NIH Public Access
Author Manuscript
Artif Intell Med. Author manuscript; available in PMC 2013 May 29.

Published in final edited form as:
Artif Intell Med. 2012 March ; 54(3): 171–179. doi:10.1016/j.artmed.2011.10.003.

Selective voting in convex-hull ensembles improves classification accuracy

Ralph L. Kodell[a,*], **Chuanlei Zhang**[a], **Eric R. Siegel**[a], and **Radhakrishnan Nagarajan**[b]

[a]Department of Biostatistics, #781, University of Arkansas for Medical Sciences, 4301 W. Markham St., Little Rock, AR 72205, United States

[b]Division of Biomedical Informatics, #782, University of Arkansas for Medical Sciences, 4301 W. Markham St., Little Rock, AR 72205, United States

Abstract

Objective—Classification algorithms can be used to predict risks and responses of patients based on genomic and other high-dimensional data. While there is optimism for using these algorithms to improve the treatment of diseases, they have yet to demonstrate sufficient predictive ability for routine clinical practice. They generally classify all patients according to the same criteria, under an implicit assumption of population homogeneity. The objective here is to allow for population heterogeneity, possibly unrecognized, in order to increase classification accuracy and further the goal of tailoring therapies on an individualized basis.

Methods and materials—Anew selective-voting algorithm is developed in the context of a classifier ensemble of two-dimensional convex hulls of positive and negative training samples. Individual classifiers in the ensemble are allowed to *vote* on test samples only if those samples are located within or behind pruned convex hulls of training samples that define the classifiers.

Results—Validation of the new algorithm's increased accuracy is carried out using two publicly available datasets having cancer as the outcome variable and expression levels of thousands of genes as predictors. Selective voting leads to statistically significant increases in accuracy from 86.0% to 89.8% ($p < 0.001$) and 63.2% to 67.8% ($p < 0.003$) compared to the original algorithm.

Conclusion—Selective voting by members of convex-hull classifier ensembles significantly increases classification accuracy compared to one-size-fits-all approaches.

Keywords

Cross-validation; Genomic prediction; Cancer screening; Individualized therapy

1. Introduction

Advancements in biotechnology in recent years have increased the availability of high-dimensional 'omic data for biomedical decision making. For such data to be informative for patient care, it must be transformed from simply a mass of raw data on each patient to a higher level of relevant electronic knowledge. Statistical learning techniques have been used to develop computational algorithms that can process such high-dimensional data to classify unknown tissue or blood samples through supervised training on samples of known class. The primary goal of these algorithms is to improve the assignment of therapies to patients in the treatment of disease, by either maximizing efficacy with respect to the intended

© 2011 Elsevier B.V. All rights reserved.

*Corresponding author. Tel.: +1 501 686 5353; fax: +1 501 526 6729. rlkodell@uams.edu, kodell.r@att.net (R.L. Kodell).

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

beneficial effect or minimizing the risk of adverse side effects. Applications of class prediction based on gene expression that are of special interest in this paper are those that specialize in classifying which tissue samples are cancerous and which are cancer-free [1] or predicting which cancer patients could benefit from chemotherapy versus which would likely only experience toxic side effects [2]. Determining the requirements for validation of prediction models and diagnostic procedures for clinical management of patients with therapeutic agents is the focus of much current effort by pharmaceutical companies and the Food and Drug Administration that regulates those companies [3].

In assessing the performance of a classification algorithm, the goal is to estimate its ability to generalize, i.e., to predict the outcomes of samples not included in the dataset used to train the classifier. Performance may be assessed on the basis of a number of different indices. For problems having a dichotomous outcome variable, the focus of the present paper, sensitivity (SEN), specificity (SPC), positive predictive value (PPV) and negative predictive value (NPV) are indices that may be of interest in addition to prediction accuracy (ACC) [4]. ACC is the overall proportion of correct predictions, SEN the proportion of correct predictions among true positives, SPC the proportion of correct predictions among true negatives, PPV the proportion of correct predictions among positive predictions and NPV the proportion of correct predictions among negative predictions.

Ensembles of classifiers have been shown to achieve higher accuracy than individual classifiers under certain conditions [5]. Many methods have been proposed for combining the outputs of classifiers in an ensemble, including methods for combining class labels and methods for combining continuous outputs [6]. Various alternatives have been proposed for ensemble selection, including selecting ensemble members from among many models of the same type as well as from among many models of different types [7–9]. A comprehensive review of ensemble-based methods is provided in [10]. Majority voting among ensemble members is a common approach for combining class labels to predict the class of an unknown sample [11]. Two well-known methods that employ majority voting are bagging and boosting [12,13]. The random forest method [14] is a popular ensemble classifier that employs the bagging method, while logitboost [15] is a popular method of boosting. Recently, an alternative to bagging and boosting, classification by ensembles from random partitions (CERP), designed specifically for tree-based classifiers [4,16], has been shown to be very competitive. Other notable methods of classification based on high-dimensional feature set partitioning include the random subspace method [17], attribute bagging [18] and the genetic algorithm-based method [19]. While all of these methods show high accuracy when classifying cases not included in the training set, they tend to follow the standard “one-size-fits-all” approach to predicting class membership for unknowns, in that every member of the trained ensemble of classifiers votes on every test sample to arrive at a prediction. However, it is generally recognized that not all patients are alike, even if they may be members of the same positive or negative class; populations of patients are heterogeneous rather than homogeneous. Allowing for population heterogeneity in a classification procedure should enhance the accuracy of class prediction. Specifically, a strategy that selects only a subset of classifiers in an ensemble to vote on a given patient can be expected to improve prediction accuracy and, ultimately, to lead to more representative subsets of predictor variables for each patient. The development and validation of such a selective-voting strategy is the goal of this paper.

Recently, Kodell et al. [20] introduced a model-free, convex-hull-based approach to ensemble building and showed it to be among the best of several methods studied. With the convex-hull approach, individual members of the ensemble are selected from all possible two-dimensional convex hulls of positive and negative cases constructed from the ${}_pC_2$ possible pairs of predictor variables, where ${}_pC_2$ is the binomial coefficient “p choose 2”and

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

p is the number of predictor variables from which to choose combinations of 2 predictors. A nearest-neighbor criterion determines the class of each training and test point. The upper portions of Tables 1 and 2 show that the convex-hull ensemble classifier performed competitively compared to several well-known classification procedures on the colon cancer data of Alon et al. [1] and the breast cancer data of van't Veer et al. [2], respectively. The tables show results for convex hull [20], classification-tree (C-T) CERP [4], logistic regression-tree (LR-T) CERP [16], random forest (RF) [14], adaboost [13], logitboost [15]k-nearest neighbor (k-NN) [21], shrunken centroid [22], support vector machine with linear kernel (SVM-linear) [23], diagonal linear discriminant analysis (DLDA) [24] and Fisher's linear discriminant analysis (FLDA) [25].

The two-dimensional geometry of the convex-hull ensemble [20] exploits second-order interactions among the predictor variables [26] while being robust to the curse of dimensionality [27], in that the prediction space needs to be populated with training-set points in only two dimensions at a time. In addition, the algorithm does not require that all samples, whether training or test cases, have complete data on all predictor variables, which sets it apart from most other classification procedures. The two-dimensional convex-hull ensemble's mathematically and statistically robust structure provides an excellent platform for a new, natural selective-voting strategy that allows for population heterogeneity, to move away from the standard one-size-fits-all approach. Such a new convex-hull strategy that uses selective voting is described in this paper.

2. Methods

2.1. Developing a selective-voting algorithm to increase classification accuracy

2.1.1. Convex-hull pruning to facilitate selective voting—In order to employ selective voting in the convex-hull ensemble, each potential member of the ensemble, that is, each pair of positive and negative convex hulls defined by each pair of predictor variables, is first pruned to produce reduced positive and negative convex hulls that do not overlap. The objective is to achieve separation of sets of positive and negative *training* samples to increase the voting accuracy of individual ensemble members on *test* samples.

Fig. 1a–d, based on two-dimensional convex hulls for cancer samples and non-cancer samples formed by a selected pair of genes from Alon et al. [1] (dataset to be described fully in Section 2.2.1), are used to illustrate the pruning steps and the voting mechanism in the proposed algorithm. Fifty-six training points in the figures are represented by filled symbols while test points are represented by unfilled symbols and labeled 1–6. The six held-out test samples are superimposed on the plots for illustrative purposes; they are not used to train the classifier. Fig. 1a is a plot of the convex hulls for cancer training samples and non-cancer training samples before any pruning begins. As such, Fig. 1a represents an individual two-dimensional member of the original convex-hull ensemble classifier [20], and the figure caption describes how this original ensemble member *votes* on each test sample. Fig. 1b shows the reduced convex hulls for positive (cancer) training samples and negative (non-cancer) training samples after the first pruning step of *disregarding training points in the intersection that are vertices of the convex hulls*. Three positive training points and one negative training point have been disregarded. As Fig. 1b shows, the intersection of the positive and negative pruned convex hulls still has two positive training points and three negative training points, including the vertices.

In the second step, one more positive training-point vertex and one more negative training-point vertex are disregarded. Fig. 1c shows the resulting reduced convex hulls. There is no overlap of the respective convex hulls, so the pruning process stops. Note that two positive test points fall within the reduced positive convex hull (points 1 and 2) and one negative test

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

point falls within the reduced negative convex hull (point 5). The new selective-voting-based classification approach is to *classify test points according to the final reduced convex hull in which they are located*; thus these three test points are classified correctly by this ensemble member (i.e., it votes correctly). Note also that three negative test points (3, 4 and 6) fall outside the reduced convex hulls. The reduced convex-hull classifier would not vote on test points 3 and 4, which were misclassified by the original algorithm (Fig. 1a). Thus, two incorrect votes have been eliminated. However, the classifier *would* vote (correctly) on test point 6, because of its position relative to the two convex hulls, as explained in the next section and illustrated in Fig. 1d.

Two pruning methods are considered here. The less-aggressive method, illustrated with reference to Fig. 1, reduces the positive and negative training-set-based convex hulls by disregarding vertices that lie in the intersection of the two convex hulls. Reduced (or pruned) convex hulls are newly constructed using all training points except those that are disregarded. This pruning process is done in a series of steps until there is no overlap. The more-aggressive strategy, in addition to disregarding the same points as the less-aggressive strategy, also disregards vertices outside the intersection that are connected to the disregarded interior points. When most ensemble members have small or modest overlap of positive and negative classes in the training set, less-aggressive pruning can produce reduced convex hulls that are highly homogeneous with respect to the class of training points they contain (e.g., colon cancer data illustrated in Fig. 1). When many or most ensemble members have considerable overlap of classes in the training set, more-aggressive pruning may be needed to produce reduced convex hulls that contain training points with high class homogeneity (e.g., breast cancer data to be described in Section 2.2.2; illustrated in Fig. 2).

2.1.2. Test points lying behind pruned convex hulls—An important feature of the proposed selective-voting method is that test points that fall “behind” one of the two reduced convex hulls for a given two-variable ensemble member are assigned to the same class as points that comprise the nearer convex hull behind which those points fall, provided there are no training points of the opposite class that lie behind the same convex hull. A test point is behind a convex hull if a straight line cannot be drawn from the test point to the farther convex hull without passing through the nearer convex hull (e.g., test point 6 in Fig. 1d). With the less-aggressive pruning method, there is no way for training points of one class to lie behind the reduced convex hull of the opposite class; thus, expanding the nearer reduced convex hull to include the test point(s), and therefore enable the classifier to vote, would not adversely affect the training thresholds used to select classifiers (to be discussed in Section 2.1.5). However, with the more-aggressive pruning method, training points from one class can lie behind the reduced convex hull of the opposite class. Because expanding the nearer reduced convex hull to include test points that lie behind it could affect a training threshold by also including training points from the opposite class, potential classifiers having training points so located are not considered further. In Fig. 1d, test point 6 is behind the pruned convex hull for negatives. Thus, the two-gene classifier assigns this point a (correct) negative vote.

2.1.3. Self-selection of voters by test points—As described, the new classifier selectively votes on test cases, depending on whether or not they fall inside or behind the reduced convex hulls. This is different from a selective voting strategy that selects the same subset of classifiers to vote on every test case [28,29]. In the proposed strategy, each test point *self-selects*, after the training phase, which classifiers will vote on it, in accordance with how representative the training points are of that test point. Thus, cancer predictions are individualized in that different subsets of classifiers with varying degrees of duplication will vote on different test cases. In the original convex-hull algorithm, the classifier depicted in Fig. 1a voted incorrectly on two test cases, 3 and 4. The reduced classifier (Fig. 1c) did not

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

vote incorrectly on any test points, although it did fail to vote on these two cases. Nevertheless, because the classification algorithm is an *ensemble* of reduced, two-dimensional convex-hull classifiers, not every member of the ensemble need necessarily vote on every test case. As long as a sufficient number of members of the ensemble do cast votes, every test point can be classified by a majority of those voting. However, if a test case cannot be classified, either due to a tied vote or to an insufficient number of votes being cast, this may indicate that such a case should not be classified by the algorithm. Such a case (patient) would need to be evaluated using different criteria. Hence, the selective-voting algorithm does not force arbitrary classification of all patients, in contrast to one-size-fits-all approaches.

2.1.4. Selection of informative subsets of predictor variables—One of the drawbacks of the ensemble classifier of Kodell et al. [20] is that it requires a large amount of CPU time to develop and train the ${}_pC_2$ potential two-dimensional convex-hull classifiers when p is large. With the present selective-voting algorithm, this problem is alleviated to a large degree by incorporating subset selection into the algorithm to identify only pairs of predictor variables with the best information for classification. The present approach is to use bivariate regression with best-subsets selection, which applies the branch-and-bound algorithm (leaps and bounds) [30] to select the q two-variable subsets having the highest regression R^2 values from p potential predictors, without having to evaluate all ${}_pC_2$ possible subsets. Selection bias [31] is not an issue in the development and cross-validation of the ensemble classifier because the selection of q two-variable subsets from p predictors is done inside the algorithm in each run for each training set. Because the objective of the convex-hull approach is to extract the best classification information contained in the two-dimensional space of a set of predictor variables, the regression R^2 is a natural criterion for ranking the information content in two-variable sets.

2.1.5. Options and thresholds for training the algorithm—Several options were considered in developing a prediction algorithm for a given dataset. These options included (1) the maximum number (q) of bivariate regression models from which to choose two-variable classifiers using p possible predictors (q = 50, 100, 200, 500, 1000 models in this paper), (2) the method of pruning of convex hulls (more-aggressive or less-aggressive pruning), and (3) the method of selection of a set of two-variable classifiers to retain from the set of q classifiers (either keep all qualified classifiers or keep only unique classifiers). To implement the unique-classifiers option, the two-variable classifiers were ranked according to their regression R^2 values, and a classifier was retained only if neither of its two variables was a member of a pair ranked higher. The third option depended on the following two thresholds to define qualified classifiers. First, each pruned convex hull of each potential classifier was required to contain at least a proportion X of the training samples of that convex hull's class to be considered for retention. Second, at least a proportion Y of the training samples contained in a pruned convex hull was required to be in the correct class for that convex hull. For the analyses reported here, these two thresholds were set at default values of X = 0.5 and Y = "proportion of training samples in the majority class of the training set" and were not varied. The default value of X was considered minimal for developing informed voters from the training set and for preventing extreme unbalance of positive and negative samples in the training set. The default value of Y was selected so that the PPV and NPV of each individual voter would be at least as high as that of a naïve voter who would classify every sample according to the majority class in the training set. The reason to consider retaining only a set of unique classifiers, as opposed to keeping all classifiers satisfying the two thresholds just described, is that unique members of a set of two-variable classifiers should tend to be less correlated with each other than

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

members that share a common predictor, and uncorrelated errors among members has been shown to be a desirable property of ensemble classifiers [20,32].

2.2. Application to cancer diagnosis and treatment with genomic predictors

To validate the selective voting strategy for increasing accuracy, two publicly available genomic datasets predictive of cancer outcomes [1,2] were used. The two datasets were previously analyzed by Kodell et al. [20], and thus provide a representative baseline from which to measure improvement with the new algorithm. It is acknowledged that continued re-use of the same data could lead to concerns about over-designing the classification scheme. To have as objective an assessment as possible for these datasets, multiple repetitions of ten-fold cross-validation (CV) were used for performance assessment. With 10-fold CV, a data set of n samples is randomly divided into 10 subsets, each having (approximately) $n/10$ samples. Each of these 10 subsets serves in turn as a test set. For each of these 10 test sets of approximate size $n/10$, a classifier is trained on the remaining $(10-1) \times (n/10)$ observations (the training set). The trained classifier is then used to classify the $n/10$ samples in the test set, and the ACC and other indices are calculated. The combined value of the ACC or other index over the 10 test sets, which is based on the prediction of all n samples one time each, is the cross-validated estimate of that index. For a given set of options, at least forty repetitions of 10-fold CV were performed based on different random permutations of the n samples in a dataset, and the average and standard deviation of these repetitions were used to assess the selective-voting algorithm's performance.

2.2.1. Colon cancer diagnosis – data—Alon et al. [1] presented gene expression data on 62 colon tissue samples, 40 samples being from cancerous colon tissue of patients with colon adenocarcinoma (positive cases) and 22 samples being from normal colon tissue of 22 of the 40 cancer patients (negative cases). The objective in this application was to develop a classification algorithm to screen for colon cancer based on patient-specific, high-dimensional genomic data to enable streamlined classification of new, undiagnosed tissue samples in a clinical setting. An initial set of 6500 genes whose expression levels were measured with an Affymetrix oligonucleotide array was reduced by Alon et al. to 2000 genes having the highest intensity levels across the 62 tissue samples. We started with a pre-processed version of Alon's data for which the expression level of each gene had been $\log_2$-transformed and normalized across the 62 samples by subtracting the mean and dividing by the standard deviation. We noted that the set of 2000 genes contains three control genes that have been replicated four times each; we did not remove these genes from the dataset. To put all genes on the same scale for forming two-dimensional convex hulls, the normalized expression levels were mapped to the unit interval by subtracting the minimum value among the 62 samples and dividing by the difference between the maximum and minimum values. The dataset of raw expression values is publicly available at http://microarray.princeton.edu/oncology/affydata/index.html (accessed: 10 June, 2002).

The results for several classifiers from twenty repetitions of 10-fold cross-validation (CV) for this subset of 2000 genes are shown in the upper portion of Table 1 (from [20]). Most of the classifiers had high accuracies (ACC), the highest being 87.4%. In light of such already-high ACC, the colon dataset represents a significant challenge for the proposed selective-voting algorithm to show improvement.

2.2.2. Breast cancer prognosis and treatment – data—van't Veer et al. [2] presented a gene-expression-based classification analysis of 78 primary breast cancer patients who had undergone surgery. The objective of their analysis was to increase the percentage of node-negative patients who are spared chemotherapy to a level above the current level of 15–20% based on conventional criteria. Thus, although it is customary to

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

subject a high percentage of post-surgery patients to adjuvant chemotherapy, it may be possible to identify patients with a good prognosis for whom the chemotherapy would not necessarily provide benefit, in which case these patients could be spared the toxic side effects of the adjuvant treatment. In the study of van't Veer et al. [2], there were 34 patients classified with a poor prognosis (developed distant metastases within 5 years: positive cases) and 44 patients classified with a good prognosis (did not develop distant metastases within 5 years: negative cases). For our classification analysis, we used fold changes and *p*-values provided by van't Veer et al. [2] on 24,481 genes to select 4741 genes that had no missing values and that had at least a two-fold difference and a p-value less than 0.01 in more than 3 tumor samples out of 78. To put all genes on the same scale for forming two-dimensional convex hulls, van't Veer's $\log_{10}$ (ratio) measures of gene expression were mapped to the unit interval by subtracting the minimum value among the 78 samples and dividing by the difference between the maximum and minimum values. The log10 expression values and associated p-values were publicly available until April 2011 at http://www.rii.com/publications/2002/vantveer.html (accessed: 17 Febraury, 2006).

The cross-validated results for several classifiers are shown in the upper portion of Table 2 (from [20]). The performance measures in Table 2 are quite low for all classification procedures, certainly too low for clinical use. Unreduced convex hulls for a representative pair of genes are displayed in Fig. 2a, showing that there is considerable overlap of positives and negatives, which is consistent with a high degree of heterogeneity. Classification accuracies based on a separate study by van de Vijver et al. [33] were considerably higher [34]; however, the dataset of van't Veer et al. [2] was analyzed here primarily because its low level of predictivity and its apparently high level of heterogeneity were thought to provide a good opportunity for selective voting to increase prediction accuracy. Fig. 2b–d shows how the unreduced convex hulls in Fig. 2a are aggressively pruned in the selective-voting algorithm to achieve separation of classes. As shown in Fig. 2d, this ensemble member casts correct votes on test points 2 and 8, but does not vote on the other six test points. In contrast, the original algorithm of Kodell et al. [20] casts votes on all eight test points, but votes incorrectly on points 5–7. Thus, the new algorithm eliminates three previously incorrect votes. This ensemble member is not forced to vote on test points 1, 3–7, for which other members may selectively cast votes.

3. Results

3.1. Colon cancer diagnosis – results

For the colon cancer data, although any combination of options with the selective-voting algorithm gave accuracies comparable to or better than the accuracies of the algorithms in the upper portion of Table 1, the best results were obtained with fewer rather than more models, with less-aggressive rather than more-aggressive pruning, and with unique sets of classifiers instead of full sets. The lower portion of Table 1 shows the results of five separate runs of 20 repetitions of 10-fold CV, along with the average of the five runs, for a total of 100 repetitions. These results are for $q = 100$ best models in terms of the bivariate regression R^2. The overall ACC of 89.8% was comparable to the ACC for 50 models (89.9%), but higher than that for 200 models (88.8%), 500 models (87.3%) and 1000 models (87.3%), each based on at least forty repetitions of 10-fold cross-validation. A two-sample *t* test comparing the accuracies of the original convex-hull algorithm [20] and the new selective-voting algorithm showed the new algorithm to have significantly higher accuracy ($p < 0.001$). The test compared the mean of the Freeman–Tukey-transformed accuracies of each of the 20 10-fold CVs that gave average accuracy of 86.0% for the old method (Table 1, row 1) to the mean of the 20 similarly transformed accuracies of the run that gave the lowest average accuracy of 89.6% for the new method (Table 1, row 16).

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

In return for the higher accuracies achieved by the selective-voting algorithm, a few samples were left unclassified in each of the five sets of twenty runs. Failure to classify occurred either because a test sample did not receive any votes at all or because of a tied vote. The average number of unclassified samples out of 62 samples per 10-fold CV was 0.76, which converts to approximately 1 out of 81 attempts. It is noted that even if all failures to classify were counted as misclassifications, the average ACC for 100 models would still be 88.7%. On the other hand, if all failures to classify were counted as half right and half wrong, average ACC would be 89.4%.

In addition to the accuracy, denoted by ACC, the other performance measures defined in the Introduction are displayed in the lower portion of Table 1. Like the results for all classifiers in the upper portion of Table 1, these results are biased in favor of the majority class (positives); however, the lack of balance between SEN and SPC and between PPV and NPV is less pronounced for the new method compared to the other methods. The actual numbers of unique voters derived from 100 models ranged between 0 and 18 from sample to sample. For 50 models the range was 0–13, for 200 models it was 0–22, for 500 it was 1–31, and for 1000 models it was 1–43.

3.2. Breast cancer prognosis and treatment – results

For the breast cancer data, in contrast to the colon cancer data, more-aggressive pruning tended to give higher accuracy than less-aggressive pruning and “keeping all qualified classifiers” tended to give higher accuracy than “keeping only unique classifiers.” The lower portion of Table 2 shows the results of five separate runs of 20 repetitions of 10-fold CV, along with the average of the five runs, for a total of 100 repetitions. All accuracies are well above the highest accuracy of 65.3% among the algorithms in the upper portion of Table 2. These results are for $q = 200$ best models in terms of the bivariate regression R^2. The overall ACC of 67.8% for 200 models was higher than the ACC for 50 models (67.1%), 100 models (67.3%), 500 models (67.5%) and 1000 models (67.4%), each based on at least forty repetitions of 10-fold cross-validation. A two-sample t test comparing the accuracies of the original convex-hull algorithm [20] and the new selective-voting algorithm showed the new algorithm to have significantly higher accuracy ($p < 0.003$). The test compared the mean of the Freeman–Tukey-transformed accuracies of each of the 20 10-fold CVs that gave average accuracy of 63.2% for the old method (Table 2, row 1) to the mean of the 20 similarly transformed accuracies of the run that gave the lowest average accuracy of 67.4% for the new method (Table 2, row 12).

As mentioned above for the colon data, in return for the higher accuracies achieved by the selective-voting algorithm, a few breast cancer samples also were left unclassified in each of the five sets of twenty runs. In all cases, all samples received votes, so failure to classify occurred only because of a tied vote. The average number of unclassified samples out of 78 samples per 10-fold CV was 0.35, which converts to approximately 1 out of 223 attempts. It is noted that even if all tied votes were counted as misclassifications, the overall accuracy for 200 models would still be 67.5%. On the other hand, if all failures to classify were counted as half right and half wrong, average ACC would be 67.7%.

In addition to ACC, the other performance measures in the lower portion of Table 2 compare very favorably to those of the algorithms in the upper portion of Table 2. Although biased toward the majority class (negatives), the lack of balance between SEN and SPC and between PPV and NPV is much less pronounced for the new method compared to the other methods. The actual numbers of voters derived from 200 models ranged between 12 and 150 from sample to sample. For 50 models the range was 1–46, for 100 models it was 7–79, for 500 models it was 28–285, and for 1000 it was 52–497.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

4. Discussion

The colon and breast cancer datasets offer two very different types of data for evaluating the potential of selective voting in convex-hull classification ensembles to increase classification accuracy. The gene-expression data are less variable from sample to sample among the colon tissue samples than among the breast tissue samples, and more predictive of whether a sample is positive or negative (Figs. 1 and 2). One possible reason for the difference in accuracy is that the colon-tissue classes already existed when gene expression was measured while the breast-tissue classes were not defined until 5 years after gene expression was measured. The contrast in accuracy and variability between colon cancer and breast cancer is further illustrated by the ROC curves in Fig. 3, which were derived using the fraction of positive votes for each test sample as a score. Values of SEN and 1-SPC from Table 1, run #3 (91.8%, 13.3%) and Table 2, run #4 (62.8%, 27.5%) give the respective coordinates of points on the upper (colon) and lower (breast) solid-line plots corresponding to a cut-off score of 0.5 (simple majority voting). The average areas under the ROC curves (with standard deviations), which represent probability measures of prediction accuracy, are 0.904 (0.016) and 0.692 (0.033) for colon and breast, respectively, indicating the substantially higher degree of predictivity for colon cancer compared to breast cancer. The scatter of individual curves around the average curves reflects the higher degree of variability for breast cancer than for colon cancer.

The combination of less-aggressive pruning along with retaining only unique classifiers, which was found to perform best for the colon samples, may perhaps be explained by the fact that a relatively small number of gene pairs gave good separation of classes without much pruning and these two-gene classifiers contain decidedly better information for classification than the majority of classifiers, so that keeping all potential classifiers would be counterproductive to getting the best majority vote. It is worth noting that in a set of twenty repetitions of 10-fold CV involving 200 training sets, 15 unique genes were repeated in at least 150 training sets. The relatively small number of genes that define the retained classifiers may show promise as biomarkers for colon cancer diagnosis. Although biomarker discovery is not the focus of this paper, using selective voting to develop gene sets as biomarkers will be a topic of future research.

In contrast to the colon cancer dataset, the gene expression data obtained from the breast tissue samples are much more variable from sample to sample and much less predictive of a patient's prognosis following surgery. The combination of more-aggressive pruning along with keeping all potential voters, which was found to perform best, can perhaps be explained by the tendency for high overlap of positive and negative samples and high variability from sample to sample, especially in the positive class (e.g., Fig. 2). That is, aggressive pruning was needed to achieve good separation of classes and the resulting classifiers were already sufficiently uncorrelated due the high variability, such that eliminating classifiers to obtain a unique set worked against the advantage of having the largest number of uncorrelated voters possible. Although the accuracies in the lower portion of Table 2 are likely still too low for meaningful clinical use, they are all well above the highest accuracy of 65.3%in the upper portion of the table, demonstrating the potential for selective voting to increase classification accuracy even for noisy, relatively uninformative data. However, the fact that many classifiers were kept and the accuracy was low seems to argue against finding a manageable set of genes as predictive biomarkers from such noisy data.

To apply the selective-voting algorithm in practice, several options must be considered: (1) the number of bivariate regression models, (2) the method of pruning, and (3) the method of retaining classifiers. One approach would be to create a "tuning" subset of the training data to find the best options experimentally. For option 3, the two default thresholds are

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

recommended (see Section 2.1.5). In the analysis of the colon- and breast-cancer datasets, all combinations of the three options gave higher accuracy than the original algorithm [20], except for the method of retention of classifiers for breast cancer, for which keeping all qualified classifiers was clearly better than keeping only unique ones. While the algorithm in [20] required 6 days of CPU time to run twenty 10-fold CVs for the colon data and 19 days to run twenty 10-fold CVs for the breast data, the new algorithm required only 1h and 13h, respectively.

A limited examination of the relationship between the number of voters for a test sample and the accuracy of the classification was inconclusive. In the colon dataset, for example, while the majority of samples with high vote totals were classified correctly, there were a few samples that tended to be misclassified from CV run to CV run even with high vote totals (non-cancer samples 8, 18 and 20; cancer samples 55 and 58). Conversely, a few samples tended to be classified correctly (or, else not classified), even with low vote totals (non-cancer samples 12 and 15; cancer sample 24). If the ensemble of potential voters is thought of as an expert committee, then low vote totals are not necessarily bad, as they could mean that even though only a few members are sufficiently knowledgeable to vote on difficult cases, the rest know they should abstain. This may be an important factor to consider in the development of biomarkers.

Selective voting in classification ensembles can potentially contribute to the treatment of patients for specific diseases by helping clinicians to assign therapies on an individualized basis, which is a goal of much current research in the development of drugs and other therapies. For personalized medicine to become a reality, methods must be developed that can distinguish patients according to relevant differences in disease types, risk factors, and responses to therapy. For example, many approved drugs are removed from the market after the post-marketing discovery of unexpected adverse events, such as liver toxicity [35]. It is conceivable that such toxicity is not predicted from pre-clinical and clinical studies because patients with high susceptibility to liver injury represent a small *hidden* subpopulation, and adverse liver events only stand out when a drug is administered to a large segment of the general population [36]. If sensitive subpopulations could be identified by genomic profiling, for example, then drugs could be approved for all patients except those in sensitive subpopulations. Similarly, if patients who are especially sensitive to adverse side effects of chemo or radiation therapy could be identified by genomic profiling, then the clinician could use this information to tailor these therapies on an individualized basis. Although classification algorithms are promising tools to support clinical diagnosis and treatment of disease, the usual implicit assumption underlying these algorithms, that all patients are homogeneous with respect to characteristics of interest, is unsatisfactory. By allowing for population heterogeneity reflected by characteristics that may not be readily apparent and thus not controlled, the proposed selective-voting approach may hold promise for improving the assignment of treatments and for identifying profiles of disease and risk sub-categories. Not only would this increase the likelihood of successful treatment, but also it could contribute to a better understanding of diseases themselves.

A positive feature of the ensemble of two-dimensional convex-hull classifiers is that it offers a convenient structure for addressing the challenge of integrating diverse types of data. Its simple structure easily facilitates the consideration of any pool of predictor variables to form two-dimensional classifier members of the ensemble, whether clinical, demographic, ‘omic, or any combination of such variables. Hence, it has the potential to enable moving beyond the use of only anatomic pathologic information to predict cancer prognosis and determine therapy, toward combining that information with clinical and demographic patient characteristics, as well as genomic and proteomic profiles, to significantly enhance predictions of cancer risk, prognosis and response to therapy.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was supported by National Cancer Institute Grant 1R01CA152667-01. The authors are grateful to Drs. Hojin Moon and Chen-an Tsai for their help with the data.

References

1. Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proceedings of the National Academy of Sciences. 1999; 96:6745–50.
2. van't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature. 2002; 415:530–6. [PubMed: 11823860]
3. Goldberg P. KRAS finding changes oncology practice but poses profound regulatory dilemma. The Cancer Letter. 2009; 35(4):1–8.
4. Moon H, Ahn H, Kodell RL, Baek S, Lin C-J, Chen JJ. Ensemble methods for classification of patients for personalized medicine with high-dimensional data. Artificial Intelligence in Medicine. 2007; 41:197–207. [PubMed: 17719213]
5. Hastie, T.; Tibshirani, R.; Friedman, J. The elements of statistical learning: data mining, inference, and prediction. New York: Springer; 2001.
6. Polikar R. Ensemble based systems in decision making. IEEE Circuits and Systems Magazine. 2006; 6:21–45.
7. Zhou Z-H, Wu J, Tang W. Ensembling neural networks: many could be better than all. Artificial Intelligence. 2002; 137:239–63.
8. Caruana, R.; Niculescu-Mizil, A.; Crew, G.; Ksikes, A. Ensemble selection from libraries of models. In: Brodley, CE., editor. Proceedings of the 21st international conference on machine learning. New York: Association for Computing Machinery; 2004. p. 8
9. Rokach L. Collective-agreement-based pruning of ensembles. Computational Statistics and Data Analysis. 2009; 53:1015–26.
10. Rokach L. Ensemble-based classifiers. Artificial Intelligence Review. 2010; 33:1–39.
11. Lam L, Suen CY. Application of majority voting to pattern recognition: an analysis of its behavior and performance. Systems, Man & Cybernetics. 1997; 27:553–68.
12. Breiman L. Bagging predictors. Machine Learning. 1996; 24:123–40.
13. Freund, Y.; Schapire, RE. Experiments with a new boosting algorithm. In: Saitta, L., editor. Proceedings of the 13th international conference on machine learning. San Francisco: Morgan Kaufmann; 1996. p. 148-56.
14. Breiman L. Random forest. Machine Learning. 2001; 45:5–32.
15. Friedman J, Hastie T, Tibshirani R. Adaptive logistic regression: a statistical view of boosting. Annals of Statistics. 2000; 28:337–74.
16. Ahn H, Moon H, Fazzari MJ, Lim N, Chen JJ, Kodell RL. Classification by ensembles from random partitions for high-dimensional data. Computational Statistics and Data Analysis. 2007; 51:6166–79.
17. Ho TK. The random subspace method for constructing decision forests. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1998; 20:832–44.
18. Bryll R, Gutierrez-Osuna R, Quek F. Attribute bagging: improving accuracy of classifier ensembles by using random feature subsets. Pattern Recognition. 2003; 36:1291–302.
19. Rokach L. Genetic algorithm-based feature set partitioning for classification problems. Pattern Recognition. 2008; 41:1676–700.
20. Kodell RL, Pearce BA, Baek S, Moon H, Ahn H, Young JF, et al. A model-free ensemble method for class prediction with application to biomedical decision making. Artificial Intelligence in Medicine. 2009; 46:267–76. [PubMed: 19081231]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

21. Cover TM, Hart PE. Nearest neighbor pattern classification. IEEE Transactions on Information Theory. 1967:21–7. IT-13.
22. Tibshirani R, Hastie T, Narasimhan B, Chu G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proceedings of the National Academy of Sciences. 2002; 99:6567–72.
23. Vapnik, V. The nature of statistical learning theory. New York: Springer; 1995.
24. Dudoit S, Fridlyand J, Speed TP. Comparison of discrimination methods for the classification of tumors using gene expression data. Journal of the American Statistical Association. 2002; 97:77–87.
25. Mardia, KV.; Kent, JT.; Bibby, JM. Multivariate analysis. San Francisco: Academic Press; 1979.
26. Foster DP, Stine RA. Variable selection in data mining: building a predictive model for bankruptcy. Journal of the American Statistical Association. 2004; 99:303–13.
27. Bellman, RE. Dynamic programming. Princeton, NJ: Princeton University Press; 1957. republished Dover:
28. Kotsiantis, SB.; Pintelas, PE. IEEE fourth international conference on intelligent systems design and applications. Budapest, Hungary: IEEE Press; 2004. Selective voting; p. 397-402.
29. Tsoumakas, G.; Partalas, L.; Vlahavas, I. A taxonomy and short review of ensemble selection. In: Okun, O.; Valentini, G., editors. Second workshop on supervised and unsupervised ensemble methods and their applications (SUEMA-2008). 2008. p. 41-6.
30. Furnival GM, Wilson RW. Regression by leaps and bounds. Technometrics. 1974; 16:499–511.
31. Ambroise C, McLachlan GJ. Selection bias in gene extraction on the basis of microarray gene-expression data. Proceedings of the National Academy of Sciences. 2002; 99:6562–6.
32. Dietterich TG. Machine-learning research: four current directions. AI Magazine. 1997 Winter;9:7–136.
33. vande Vijver MJ, He YD, van't Veer LJ, Dai H, Hart AA, Voskuil DW, et al. A gene-expression signature as a predictor of survival in breast cancer. New England Journal of Medicine. 2002; 347:1999–2009. [PubMed: 12490681]
34. Moon H, Ahn H, Kodell RL, Lin C-J, Baek S, Chen JJ. Classification methods for the development of genomic signatures from high-dimensional data. Genome Biology. 2006; 7:R121.1–7. [PubMed: 17181863]
35. Kaplowitz N. Idiosyncratic drug hepatotoxicity. Nature Reviews Drug Discovery. 2005; 4:489–99.
36. Boelsterli UA. Diclofenac-induced liver injury: a paradigm of idiosyncratic drug toxicity. Toxicology and Applied Pharmacology. 2003; 192:307–22. [PubMed: 14575648]
37. Ihaka R, Gentleman RR. A language for data analysis and graphics. Journal of Computational and Graphical Statistics. 1996; 5(3):299–314.
38. R Development Core Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2009. http://www.R-project.org

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

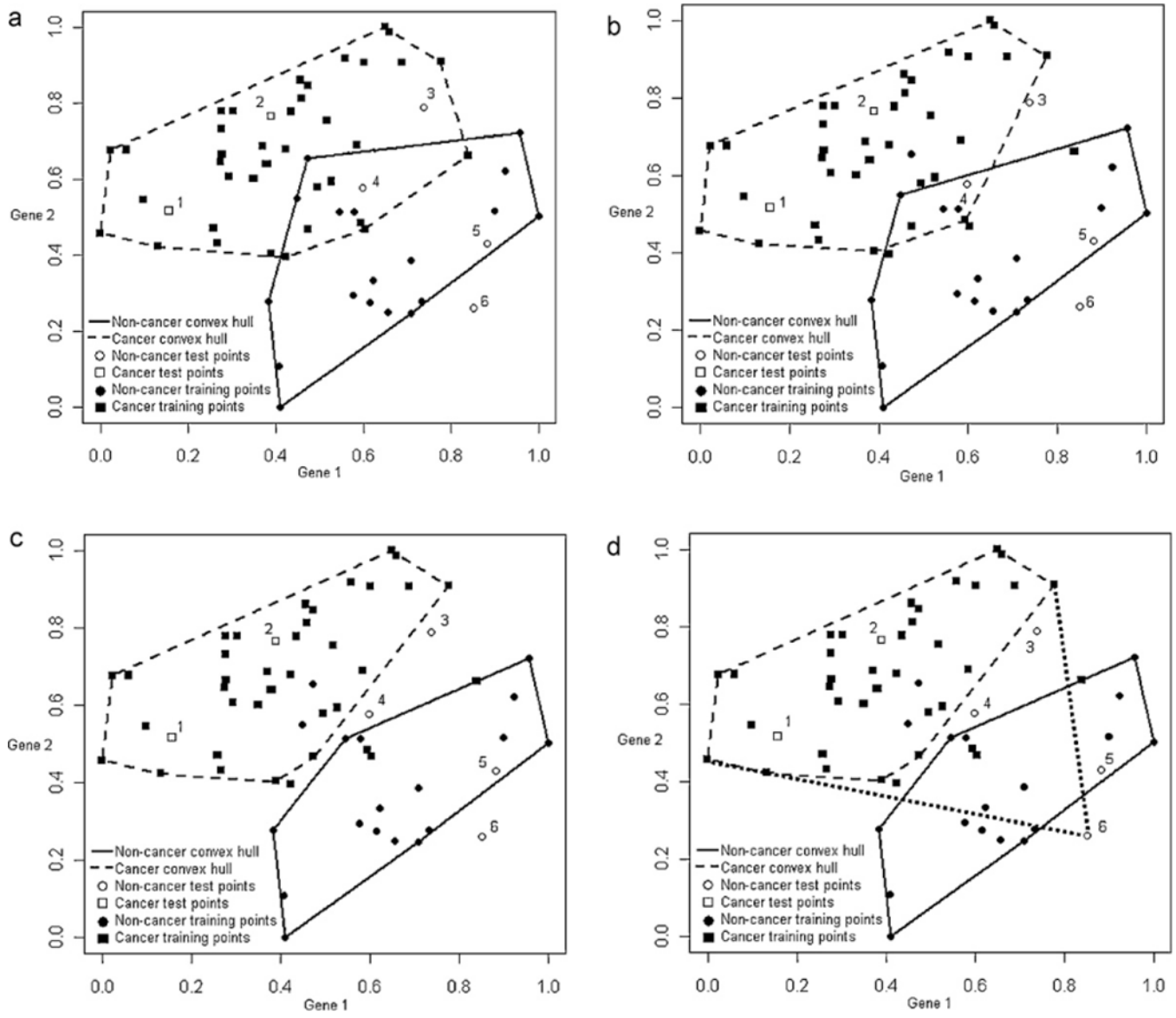

Fig. 1.

Two-dimensional convex hulls formed by normalized expression values of two representative genes for 56 training samples classified as cancerous or non-cancerous, with 6 numbered test samples superimposed. Panel (a) gives the unreduced convex hulls, by which this ensemble member votes correctly on test points 1, 2, 5, and 6 in the original algorithm of Kodell et al. [20], but votes incorrectly on test points 3 and 4. Panel (b) shows the first step of pruning with the new selective-voting algorithm and panel (c) shows the final step of pruning. With the new algorithm, panel (c) shows that this ensemble member votes correctly on test points 1, 2, and 5, but does not vote on test points 3 and 4. Panel (d) shows that test point 6, although outside both reduced convex hulls, also receives a correct vote because it lies "behind" a reduced convex hull.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

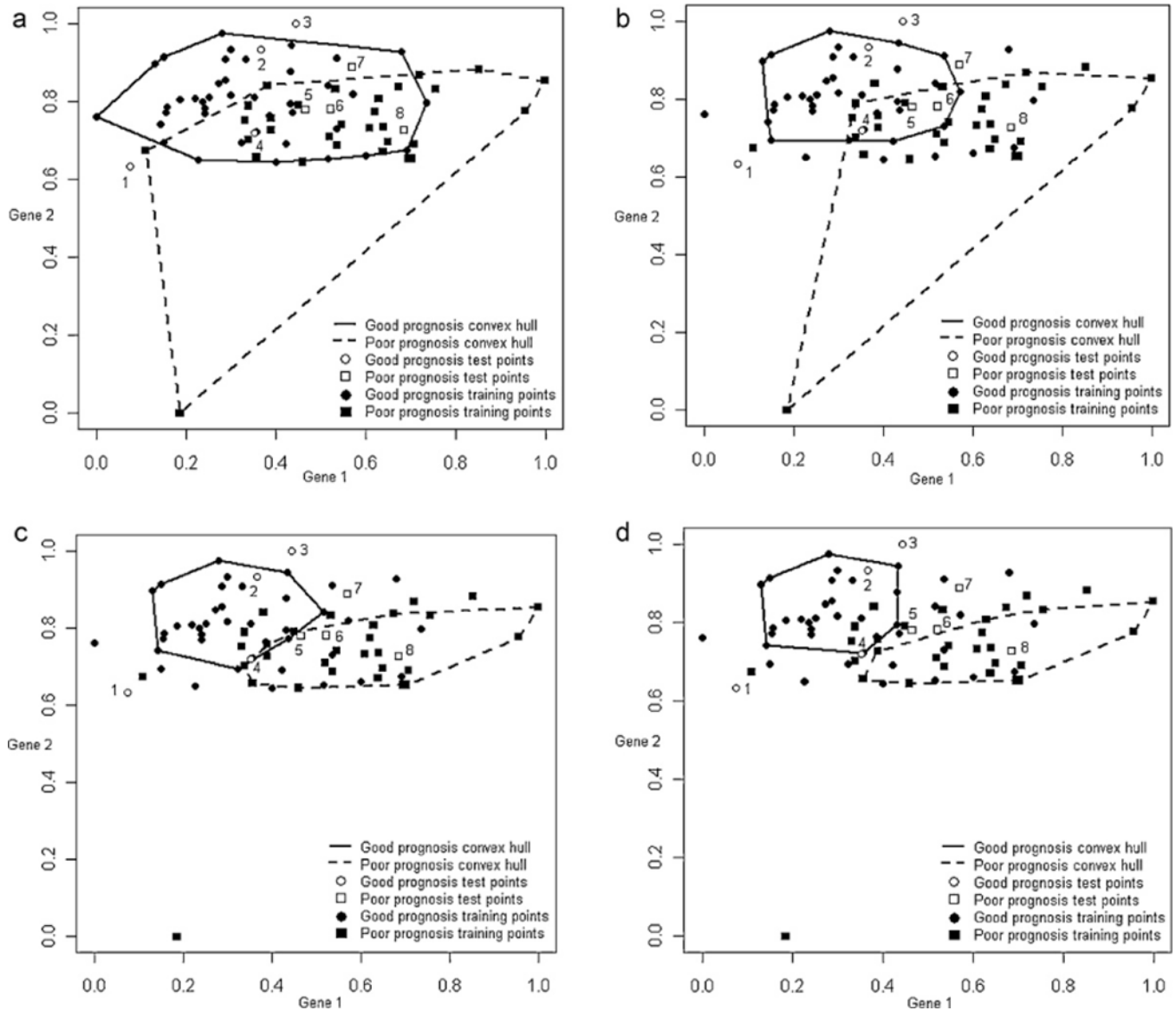

Fig. 2.

Two-dimensional convex hulls formed by normalized expression values of two representative genes for 70 training samples classified as having a good prognosis or a poor prognosis, with 8 numbered test samples superimposed. Panel (a) gives the unreduced convex hulls, by which this ensemble member votes correctly on test points 1–4 and 8 in the original algorithm of Kodell et al. [20], but votes incorrectly on test points 5–7. Panels (b)–(d) show the pruning steps of the new selective-voting algorithm. With the new algorithm, panel (d) shows that this ensemble member votes correctly on test points 2 and 8, but does not vote on the other six test points.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

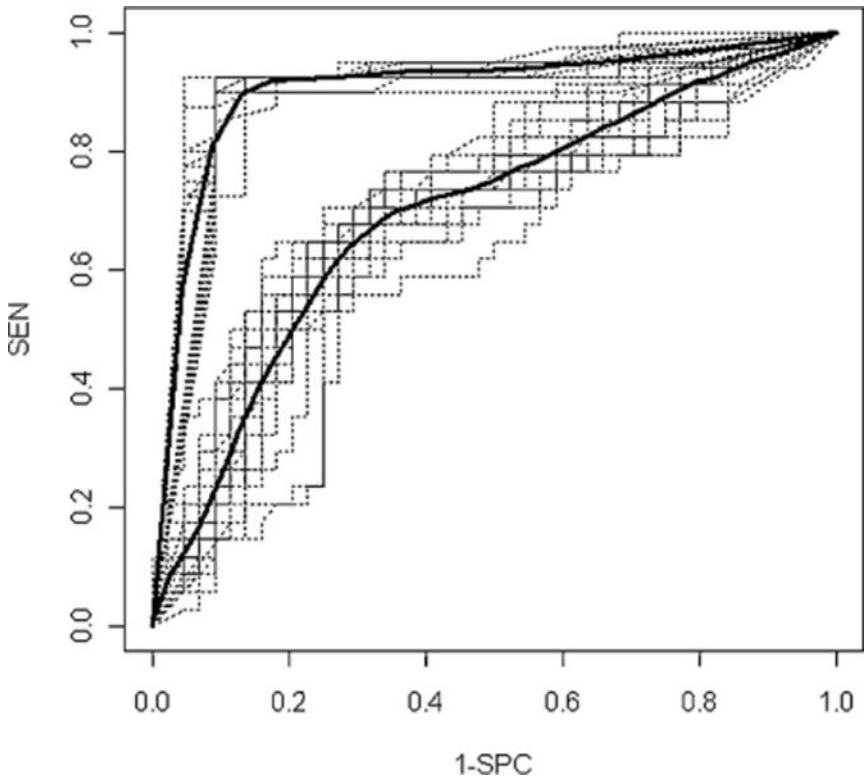

Fig. 3.

Receiver operating characteristic (ROC) curves for colon and breast cancer. Upper solid line is the average of twenty ROC curves derived from selective-voting run #3 for colon cancer in Table 1. Lower solid line is the average of twenty ROC curves derived from selective-voting run #4 for breast cancer in Table 2. Broken-line plots surrounding each solid-line average plot are ROC curves from individual runs.

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

Table 1

Performance (%) of classification algorithms for colon cancer data (62 samples: 40 positives, 22 negatives).[a]

Algorithm	ACC	SEN	SPC	PPV	NPV
Convex hull	86.0 (1.1)	88.0 (1.0)	82.3 (2.5)	90.0 (1.3)	79.0 (1.5)
C-T CERP	84.4 (1.4)	86.9 (2.0)	79.7 (3.3)	88.8 (1.3)	76.8 (2.9)
LR-T CERP	84.8 (2.3)	87.1 (1.5)	80.5 (4.9)	89.1 (2.5)	77.4 (2.5)
Random forest	81.2 (2.7)	87.8 (1.1)	69.3 (6.4)	83.9 (2.9)	75.6 (3.0)
AdaBoost	74.4 (2.9)	82.1 (3.0)	60.2 (4.9)	79.0 (2.3)	65.0 (4.5)
LogitBoost	74.0 (2.4)	82.3 (3.2)	59.1 (4.9)	78.6 (2.0)	64.9 (4.2)
k-NN	83.5 (3.5)	88.3 (2.3)	75.0 (8.0)	86.6 (3.7)	77.8 (4.4)
Shrunken centroid	84.7 (2.3)	86.9 (2.3)	80.7 (7.2)	89.3 (3.4)	77.2 (2.8)
SVM-linear	84.1 (2.1)	86.9 (1.6)	79.1 (4.8)	88.4 (2.4)	76.8 (2.5)
DLDA	85.1 (2.4)	86.0 (1.7)	83.4 (4.7)	90.4 (2.6)	76.6 (2.9)
FLDA	87.4 (2.1)	88.6 (1.7)	85.2 (4.6)	91.7 (2.4)	80.5 (2.6)
Selective voting					
1	89.8 (1.7)	91.5 (1.2)	86.7 (4.7)	92.7 (2.3)	84.8 (1.8)
2	90.0 (1.4)	92.0 (1.3)	86.3 (2.7)	92.6 (1.4)	85.3 (2.1)
3	90.0 (1.3)	91.8 (1.1)	86.7 (2.4)	92.6 (1.3)	85.3 (1.7)
4	89.8 (1.4)	91.3 (1.5)	87.1 (2.7)	92.9 (1.5)	84.5 (2.2)
5	89.6 (1.4)	91.1 (1.2)	86.9 (2.8)	92.8 (1.5)	83.9 (2.1)
1–5	89.8 (0.2)	91.5 (0.4)	86.7 (0.3)	92.7 (0.1)	84.8 (0.6)

[a]Upper portion is reproduced from [20], where each row gives the average (standard deviation) of twenty replications of 10-fold cross-validation for a given method. Parameter selection for each method is described in [16,20]. Lower portion gives results for five separate runs of twenty replications of 10-fold cross-validation for the selective-voting method, with the last row giving the average (standard deviation) of the five runs. ACC, SEN, SPC, PPV and NPV are defined in the Introduction.

Artif Intell Med. Author manuscript; available in PMC 2013 May 29.

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

Table 2

Performance (%) of classification algorithms for breast cancer data (78 samples: 34 positives, 44 negatives).[a]

Algorithm	ACC	SEN	SPC	PPV	NPV
Convex hull	63.2 (3.6)	52.1 (4.8)	70.2 (4.5)	58.4 (4.3)	66.4 (2.7)
C-T CERP	65.3 (2.1)	54.3 (3.9)	73.8 (3.6)	61.6 (3.1)	67.6 (1.8)
LR-T CERP	60.6 (3.0)	55.1 (5.6)	64.6 (3.7)	54.7 (3.6)	65.1 (3.1)
Random forest	62.5 (1.9)	46.8 (3.2)	74.7 (3.2)	58.9 (2.9)	64.5 (1.4)
AdaBoost	58.8 (4.1)	32.1 (8.9)	79.4 (6.9)	55.0 (9.4)	60.3 (2.8)
LogitBoost	65.2 (4.9)	55.6 (8.4)	72.6 (6.1)	61.6 (6.7)	68.0 (4.3)
k-NN	61.7 (3.6)	50.6 (8.4)	70.3 (4.7)	56.8 (4.5)	65.0 (3.7)
Shrunken centroid	60.9 (1.9)	50.6 (2.6)	68.9 (2.3)	55.7 (2.4)	64.3 (1.6)
SVM-linear	56.5 (2.9)	39.6 (5.3)	69.7 (2.7)	50.1 (4.2)	59.9 (2.5)
DLDA	62.5 (1.9)	52.4 (2.3)	70.3 (2.6)	57.8 (2.6)	65.6 (1.5)
FLDA	62.3 (2.6)	55.1 (4.9)	67.8 (2.0)	56.9 (2.9)	66.3 (2.6)
Selective voting					
1	67.4 (4.4)	59.3 (6.4)	73.6 (4.5)	63.4 (5.2)	70.2 (4.0)
2	67.5 (2.9)	61.4 (4.9)	72.1 (3.8)	63.1 (3.5)	70.7 (2.8)
3	67.7 (3.0)	61.7 (5.2)	72.4 (3.7)	63.3 (4.0)	71.1 (2.8)
4	68.3 (2.3)	62.8 (4.7)	72.5 (3.2)	63.8 (2.8)	71.8 (2.4)
5	68.1 (2.7)	60.8 (4.5)	73.8 (4.4)	64.3 (3.6)	71.0 (2.3)
1–5	67.8 (0.4)	61.2 (1.3)	72.9 (0.8)	63.6 (0.5)	71.0 (0.6)

[a]Upper portion is reproduced from [20], where each row gives the average (standard deviation) of twenty replications of 10-fold cross-validation for a given method. Parameter selection for each method is described in [4,20]. Lower portion gives results for five separate runs of twenty replications of 10-fold cross-validation for the selective-voting method, with the last row giving the average (standard deviation) of the five runs. ACC, SEN, SPC, PPV and NPV are defined in the Introduction.

Artif Intell Med. Author manuscript; available in PMC 2013 May 29.