 [26] Molinaro AM, Simon R, Pfeiffer RM. Prediction error estimation: A comparison of
9 resampling methods. *Bioinformatics* 21 (2005) 3301-3307.
- 10 [27] Vose JM. Current approaches to the management of non-Hodgkin's lymphoma.
11 *Semin. Oncol.* 25 (1998) 483-491.
- 12 [28] Hastie T, Tibshirani R, Sherlock G, Eisen M, Brown P, Botstein D. Imputing
13 missing data for gene expression arrays. Stanford University Statistics Department
14 Technical report, Stanford University, CA, 1999.
- 15 [29] Tan AC, Gilbert D. Ensemble machine learning on gene expression data for cancer
16 classification. *Appl. Bioinformatics* 2 (2003) S75-S83.
- 17 [30] Ambroise C, McLachlan G. Selection bias in gene extraction on the basis of
18 microarray gene-expression data. *Proc. Natl Acad. Sci. USA* 99 (2002) 6562-6566.
- 19 [31] Ridge JR. Reimbursement and coverage challenges associated with bringing
20 emerging molecular diagnostics into the personalized medicine paradigm.
21 *Personalized Medicine* 3(3) (2006) 345-348.

22
23

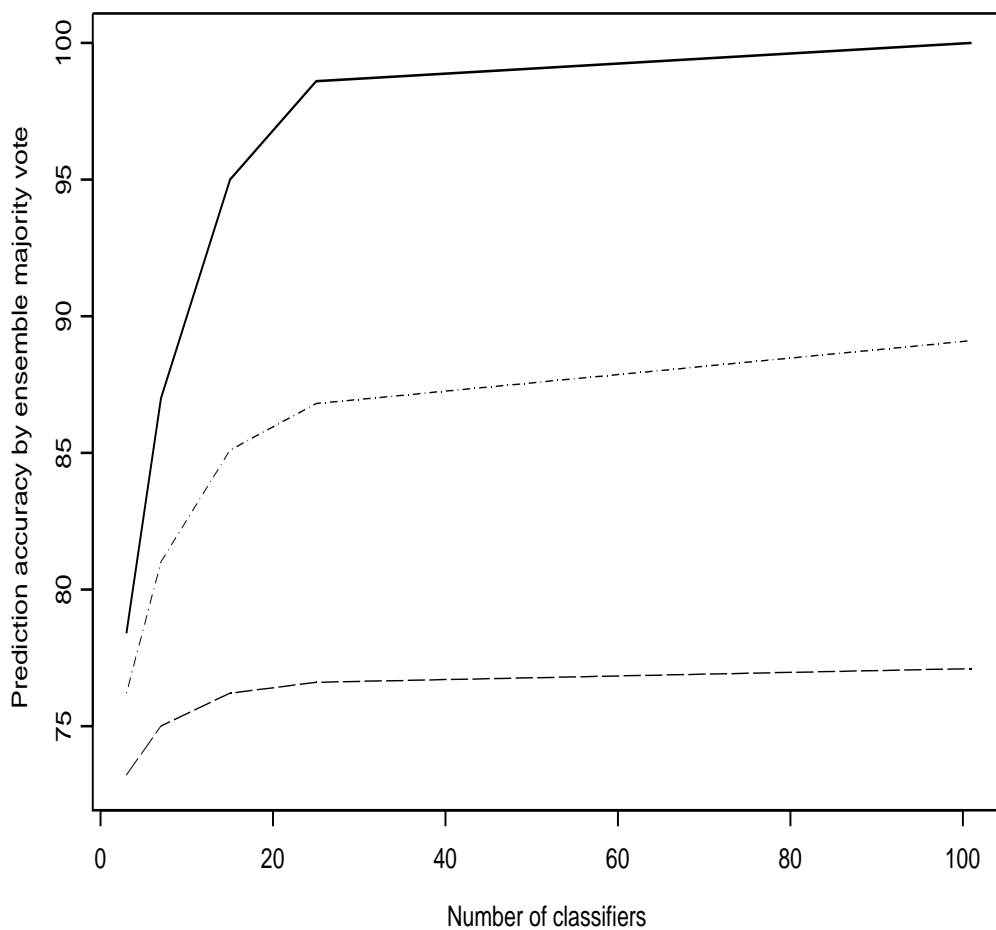


Figure 1 Enhancement of prediction accuracy (%) by majority voting when $\mu = 0.7$: solid line indicates $\rho = 0$; dotdash line indicates $\rho = 0.1$; dashed line indicates $\rho = 0.3$.

- 1
- 2
- 3
- 4
- 5
- 6

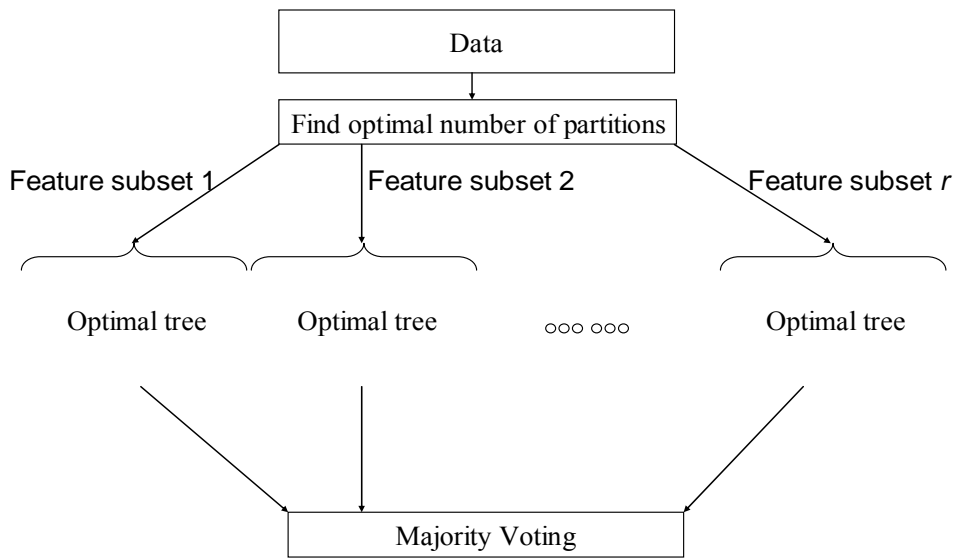
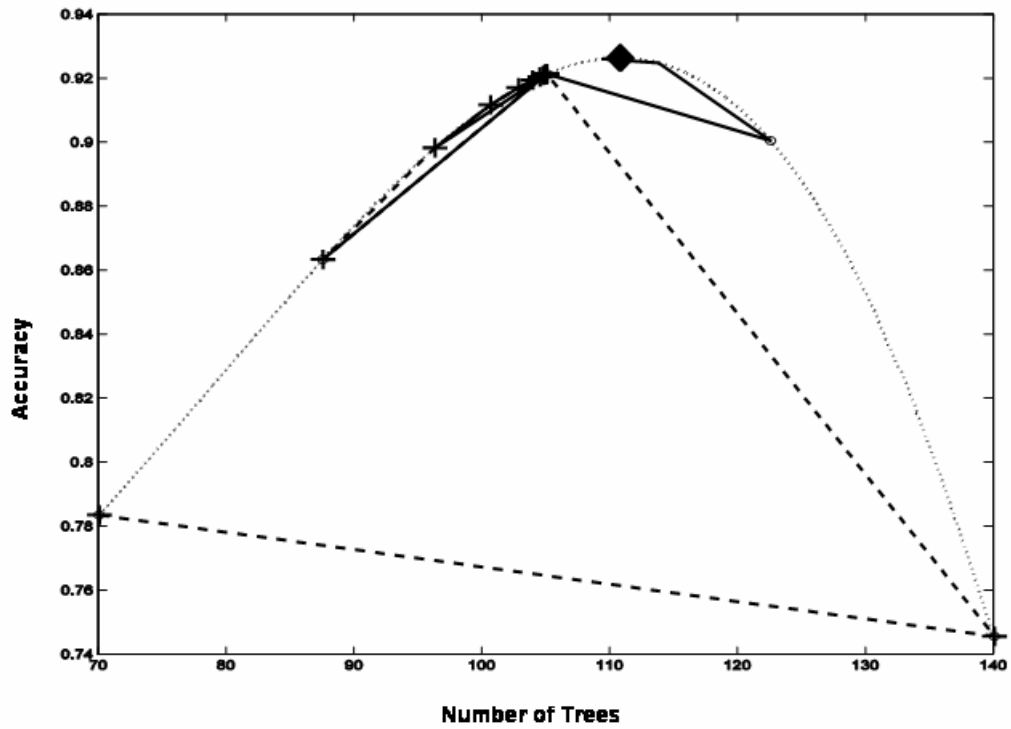


Figure 2 Schematic of an ensemble of C-T CERP

- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8
- 9
- 10
- 11
- 12
- 13



1

2

Figure 3 Our adaptive bisection method (solid line) versus a conventional bisection method (dashed line); ♦ indicates maximum

3

4

5

6

7

8

9

10

11

12

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23

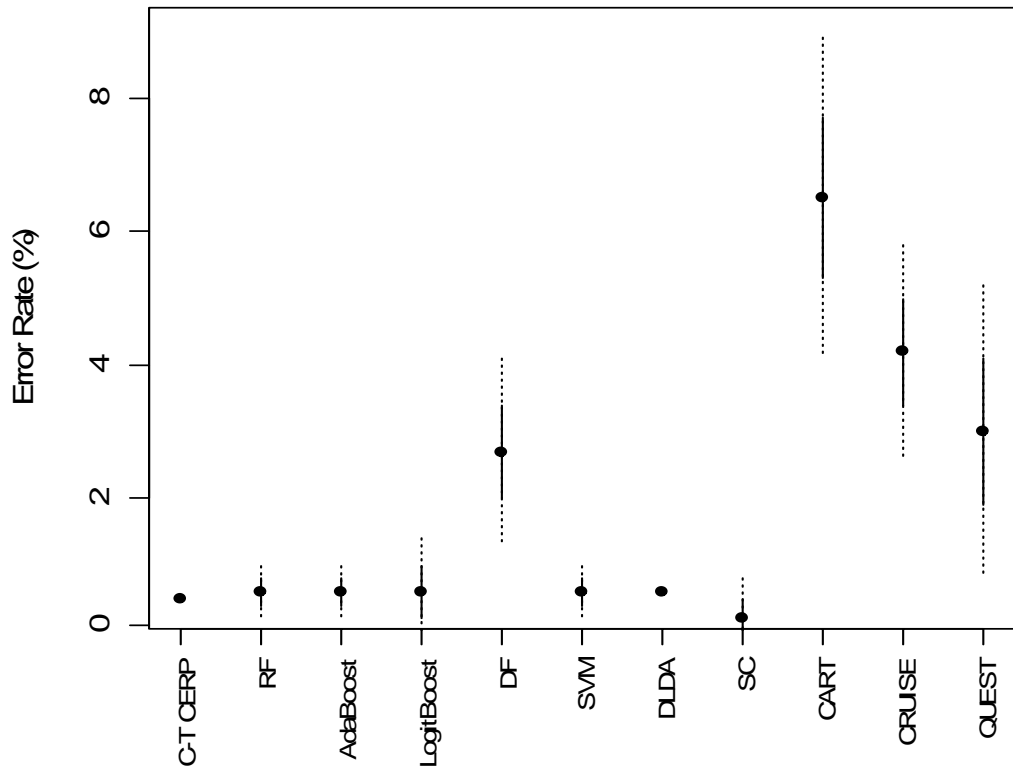


Figure 4 Cross-validated accuracy of classification algorithms for two tumor classes on the lung based on 20 replications of 10-fold CV (solid line indicates within 1 sd and dotted line shows within 2 sd)

Table 1 Adaptive bisection algorithm (ABA)

Input: N_{max} : the maximum number of trees

N_0 : the minimum number of trees

m : the number of checking points between N_0 and N_{max}

Define $f(n)$ be the accuracy of an ensemble classifier which has n trees.

Let $\{n_i\}$ be the set of checking points including N_0 and N_{max} , where

$$n_i = \text{NearestOdd}(N_0 + (i - 1)/(m - 1) * (N_{max} - N_0)), 1 \leq i \leq m$$

Find $i^* = \text{argmax}_{1 \leq i \leq m} f(n_i)$

If $i^* = 1$, set $n^* = \text{bisection}(f, n_1, n_2)$

If $i^* = m$, set $n^* = \text{bisection}(f, n_{m-1}, n_m)$

Otherwise, find $n_1^* = \text{bisection}(f, n_{i^*-1}, n_{i^*})$ and $n_2^* = \text{bisection}(f, n_{i^*}, n_{i^*+1})$

Set $n^* = \text{argmax}(f(n_1^*), n_2^*)$

- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8
- 9
- 10
- 11

Table 2 Classification by Ensembles from Random Partitions (CERP)

Input: Training set $\langle \mathbf{x}, \mathbf{y} \rangle$

Repeat

 Shuffle the predictors $x_i \in R^m$ in the feature space

 Find the optimal number of classifiers (r) via ABA

 Partition the feature space into r subspaces

 For each iteration $j = 1 \dots r$ { Implement an optimal tree $C_j(\mathbf{x})$ }

 Take majority vote on an ensemble

Until the number of ensembles

Take majority vote across ensembles

1
2
3
4
5
6
7
8
9
10
11

Table 3 Performance (sd in parentheses) of classification algorithms for the **lymphoma data** based on 20 repetitions of 10-fold CV

Algorithms	Accuracy	Sensitivity	Specificity	PPV ¹	NPV ²
C-T CERP	.968 (.015)	.996 (.013)	.942 (.025)	.943 (.023)	.996 (.013)
RF	.957 (.021)	.983 (.026)	.933 (.031)	.935 (.028)	.983 (.025)
AdaBoost	.931 (.021)	.954 (.026)	.908 (.037)	.910 (.032)	.955 (.025)
LogitBoost	.900 (.036)	.863 (.071)	.935 (.037)	.929 (.038)	.881 (.055)
DF	.905 (.018)	.933 (.036)	.879 (.027)	.882 (.021)	.933 (.031)
SVM	.931 (.024)	.941 (.043)	.921 (.030)	.920 (.027)	.944 (.039)
DLDA	.926 (.022)	.950 (.016)	.902 (.039)	.904 (.035)	.950 (.015)
SC	.956 (.011)	.987 (.020)	.927 (.019)	.929 (.017)	.987 (.020)
CART	.816 (.050)	.804 (.080)	.827 (.053)	.818 (.051)	.819 (.064)
CRUISE	.888 (.040)	.907 (.060)	.871 (.049)	.872 (.042)	.909 (.052)
QUEST	.887 (.043)	.867 (.097)	.906 (.060)	.904 (.054)	.885 (.070)

¹Positive Predictive Value;

²Negative Predictive Value

1
2
3
4
5
6
7

Table 4 Performance (sd in parentheses) of classification algorithms for the **breast cancer genomic data** based on 20 repetitions of 10-fold CV

Algorithms	Accuracy	Sensitivity	Specificity	PPV ¹	NPV ²
C-T CERP	.653 (.021)	.543 (.039)	.738 (.036)	.616 (.031)	.676 (.018)
RF	.625 (.019)	.468 (.032)	.747 (.032)	.589 (.029)	.645 (.014)
AdaBoost	.588 (.041)	.321 (.089)	.794 (.069)	.550 (.094)	.603 (.028)
LogitBoost	.652 (.049)	.556 (.084)	.726 (.061)	.611 (.067)	.680 (.043)
DF	.596 (.028)	.497 (.050)	.672 (.055)	.542 (.040)	.634 (.022)
SVM	.565 (.029)	.396 (.053)	.697 (.027)	.501 (.042)	.599 (.025)
DLDA	.625 (.019)	.524 (.023)	.703 (.026)	.578 (.026)	.656 (.015)
SC	.609 (.019)	.506 (.026)	.689 (.023)	.557 (.024)	.643 (.016)
CART	.546 (.028)	.175 (.058)	.832 (.047)	.446 (.084)	.566 (.018)
CRUISE	.551 (.048)	.215 (.100)	.810 (.059)	.456 (.112)	.573 (.034)
QUEST	.565 (.044)	.228 (.080)	.826 (.077)	.510 (.117)	.581 (.027)

¹Positive Predictive Value;

²Negative Predictive Value

1
2
3
4
5
6
7

Table 5 Performance (sd in parentheses) of classification algorithms for the **breast cancer genomic and clinical, demographical and histopathological data** based on 20 repetitions of 10-fold CV

Algorithms	Accuracy	Sensitivity	Specificity	PPV ¹	NPV ²
C-T CERP	.656 (.028)	.553 (.041)	.741 (.047)	.619 (.039)	.687 (.033)
RF	.630 (.023)	.482 (.034)	.744 (.034)	.594 (.034)	.651 (.016)
AdaBoost	.619 (.045)	.387 (.090)	.798 (.065)	.599 (.085)	.628 (.034)
LogitBoost	.656 (.040)	.568 (.049)	.725 (.063)	.618 (.054)	.684 (.033)
DF	.597 (.021)	.490 (.048)	.680 (.045)	.543 (.029)	.633 (.017)
SVM	.574 (.027)	.403 (.044)	.707 (.037)	.515 (.040)	.605 (.021)
DLDA	.629 (.017)	.526 (.025)	.709 (.027)	.584 (.023)	.660 (.013)
SC	.622 (.018)	.538 (.025)	.688 (.018)	.571 (.022)	.658 (.016)
CART	.547 (.031)	.216 (.096)	.803 (.063)	.443 (.103)	.572 (.022)
CRUISE	.575 (.047)	.240 (.100)	.834 (.063)	.519 (.120)	.588 (.032)
QUEST	.563 (.036)	.218 (.062)	.831 (.071)	.507 (.082)	.578 (.021)

¹Positive Predictive Value;

²Negative Predictive Value