Research Article

Statistics in Medicine

Received 26 February 2009, Accepted 7 November 2011 Published online 3 February 2012 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.4492

Identifying representative trees from ensembles

Mousumi Banerjee,[a*†] Ying Ding[a] and Anne-Michelle Noone[b]

Tree-based methods have become popular for analyzing complex data structures where the primary goal is risk stratification of patients. Ensemble techniques improve the accuracy in prediction and address the instability in a single tree by growing an ensemble of trees and aggregating. However, in the process, individual trees get lost. In this paper, we propose a methodology for identifying the most representative trees in an ensemble on the basis of several tree distance metrics. Although our focus is on binary outcomes, the methods are applicable to censored data as well. For any two trees, the distance metrics are chosen to (1) measure similarity of the covariates used to split the trees; (2) reflect similar clustering of patients in the terminal nodes of the trees; and (3) measure similarity in predictions from the two trees. Whereas the latter focuses on prediction, the first two metrics focus on the architectural similarity between two trees. The most representative trees in the ensemble are chosen on the basis of the average distance between a tree and all other trees in the ensemble. Out-of-bag estimate of error rate is obtained using neighborhoods of representative trees. Simulations and data examples show gains in predictive accuracy when averaging over such neighborhoods. We illustrate our methods using a dataset of kidney cancer treatment receipt (binary outcome) and a second dataset of breast cancer survival (censored outcome).

Keywords: bagging; random forest; tree similarity metric; representative trees; out-of-bag error

1. Introduction

Tree-based methods have become one of the most flexible, intuitive, and powerful data analytic tools for exploring complex data structures [1–4]. The applications of these methods are far reaching. The best documented and arguably most popular uses of tree-based methods are in biomedical research where classification is a central issue. Researchers [5–8] have also studied extensions to the censored data setting. Applications of tree-based analyses for classification and prognostication are abundant in the clinical literature [9–14].

One exciting development in recent years is the expansion of trees into ensembles of trees [15–19]. The mechanism of selecting a best split and the recursive partitioning of data leads to smaller and smaller datasets. This can lead to instability in the tree structure [15], whereby small changes in the data and/or algorithm inputs can have dramatic effects on the nature of the solution (variables and splits selected). Another major shortcoming of tree-based methods is their modest prediction performance, attributable to algorithm greediness and constraints that, while enhancing interpretability, reduce flexibility of the fitted functional forms. Growing an ensemble of trees and aggregating is a way to fix these problems. The advantage in growing many trees and using an aggregated estimate is that it is a way to reduce variance [16]. It also leads to classifiers and predictors that are drawn from a richer class of models [3]. Ensemble methods such as bagging [15, 19], boosting [19, 20], and random forest [16] yield substantial performance improvement over a single tree and address the inherent instability of a single tree.

Bagging [15, 19] involves bootstrapping the training data. A large number of pseudo datasets are generated by resampling the original observations with replacement and a tree grown on each pseudo dataset.

[a]*Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, U.S.A.*
[b]*Department of Biostatistics, Bioinformatics, and Biomathematics, Lombardi Comprehensive Cancer Center, Georgetown University, Washington, DC 20057, U.S.A.*
**Correspondence to: Mousumi Banerjee, Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, U.S.A.*
[†]*E-mail: mousumib@umich.edu*

This results in an ensemble of trees. The simple mechanism, whereby bagging reduces prediction error, is well understood in terms of variance reduction resulting from averaging [3].

Random forest [16] is an ensemble of unpruned classification or regression trees, induced from bootstrap samples of the training data, using random feature selection in the tree induction process. Correlation reduction is achieved by the random feature selection. Instead of determining the optimal split of a given node of a tree by evaluating all allowable splits on all covariates, as is performed with growing a single tree, a subset of the covariates drawn at random is employed. Once again, prediction is made by aggregating (majority vote for classification or averaging for regression) the predictions of the ensemble.

Compared with a single tree, ensemble methods typically offer substantial gains in predictive accuracy and are known to be stable. However, individual trees are lost in the ensemble. This is a significant loss, given that a tree is often used for decision making in the clinical setting. In this paper, we provide a methodology for identifying the most representative trees in an ensemble. Quite often, the problem may not be as bad as it seems; although hundreds of distinct trees are identified, many will differ only at a few nodes. Other trees may have different architectures but produce similar partitions of the covariate space. By defining several distance metrics on trees, we summarize an ensemble by several representative trees. Earlier work by other authors [21, 22] considered a similar approach in the context of exploring a set of plausible models, when either selection or averaging is the goal.

We organize this paper as follows. Section 2 describes our approach for growing an ensemble of trees. In Section 3, we introduce several distance metrics that capture similarity between two trees in an ensemble, focusing on architectural similarity as well as similarity in predictions. We choose the most representative trees on the basis of the average distance between a tree and all other trees in the ensemble. In Section 4, we present a method for estimating out-of-bag (OOB) error rate on the basis of a neighborhood of representative trees. To illustrate our methods, in Section 5, we present analyses of data on treatment patterns for kidney cancer where the outcome of interest is binary. As a second illustration, we present in Section 6 analyses of data from a breast cancer cohort study where the outcome of interest is recurrence-free survival (censored). Section 7 describes results from simulation studies. Finally, Section 8 contains concluding remarks.

2. Growing trees in the ensemble

First, we introduce some terminology. A tree T has a *root*, which is the top node, and observations are passed down the tree, with decisions being made at each *node* (also called daughters) until a *terminal node* or *leaf* is reached. Each nonterminal node (also called *internal node*) contains a question on which a split is based. The terminal nodes of a tree T are collectively denoted by $\tilde{T}$, and the number of terminal nodes is denoted by $|\tilde{T}|$. The branch T_h that stems from node h includes h itself and all its daughters.

In growing a tree, the natural question that arises is how and why a parent node is split into daughter nodes. Trees use binary splits, phrased in terms of the covariates, that partition the covariate space recursively. Each split depends upon the value of a single covariate. The partitioning is intended to increase within-node homogeneity. Goodness of a split must therefore weigh the homogeneities in the two daughter nodes. We measure quantitatively extent of node homogeneity using an 'impurity' function. We evaluate potential splits for each of the covariates and choose the covariate and split value resulting in the greatest reduction in impurity.

Corresponding to a split s at node h into left and right daughter nodes h_L and h_R, the reduction in impurity is given by

$$\Delta I(s,h) = i(h) - P(h_L)i(h_L) - P(h_R)i(h_R),$$

where $i(h)$ is the impurity in node h, and $P(h_L)$ and $P(h_R)$ are the probabilities that a subject falls in nodes h_L and h_R, respectively. For binary outcomes, $i(h)$ is measured in terms of entropy or Gini impurity [1, 2]. For continuous outcomes, $i(h)$ is typically the mean residual sum of squares [1, 2]. For censored outcomes, LeBlanc and Crowley [5] proposed using the deviance under the assumption that the hazard functions in two daughter nodes are proportional but unknown. The splitting rule that maximizes $\Delta I(s,h)$ over the set S of all possible splits is chosen as the best splitter for node h.

We grow an ensemble of trees using repeated bootstrap samples to inject randomness into the process. In general, each tree is grown to its full size without pruning: the only restriction being that a node contain no fewer than m observations; where m is typically some small number fixed a priori. This avoids

two important problems with trees, namely, when to stop growing a tree and how to optimally prune a tree. In contrast to using a random selection of covariates as is performed in the implementation of random forest, we use all covariates at each step of the splitting process. From that respect, our approach in growing the ensemble is akin to bagging.

3. Tree metrics

We think of each tree as a point in a high-dimensional space and cluster the trees according to some measure of proximity. Note that this space is much more complex than the Euclidean space, and distances between trees can be quantified in a number of ways. Any tree in the above space can be identified by a finite set of parameters, and these parameters can be broadly divided into two groups: the tree itself and the terminal node parameters. For example, the tree parameters could include the splitting rules and the architecture of the tree. The terminal node parameters could include the partition of the covariate space (defined by the terminal nodes) and the predictions from each terminal node. Metrics may be defined on either the tree or the terminal node parameters, or both. In the following discussion, we propose three different metrics, which capture different aspects of the tree.

Let T_1 and T_2 be two trees with b_1 and b_2 terminal nodes. Assume they have been trained using the same n observations $(y_i, \mathbf{x}_i), i = 1, \ldots, n$, where y_i denote the response for subject i and $\mathbf{x}_i = (x_{i1}, \ldots, x_{ik})$ is the vector of k covariates for the ith subject. We argue that two trees are similar if they use the same covariates for splitting. Towards that end, we define a metric that measures covariate mismatch between the two trees. In information theory, the Hamming distance between two strings of equal length is the number of positions for which the corresponding symbols are different. We define the Hamming distance between trees T_1 and T_2 as follows: suppose tree T_1 uses covariates x_1, x_5, x_6, and x_8 for splitting and tree T_2 uses covariates x_1, x_3, x_7, x_8, and x_9 for splitting. The k-dimensional binary strings corresponding to T_1 and T_2 are **1000110100 ... 0** and **1010001110 ... 0**, respectively, and the Hamming distance between T_1 and T_2 is 5. We define the standardized Hamming distance as

$$d_0(T_1, T_2) = \frac{\#\text{ of covariate mismatches between} T_1 \text{and} T_2}{\#\text{ of covariates in the study}} \tag{3.1}$$

In the aforementioned example, $d_0 = 5/k$.

The second metric emphasizes 'similarity' in the terminal nodes. Trees that are similar will place the same subjects together in a terminal node and separate the same subjects in different terminal nodes (i.e., if subjects i and j are placed in two different terminal nodes by T_1, then these two subjects should also be placed in two different terminal nodes by tree T_2 for it to be similar to T_1). Towards that end, we define a metric that captures how subjects are clustered in the terminal nodes. For all $\binom{n}{2}$ pairs of subjects, if subjects i and j are in the same terminal node by tree T_1, then $I_{T_1}(i, j) = 1$, otherwise $I_{T_1}(i, j) = 0$. We define the metric

$$d_1(T_1, T_2) = \frac{\sum_{i>j} \sum_j |I_{T_1}(i, j) - I_{T_2}(i, j)|}{\binom{n}{2}}. \tag{3.2}$$

The factor $\binom{n}{2}$ scales the metric to the range $(0, 1)$ such that 0 indicates perfect agreement. A pair of subjects contributes a positive amount to d_1 if and only if one tree places the subjects together and the other tree places them apart. Thus, d_1 is 0 if the two trees partition the covariate space in exactly the same way.

Instead of using the partition induced by the terminal nodes, one could define a metric that captures the similarity in predictions. Two trees are similar if the predictions from the two trees are the same for all subjects. For each subject i, we have a predicted value $\hat{y}_{ji}$ from tree T_j. For binary outcomes, $\hat{y}_{ji}$ represents the predicted probability of the event of interest in the terminal node that subject i is assigned to by tree T_j. For censored outcomes, $\hat{y}_{ji}$ could represent any summary statistic of the survival distribution (e.g., estimated 1-, 3-, and 5-year survival probabilities, median survival) in the terminal node that subject i is assigned to by tree T_j. We define the metric

$$d_2(T_1, T_2) = \frac{1}{n} \sum_{i=1}^{n} (\hat{y}_{1i} - \hat{y}_{2i})^2. \tag{3.3}$$

Thus, d_2 is 0 if the predictions from T_1 and T_2 are the same for all subjects.

Whereas d_2 captures prediction similarity, the metrics d_0 and d_1 capture similarity in tree architecture in terms of their focus on the splitting covariates and the partition induced by the terminal nodes. One could think of additionally rewarding two trees that not only cluster subjects similarly but use the same covariates in doing so. Towards that end, we define the metric

$$d_1^*(T_1, T_2) = \frac{d_0(T_1, T_2) \sum_{i>j} \sum_j |I_{T_1}(i, j) - I_{T_2}(i, j)|}{\binom{n}{2}}, \tag{3.4}$$

which is a weighted version of d_1 using Hamming distance between the two trees as the weight. Note that if two trees cluster the subjects exactly the same (can happen only with a small chance in practice), then the numerator of d_1^* will be 0 no matter what $d_0(T_1, T_2)$ is. Thus, d_1^* really focuses on the pairs of trees that do not match exactly in terms of patient clustering.

3.1. Choosing the most representative trees

We compute the similarity score $D(T)$ for a tree T by averaging the individual distance metrics between tree T and all other trees in the ensemble. This is the average distance between tree T and all other trees in the ensemble. So, a low score for a tree indicates its similarity to all other trees in the ensemble. We compute the score $D(T)$ for each of the distance metrics (i.e., d_0, d_1, d_2, and d_1^*) and choose the representative trees in the ensemble on the basis of the smallest $D(T)$ values.

4. Out-of-bag error

Error rate performance is calculated on the basis of the ensemble estimate. For binary outcomes, our key deliverable is the predicted probability of the event of interest. We derive the ensemble estimate as follows. First, from each tree grown in the ensemble, we estimate the probability of event for each subject i by grouping subjects by terminal nodes. To estimate the probability of event for patient i with predictor $\mathbf{x}_i$, simply drop their covariate $\mathbf{x}_i$ down the tree. If $\mathbf{x}_i$ lands in terminal node h, then define

$$\hat{p}(\mathbf{x}_i) = \hat{p}_h, \text{ if } \mathbf{x}_i \in h.$$

Note that we compute this value for all individuals i in the data.

We base the aforementioned estimate on one tree. To obtain the ensemble estimate, we average over all trees in the ensemble. Let $\hat{p}_b(\mathbf{x}_i)$ be the predicted probability of event for subject i from tree b. Then, the ensemble estimate is

$$\hat{p}_{\text{ens}}(\mathbf{x}_i) = \frac{1}{B} \sum_{b=1}^{B} \hat{p}_b(\mathbf{x}_i),$$

where B is the total number of trees in the ensemble.

Now, define $I_{i,b} = 1$ if i is OOB for b (i.e., i is not included in the bootstrap sample used to grow tree b), otherwise $I_{i,b} = 0$. The OOB ensemble estimator for patient i is

$$\hat{p}_{\text{oob}}(\mathbf{x}_i) = \frac{\sum_{b=1}^{B} I_{i,b} \hat{p}_b(\mathbf{x}_i)}{\sum_{b=1}^{B} I_{i,b}}. \tag{4.1}$$

Note that we obtain $\hat{p}_{\text{oob}}(\mathbf{x}_i)$ by averaging over only those trees for which the bootstrap samples did not include i, that is, those bootstrap samples for which subject i was OOB. This is in contrast to the ensemble estimator $\hat{p}_{\text{ens}}(\mathbf{x}_i)$ that uses all the trees.

Because we ultimately want to assign OOB error rates to the representative trees, we compute $\hat{p}_{\text{oob}}(\mathbf{x}_i)$ using neighborhoods of trees that are deemed similar based on $D(T)$. In essence, the averaging in Equation (4.1) is carried out over neighborhoods of representative trees. For example, using the top $l\%$ of the most representative trees to define a neighborhood R_l, we compute

$$\hat{p}_{\text{oob}}^{(l)}(\mathbf{x}_i) = \frac{\sum_{b \in R_l} I_{i,b} \hat{p}_b(\mathbf{x}_i)}{\sum_{b \in R_l} I_{i,b}}, \tag{4.2}$$

where $\hat{p}_{\text{oob}}^{(l)}(\mathbf{x}_i)$ is the OOB ensemble estimator of the probability of event for patient i based on a neighborhood R_l of similar trees.

Given the OOB estimator $\hat{p}^{(l)}_{\text{oob}}(\mathbf{x}_i)$, we finally obtain the error rate using Harrell's concordance index [23]. The latter is calculated using all possible pairs of subjects, for whom one experienced the event ($y = 1$) and the other did not ($y = 0$). Denote the total number of such subject pairs by a. We judge prediction performance by the ability of the model to allocate to patients who experience the event a higher predicted probability of experiencing the event (i.e., worse predicted outcome) than that allocated to those who did not experience the event. Denote by e the number of subject pairs for whom the subject who experienced the event had a higher predicted probability of event. Finally, denote by t the number of subject pairs for whom the predicted probabilities for the subjects are tied. We then derive Harrell's concordance (c) index as $c = (e + 0.5t)/a$ and calculate the OOB estimate of error rate as $1 - c$. A value of 0.5 for the error rate corresponds to a procedure doing no better than random guessing, whereas 0 indicates perfect accuracy. Given that enough trees have been grown, the OOB estimate of error rate is an accurate estimate of test set prediction error rate [16].

In the censored data setting, we use the OOB ensemble estimate of the cumulative hazard function to compute error rates. Consider a specific node h. Let $\{y_{l,h}\}$ be the distinct death times for patients in node h, and let $d_{l,h}$ and $r_{l,h}$ be the number of deaths and individuals at risk at time $y_{l,h}$. The cumulative hazard function estimate for node h is defined as

$$\hat{H}_h(y) = \sum_{y_{l,h} \leq y} \frac{d_{l,h}}{r_{l,h}}.$$

Each tree provides a sequence of such estimates $\hat{H}_h(y)$. If there are M terminal nodes in the tree, then there are M such estimates. Once again, to estimate the cumulative hazard function for patient i with predictor $\mathbf{x}_i$, we simply drop their covariate $\mathbf{x}_i$ down the tree. As before, let $\hat{H}_b(y \mid \mathbf{x}_i)$ be the cumulative hazard estimate for subject i from tree b. Then, the OOB ensemble cumulative hazard estimator for patient i based on the neighborhood R_l of similar trees is given by

$$\hat{H}^{(l)}_{\text{oob}}(y \mid \mathbf{x}_i) \quad = \quad \frac{\sum_{b \in R_l} I_{i,b} \hat{H}_b(y \mid \mathbf{x}_i)}{\sum_{b \in R_l} I_{i,b}}. \tag{4.3}$$

To compute Harrell's concordance index in the censored data setting, one must define what constitutes a worse predicted outcome. We take the following approach. Let $y^*_1, y^*_2, \ldots, y^*_N$ denote all unique event times in the data. Subject i is said to have worse predicted outcome than subject j if

$$\sum_{k=1}^{N} \hat{H}^{(l)}_{\text{oob}}(y^*_k \mid \mathbf{x}_i) > \sum_{k=1}^{N} \hat{H}^{(l)}_{\text{oob}}(y^*_k \mid \mathbf{x}_j).$$

Harrell's c index is then defined as the proportion of all allowable subject pairs in which the predicted and observed outcomes are concordant [24]. Unlike other measures of survival performance, Harrell's c index does not depend on choosing a fixed time for evaluation of the model and specifically takes into account censoring of individuals.

5. Treatment patterns for kidney cancer

As an illustrative example, we present analyses of data from a population-based study of kidney cancer where the outcome of interest is (binary) receipt of treatment. Radical nephrectomy is the traditional gold standard for treating patients with organ-confined kidney cancer. During the last two decades, however, the introduction of a nephron-sparing alternative (i.e., partial nephrectomy) to radical excision has appreciably modified the therapeutic options for patients with kidney cancer. Partial nephrectomy yields oncologic outcomes that are indistinguishable from those achieved by radical excision and also preserves long-term renal function while reducing overtreatment of patients with benign tumors. Despite these potential benefits to patients, population-based data suggest that the adoption of partial nephrectomy has been slow. Consequently, radical nephrectomy remains the predominant surgical therapy for patients with kidney cancer. The goal of the current study was to identify and characterize patient subgroups who are likely to receive partial nephrectomy based on clinical and sociodemographic factors.

The analysis cohort comprised 1507 Medicare beneficiaries treated surgically for kidney cancer diagnosed between 1997 and 2002. The outcome of interest is receipt of partial versus radical nephrectomy (i.e., binary outcome). A total of 10 covariates were considered for the analysis. These included

sociodemographic variables (age, race, gender, marital status, and socioeconomic status), tumor size, and preexisting comorbidity (using a modification of the Charlson index). In addition, we included prior history of hypertension, urolithiasis, and/or renovascular disease, given their relevance to surgical decision making among patients with kidney cancer. On the basis of standard clinical guidelines, we categorized tumor size as $\leqslant$4 cm and $>$4 cm. We used median census-tract income and census-tract percentage of non-high school graduates as patient-level measures of income and education, respectively, to obtain socioeconomic status.

We implemented the analysis using the RPART algorithm in R [25, 26]. We grew an ensemble of 500 classification trees using the Gini index as the measure of within-node impurity for the goodness-of-split criterion. The size of the individual trees constituting the ensemble is controlled by a tuning parameter, which specifies the minimum number of observations in a node for which the tree will try to split. This was set to the value of 10, which generally gives good results (as corroborated by our simulations that follow). For computing the metric d_2, we used the predicted probability of partial nephrectomy. We based prediction performance on the ensemble estimate of the probability of receiving partial nephrectomy.

Figures 1–4 present a subset of the most representative trees chosen on the basis of D_{d_0}, D_{d_1}, $D_{d_1^*}$, and D_{d_2}, respectively. At each level of the tree, we show the best splitter (covariate with cutpoint). The letters in each terminal node denote the predicted class for patients in that node (RAD for radical and PAR for partial nephrectomy). Note that both trees in Figure 1 (representative trees based on D_{d_0}) use the same covariates to split (namely, tumor size, age, prior history of hypertension, and comorbidity index), although their complexities are different (e.g., tree 28 has eight terminal nodes, whereas tree 67 has five terminal nodes). Furthermore, the aforementioned four covariates turned out to be the most prominent (ranked topmost) in terms of variable importance for the forest. In general, the metric D_{d_0} was fairly discrete and may therefore be less adept at picking up architectural differences between trees than D_{d_1}. Figure 2 presents two most representative trees chosen on the basis of D_{d_1} (i.e., in terms of how they cluster patients in the terminal nodes). Some of the terminal node characterizations in these two trees are exactly the same. Yet, tree 34 uses two additional covariates (marital status and gender) for splitting. The D_{d_0} value for tree 92 is 0.253 (of the 500 trees in the ensemble, 12% had a value for D_{d_0} that is smaller than 0.253). Thus, tree 92 is also representative of other trees in the ensemble in terms of the covariates used for splitting. Figure 3 presents trees that are most similar to other trees in the ensemble from both aspects ($D_{d_1^*}$ rewards architectural similarity, i.e., trees that not only cluster subjects similarly but also use the same covariates in doing so). It is therefore not surprising that tree 92 emerges as the most representative tree in terms of architectural similarity with other trees in the ensemble. Figure 4 presents trees that are most similar to other trees in the ensemble in terms of predictions. Note that the complexities of these trees are different (tree 158 has six terminal nodes, and tree 451 has four terminal nodes). Both trees use tumor size and age; however, tree 158 splits on hypertension as well, whereas tree 451 splits on comorbidity index. Although similar clustering of patients will yield similar predictions, and there is some overlap in how patients are being clustered by these trees, it is not necessary for trees to cluster patients the same way to achieve small values of D_{d_2}.

Table I presents OOB error rates using representative tree neighborhoods. We use the top 5%, 10%, and 25% of the most representative trees (on the basis of the smallest value of each metric d_0, d_1, d_2, and d_1^*), as well as all 500 trees in the forest to define various neighborhoods. For a fixed neighborhood

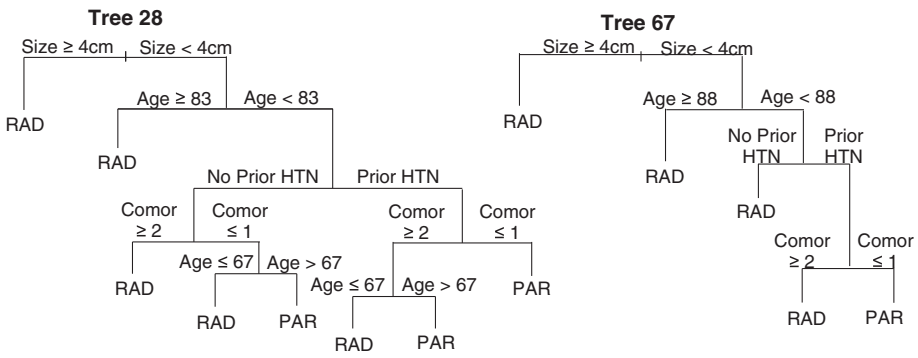

Figure 1. Kidney cancer data: representative trees chosen by d_0.

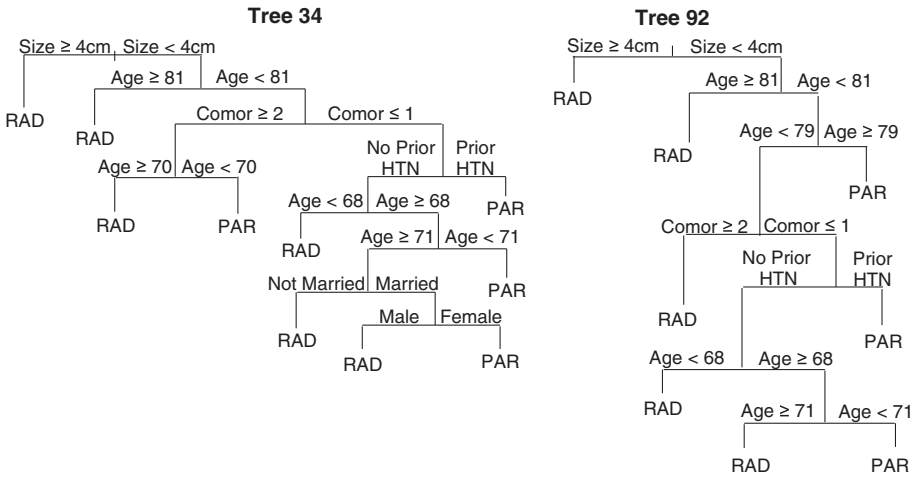

Figure 2. Kidney cancer data: representative trees chosen by d_1.

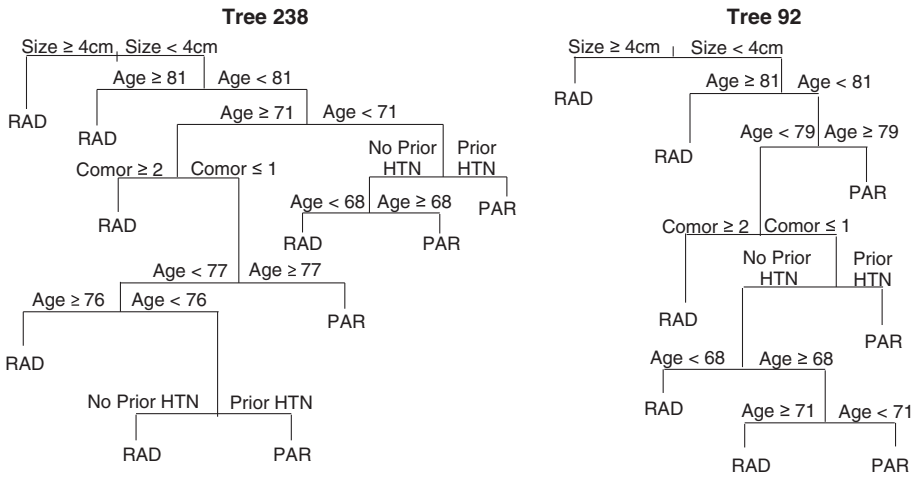

Figure 3. Kidney cancer data: representative trees chosen by d_1^*.

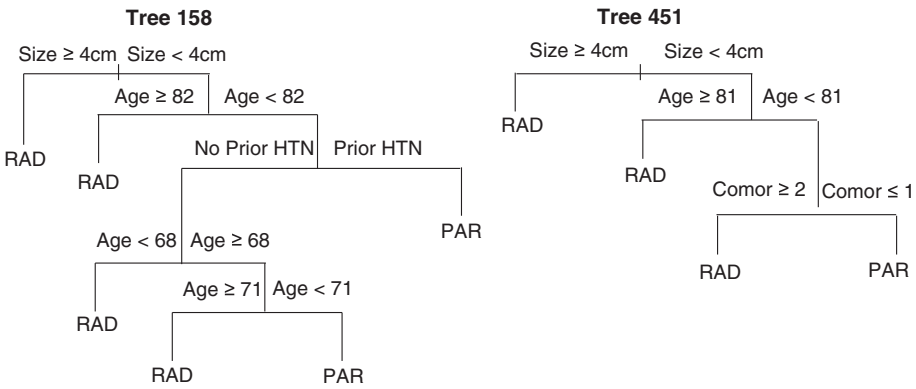

Figure 4. Kidney cancer data: representative trees chosen by d_2.

size, the error rates obtained using the different metrics are very similar. There is modest gain in predictive accuracy when averaging over neighborhoods of representative trees compared with averaging over the entire ensemble. For example, the OOB error rate based on averaging over the top 10% most representative trees chosen by D_{d_2} was 30.4%, which is roughly a 12% relative reduction from the error rate

Table I. Kidney cancer data: out-of-bag error rates based on Harrell's concordance index and with the use of neighborhoods of representative trees.

Neighborhood (%)	D_{d_0} (%)	D_{d_1} (%)	D_{d_2} (%)	$D_{d_1^*}$ (%)
5	*	30.0	29.2	30.2
10	30.7	30.4	30.4	30.5
25	31.7	31.7	31.6	31.8
All trees in the ensemble			34.5	

*There are 59 trees in the ensemble with the smallest value of $D_{d_0}(0.248)$. Thus, the smallest unique neighborhood is the top 11% neighborhood of representative trees.

obtained on the basis of averaging over all 500 trees in the ensemble. Overall, this pattern is consistent across all the metrics, with relative reductions in OOB error rates ranging from 8% to 15% roughly.

6. Breast cancer prognostic study

To illustrate the application of our methodology in the censored data setting, we present a second data analyses of a cohort study of breast cancer patients. Women eligible for this study were newly diagnosed patients with stage I, II, or III breast cancer, diagnosed between January 1990 and December 1996 at Harper Hospital in Detroit, Michigan. Detailed demographic, clinical, pathological, treatment, and follow-up information were obtained from the Surveillance, Epidemiology, and End Results database, hospital and clinic records. Recurrence-free survival (RFS) was the primary endpoint of the study, defined as the interval between diagnosis and documented regional/local or distant recurrence. The goals of the study were to analyze the relative contributions of patient-related and tumor-related prognostic factors on RFS and to identify patient subgroups with homogeneous RFS within a group but different RFS between groups.

The analysis cohort comprised 764 patients. A total of 10 covariates were considered for the analysis. These included sociodemographic variables (age, race, marital status, and socioeconomic status), factors characterizing tumor (tumor size, number of positive lymph nodes, tumor differentiation, estrogen receptor status, and progesterone receptor status), and body mass index as a comorbid factor. Patients were classified as obese if their body mass index was >30, per the standard guideline recommended by the World Health Organization. On the basis of standard clinical guidelines, tumor size was categorized as $\leqslant 2$ cm and >2 cm. Number of positive lymph nodes was categorized as 0, 1–3, 4–9, and >10 positive nodes. Tumor differentiation was also categorized as moderate and poor. Estrogen and progesterone receptors were binary categorical variables (positive/negative).

We implemented our analysis using the RPART algorithm in R devised by Therneau and Atkinson [25, 26]; but this can be implemented in any recursive partitioning software that fits Poisson trees, by capitalizing on the equivalence noted by LeBlanc and Crowley between their relative risk tree and Poisson tree likelihoods [5]. We grew 500 trees in the ensemble. We set the minimum number of observations in a node for which the tree will try to split to the value of 10, as in the kidney cancer study.

Figure 5 presents a subset of the most representative trees chosen on the basis of D_{d_2}. At each level of a tree, we show the best splitter (covariate with cutpoint). Numbers in the terminal nodes of each tree denote the Kaplan–Meier estimates of the 5-year survival probabilities. Note that the complexities of these trees are different (tree 83 has seven terminal nodes, and tree 375 has five terminal nodes). Table II presents five randomly selected patients and their corresponding 5-year survival predictions based on the trees in Figure 5. Although D_{d_2} is not based on comparing these four trees to each other (but rather each tree to all other trees in the ensemble), note the similarity in predictions, most evident for trees 83 and 375 that had the smallest values of D_{d_2}. The sets of representative trees picked by D_{d_0} and D_{d_1} are largely different from the ones picked up by D_{d_2}. This is not surprising, given that d_0, d_1, and d_2 focus on different aspects of tree similarity. The metric d_0 once again turned out to be fairly coarse.

Table III presents OOB error rates using neighborhoods of similar trees. Once again, we use the top 5%, 10%, and 25% of the most representative trees (on the basis of the smallest value of each metric d_0, d_1, d_2, and d_1^*), as well as all 500 trees in the ensemble to define various neighborhoods. In addition to the OOB error rates using Harrell's concordance index, we also computed OOB error rates using

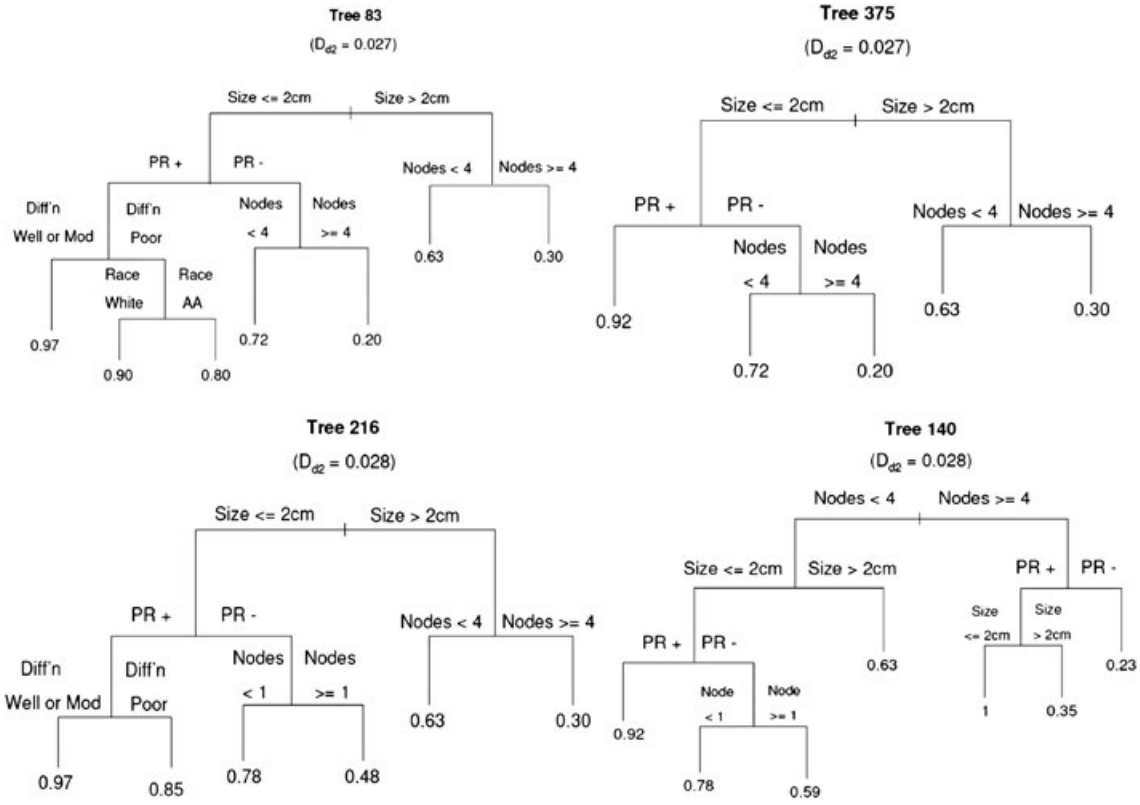

Figure 5. Breast cancer data: representative trees chosen by d_2.

Table II. Breast cancer data: 5-year survival predictions from representative trees selected by d_2.

Subject	Tree 83	Tree 375	Tree 216	Tree 140
18	0.30	0.30	0.30	0.23
123	0.90	0.92	0.85	0.92
368	0.63	0.63	0.63	0.63
506	0.72	0.72	0.78	0.78
654	0.97	0.92	0.97	0.92

the Brier score [24, 27], another measure commonly used in the survival data setting. In an ordinary least squares regression model, the standard approach to evaluating accuracy of the model is to look at the residual mean square error, which is a measure of discrepancy between the observed and predicted values. The Brier score is based on the aforementioned idea. For a fixed time point y^*, the Brier score is interpreted as a mean square error of prediction when the estimated event free probabilities $\hat{\pi}(y^* \mid \boldsymbol{x}_i)$, which takes values in the interval $[0, 1]$, are viewed formally as predictions of the event status at y^*, $I(y_i > y^*) \in \{0, 1\}$. Graf *et al.* recommended using an integrated version of the Brier score that incorporates quadratic loss averaged over time [27]. Table III presents results on the basis of both approaches. Note that with the use of either approach, the error rates obtained by averaging over neighborhoods of representative trees are only marginally smaller compared with the error rate based on averaging over all trees in the ensemble. Within each neighborhood, the error rates corresponding to using the different metrics are very similar.

We also computed the OOB error rate from Ishwaran and Kogalur's random survival forest [18] using log-rank splitting rule. In an earlier paper [8], we reported strong concordance between the log-rank splitting rule and the splitting rule based on proportional hazards model that we used to grow our ensemble. In contrast to our approach, Ishwaran and Kogalur [18] used a random selection of *ptry* covariates at each step of the splitting process. The OOB error rate (based on Harrell's concordance index) obtained

Table III. Breast cancer data: out-of-bag error rates based on Harrell's concordance index and integrated Brier score using neighborhoods of representative trees.

Neighborhood (%)	Error rate	D_{d_0}	D_{d_1}	D_{d_2}	$D_{d_1^*}$
5	HC	*	30.0%	28.9%	29.4%
	IBS	*	0.176	0.170	0.174
10	HC	29.5%	29.2%	28.8%	28.9%
	IBS	0.175	0.176	0.170	0.172
25	HC	29.6%	29.3%	29.6%	29.5%
	IBS	0.174	0.175	0.172	0.174
All trees in the ensemble	HC		30.0%		
	IBS		0.177		

*There are 50 trees in the ensemble with the smallest value of D_{d_0}(0.239). Thus, the smallest unique neighborhood is the top 10% neighborhood of representative trees.
HC, Harrell's concordance index; IBS, integrated Brier score.

using their method was 29.5% (achieved by the optimal value of $ptry = 2$), compared with 30.0% (based on all 500 trees) that we obtained using our method.

7. Simulation study

Because in a real dataset we can only illustrate the methods and the truth about the parameters is unknown, we conducted a simulation study to evaluate our proposed methods, mimicking the real data analysis in certain aspects. Firstly, we undertook simulations to study the effect of the different tuning parameters (i.e., number of trees grown in the ensemble, minimum number of observations in a node for which the tree will try to split, etc.) on the set of representative trees picked by the various distance measures. We considered three different covariates: X_1 generated from a Bernoulli(0.3) distribution, X_2 generated from a discrete distribution taking three values 1, 2, and 3 with probabilities 0.45, 0.28, and 0.27, respectively, and X_3 generated from a Normal(50,16) distribution. We generated data from Weibull distribution, with the following choices of the shape parameter: $\alpha = 0.5,\ 1$, and 2. We considered three different models for the scale parameter λ: (i) $\lambda = 1$; (ii) $\lambda = 1$ if $X_1 = 1$, and $=$ exp(1) (=2.72) if $X_1 = 0$; (iii) $\lambda = 1$ if $X_2 = 1$, and $=$ exp(1) if $X_2 > 1$; (iv) $\lambda = 1$ if $X_2 = 1$; or 3, and $=$ exp(1) if $X_2 = 2$; and (v) $\lambda =$ exp(1) if $46 \leqslant X_3 \leqslant 54$ (within mean $\pm$ SD), and $= 1$ otherwise. We assumed independent censoring with censoring proportions of 0%, 25%, and 50%. For each simulation design scenario, we generated samples of size n (number of observations) $= 500$. We varied the tuning parameters as follows: number of trees grown in the ensemble ($ntree$) was set at 250, 500, and 1000; and minimum number of observations in a node for which the tree will try to split ($minsplit$) was set at 10, 15, and 20. We also varied the selected neighborhood of similar trees using top 5%, 10%, and 25% of the most representative trees chosen on the basis of each distance measure. Note that we are not interested in how the tuning parameters alter the absolute values of the distance measures per say. Rather, we are interested in how the tuning parameters affect the sets of representative trees selected, which are ultimately determined by the relative ranking of the trees based on the distance measures. Tables IV and V present results from selected simulation scenarios in the 25% censored situation. Although not presented here, the results obtained in the other scenarios are consistent with the general patterns mentioned in the following text.

In Table IV, we fix $minsplit = 10$ and vary $ntree$ and neighborhood size. For each distance measure, corresponding to a fixed choice of l (for defining the top l% neighborhood), of the trees that are picked by using the specified $ntree$, we present in Table IV the percent of trees that were also included in the set of trees picked by using reference $ntree = 1000$. Overall, the results are fairly similar across the different models. The metric D_{d_2} is most consistent in terms of picking the same set of representative trees across varying number of trees in the ensemble and neighborhood size. In most instances, the

Statistics in Medicine

Table IV. Simulation results showing the effect of varying $ntree$ and neighborhood size, for fixed value of $minsplit$ (10).

Scenario	*ntree*	Top % trees	D_{d_1}	$D_{d_1^*}$	D_{d_2}	D_{d_0}
Model D for $\lambda, \alpha = 0.5$	250	5	0.69	0.69	0.69	
		10	0.92	0.84	0.96	
		25	0.98	1.00	1.00	
						1.0
	500	5	0.84	0.42	0.88	
		10	0.94	0.84	0.94	
		25	1.00	1.00	0.98	
						1.0
Model E for $\lambda, \alpha = 1$	250	5	0.54	0.78	1.00	
		10	0.76	0.88	1.00	
		25	0.87	0.92	0.98	
						1.0
	500	5	0.88	1.00	1.00	
		10	0.90	1.00	1.00	
		25	0.98	1.00	1.00	
						1.0
Model E for $\lambda, \alpha = 2$	250	5	1.00	0.92	0.92	
		10	1.00	0.80	1.00	
		25	0.98	0.97	1.00	
						1.0
	500	5	0.96	0.92	0.92	
		10	0.88	0.94	0.88	
		25	0.98	0.94	0.98	
						1.0

For each distance measure, corresponding to a fixed top $l\%$ neighborhood, of the trees that are picked by using the specified $ntree$, the percent of trees that were also included in the set of trees picked by using reference $ntree = 1000$ are presented in the table. For D_{d_0}, the table entries are based on comparing the trees in the unique smallest neighborhoods obtained from using the specified $ntree$ and reference $ntree = 1000$. Because of the coarseness of D_{d_0}, the unique smallest neighborhoods range from top 53% to top 79% of the most representative trees in the various scenarios.

metrics D_{d_1} and $D_{d_1^*}$ are able to capture at least 80% of the overlapping trees across varying values of $ntree$ and neighborhood size. The metric D_{d_0}, on the other hand, is quite coarse (as was also observed in our data analysis), and only yields unique smallest neighborhoods in the range of top 53% to top 79% of the most representative trees. Therefore, although results for D_{d_0} (based on comparing the trees in the unique smallest neighborhoods obtained from using the specified $ntree$ and reference $ntree = 1000$) yield 100% overlap in all scenarios, the use of D_{d_0} may be limited in the practical setting.

In general, for a fixed neighborhood size, the set of representative trees picked from an ensemble that was grown using 1000 trees correspond very well to that picked from an ensemble grown using 500 trees, with at least 85% overlap in most scenarios. Representative trees grown from an ensemble using 250 trees had variable overlap (ranging from 54% to 100%) with those grown from an ensemble using 1000 trees. Results were fairly stable across various neighborhood sizes when comparing representative trees grown from an ensemble using 500 trees versus that grown from an ensemble using 1000 trees. Choosing a small neighborhood (top 5%) yielded variable overlap in the sets of representative trees selected, when the ensemble was grown using fewer trees (250).

Table V presents simulation results based on varying the $minsplit$ and neighborhood size, for fixed value of $ntree = 500$. For each distance measure, corresponding to a fixed choice of l (for defining the top $l\%$ neighborhood), the entries in Table V represent the percent overlap in the sets of trees picked using the specified $minsplit$ and reference $minsplit = 10$. Overall, the patterns related to the specific distance measures are similar to that demonstrated in Table IV. For the top 5% neighborhood, the percent overlap in the sets of representative trees ranged from 60% to 96% across varying values of $minsplit$. However, for the top 10% and top 25% neighborhoods, the percent overlap across varying values of $minsplit$ was at least 75% in most scenarios.

Table V. Simulation results showing the effect of varying $minsplit$ and neighborhood size, for fixed value of $ntree$ (500).

Scenario	$minsplit$	Top % trees	D_{d_1}	$D_{d_1^*}$	D_{d_2}	D_{d_0}
Model D for $\lambda, \alpha = 0.5$	15	5	0.96	0.76	0.92	
		10	0.94	0.78	0.92	
		25	0.95	0.94	0.90	
						0.98
	20	5	0.76	0.72	0.76	
		10	0.80	0.76	0.74	
		25	0.90	0.87	0.76	
						0.96
Model E for $\lambda, \alpha = 1$	15	5	0.72	0.76	0.72	
		10	0.80	0.84	0.82	
		25	0.94	0.93	0.88	
						0.98
	20	5	0.60	0.60	0.64	
		10	0.68	0.72	0.66	
		25	0.86	0.82	0.74	
						0.98
Model E for $\lambda, \alpha = 2$	15	5	0.80	0.64	0.92	
		10	0.82	0.88	0.86	
		25	0.89	0.90	0.90	
						0.97
	20	5	0.68	0.68	0.72	
		10	0.70	0.86	0.78	
		25	0.84	0.88	0.80	
						0.93

For each distance measure, corresponding to a fixed top $l\%$ neighborhood, the table entries represent the percent overlap in the sets of trees picked using the specified $minsplit$ and reference $minsplit = 10$. For D_{d_0}, the table entries are based on comparing the trees in the unique smallest neighborhoods obtained from using the specified $minsplit$ and reference $minsplit = 10$. Because of the coarseness of D_{d_0}, the unique smallest neighborhoods range from top 53% to top 79% of the most representative trees in the various scenarios.

We performed a second set of simulations to assess the effect of averaging over neighborhoods of representative trees on predictive accuracy. Specifically, we fixed $ntree = 500$ and $minsplit = 10$ and computed OOB error rates using various neighborhoods. We used the top 5%, 10%, and 25% of the most representative trees, as well as all 500 trees in the ensemble to define various neighborhoods. For these simulations, we started with the same three covariates X_1, X_2, and X_3 as in our first set of simulations and added two additional covariates: X_4 generated from an Exponential(3) distribution, and X_5 generated from a Normal(70, 25) distribution. We considered four different scenarios; the first two corresponding to binary data models, and the third and the fourth corresponding to censored data models. For the first and second (binary outcome) scenarios, we generated data based on the following models: (I) p (probability of success) $= 0.8$ if $X_1 = 1$, $X_2 = 1$;$= 0.6$ if $X_1 = 1$, $X_2 > 1$; $= 0.35$ if $X_1 = 0$, $X_2 = 1$; and $= 0.15$ if $X_1 = 0$, $X_2 > 1$; and (II) $p = 0.05$ if $X_2 = 2$, $X_5 < 67$;$= 0.1$ if $X_2 = 2$, $X_5 \geqslant 67$, $X_1 = 0$; $= 0.3$ if $X_2 = 2$, $X_5 \geqslant 67$, $X_1 = 1$; $= 0.4$ if $X_2 = \{1, 3\}$, $X_4 < 1$; $= 0.6$ if $X_2 = \{1, 3\}$, $1 \leqslant X_4 < 2.5$; $= 0.8$ if $X_2 = \{1, 3\}$, $X_4 \geqslant 2.5$, $X_5 < 71$; $= 0.9$ if $X_2 = \{1, 3\}$, $X_4 \geqslant 2.5$, $X_5 \geqslant 71$. For the third and fourth (censored outcome) scenarios, survival distributions were assumed Weibull: $S(t; \lambda, \alpha) = \exp(-\lambda t^{\alpha})$, with three values chosen for the shape parameter: $\alpha = 0.5$, 1, and 2. We considered the following two models for λ: (III) $\lambda = 1$ if $X_1 = 1, X_2 = 1$; $= 3$ if $X_1 = 1, X_2 > 1$; $=5$ if $X_1 = 0, X_2 = 1$; $=10$ $X_1 = 0, X_2 > 1$; and (IV) $\lambda = 1$ if $X_1 = 0, X_2 = 1, X_3 \leqslant 53$; $= 1.4$ if $X_1 = 0, X_2 = 1, X_3 > 53$; $= 1.8$ if $X_1 = 0$, $X_2 = 2$, $X_3 \leqslant 53$; $= 1.9$ if $X_1 = 0, X_2 = 2, X_3 > 53$; $= 2.6$ if $X_1 = 0, X_2 = 3, X_3 \leqslant 53$; $= 2.7$ if $X_1 = 0, X_2 = 3, X_3 > 53$; $= 3.3$ if $X_1 = 1, X_2 = 1, X_3 \leqslant 53$; $= 3.4$ if $X_1 = 1, X_2 = 1, X_3 > 53$; $= 4.1$ if $X_1 = 1, X_2 = 2, X_3 \leqslant 53$; $= 4.5$ if $X_1 = 1, X_2 = 2, X_3 > 53$; $= 4.8$ if $X_1 = 1, X_2 = 3, X_3 \leqslant 53$; $= 5.0$ if $X_1 = 1, X_2 = 3, X_3 > 53$. We assumed independent censoring with censoring proportions of 0%, 25%, and 50%. For each simulation design scenario, once again we generated samples of size n (number

Statistics in Medicine

of observations) = 500. Table VI presents results from these simulations: models I and II correspond to the aforementioned binary data models, and models III and IV correspond to the aforementioned censored data models with $\alpha = 1$ and 25% censoring. Although not presented here, the results obtained in the other scenarios are consistent with the general patterns mentioned in the following text. The metric D_{d_0} once again only yielded unique smallest neighborhoods of top 53% to top 88% across the various scenarios. Therefore, we do not report the error rates based on D_{d_0} in Table VI.

For binary data, there appears to be gains in predictive accuracy by averaging over neighborhoods of representative trees instead of averaging over all trees in the ensemble. The relative gains in predictive accuracy in model I range from 9.3% to 36.9%, whereas that in model II range from 2.8% to 12.8%. In general, averaging over representative tree neighborhoods chosen on the basis of D_{d_2} yielded the smallest error rates. For model I, the smallest OOB error rate was achieved by averaging over the top 5% neighborhood, whereas that for model II was achieved by averaging over the top 10% neighborhood. For censored data, the relative gains in predictive accuracy are small ranging from 3.3% to 9.8% for model III and $< 1\%$ to 7.9% for model IV. In general, averaging over representative tree neighborhoods chosen on the basis of D_{d_2} yielded the smallest error rates; although in this case, the error rates for other distance metrics were only slightly different, given a fixed neighborhood size.

To the best of our knowledge, no reported studies in the literature comparing various classification tools have performed formal significance testing of error rates [4]. However, in light of a reviewer's comment regarding whether the error rates obtained using representative tree neighborhoods are 'significantly' different from the overall error rate, we extended our simulations as follows. Of the 500 trees grown in the ensemble, we randomly selected sets of 25, 50, and 125 trees (i.e. 5%, 10%, and 25% of the trees) and calculated OOB error rates on the basis of these. For each set of 25, 50, and 125 trees, we repeated the aforementioned step 100 times to obtain a 'null' empirical distribution of the error rate. We then compared the top 5%, 10%, and 25% representative tree neighborhood-based error rates and the overall error rate to this empirical distribution. Although not intended to mimic significance testing, the results are nonetheless interesting. For example, for model II, the error rates based on randomly selected sets of 125 trees ranged from 38.4% to 42.2%, and the median of the distribution was 40.1%. The error rates based on the corresponding (i.e., top 25%) representative tree neighborhoods chosen using D_{d_1}, $D_{d_1^*}$, and D_{d_2} were 38.9%, 37.2%, and 36.5%, respectively. Compared with the empirical distribution, the proportion of times one could obtain an error rate smaller than that obtained using the

Table VI. Simulation results showing the effect of averaging over neighborhoods of representative trees on predictive accuracy, for fixed values of $ntree$ (500) and $minsplit$ (10).

Scenario	Neighborhood (%)	D_{d_1} (%)	$D_{d_1^*}$ (%)	D_{d_2} (%)
Model I	Top 5	41.7	41.7	34.6
	Top 10	44.5	44.5	37.9
	Top 25	49.8	49.8	42.6
	All trees in the ensemble		54.9	
Model II	Top 5	37.6	37.9	35.5
	Top 10	37.4	37.7	34.9
	Top 25	38.9	37.2	36.5
	All trees in the ensemble		40.0	
Model III	Top 5	36.0	36.2	36.0
	Top 10	37.1	37.2	37.1
	Top 25	38.5	38.5	38.6
	All trees in the ensemble		39.9	
Model IV	Top 5	41.2	39.7	40.3
	Top 10	39.2	38.1	38.0
	Top 25	39.3	38.9	38.3
	All trees in the ensemble		41.3	

Entries in the table represent the out-of-bag error rates based on Harrell's concordance index using specified neighborhoods defined by each distance measure, as well as all trees in the ensemble.

representative tree neighborhoods was 0.03, < 0.01, and < 0.01, respectively. In contrast, the overall error rate of 40.0% for this model (based on all 500 trees in the ensemble) was right around the median of the empirical distribution. Coupled together, these observations seem to suggest that it is possible to improve the overall error rate by averaging over the 'right' subset of trees, and representative tree neighborhoods offer such a solution. Once again, this result was most consistent for representative tree neighborhoods chosen on the basis of D_{d_2} across varying neighborhood sizes and scenarios.

8. Conclusions

It is well recognized that the mechanism of selecting a best split and the recursive partitioning of data leading to smaller and smaller datasets can lead to instability in a tree structure. Ensemble methods such as bagging and random forests can be useful in understanding the stability of tree structures and improving predictive performance. However, individual trees are lost in the process. In this paper, we provided a methodology for identifying the most representative trees in an ensemble. Our ideas are applicable to classification as well as regression trees.

Because our focus is on measuring similarity between trees in the ensemble, we wanted the comparison between any two trees to be unaffected by perturbations such as random covariate selection at each step of the splitting process. Using all covariates at each step of the splitting process puts the comparison on equal grounds. By defining several distance metrics that capture architectural as well as prediction similarity between trees, we summarize an ensemble by several representative trees. The most representative tree among all trees in the ensemble is the one with the smallest value of $D(T)$, that is, the average distance between tree T and all other trees in the ensemble.

Earlier work by other authors have considered related approaches. Shannon and Banks [22] proposed a tree metric, which accounts for the manner in which the tree is constructed. This metric compares rules at nodes in the same position in two trees. That is, if two plots are constructed on transparent paper so that nodes in the same position overlap and the plots are held up to the light, the metric counts the number of nodes at which the splitting rules are discrepant. The distance between trees is then a weighted sum of the discrepancies at each location. This metric can capture tree structure, but two isomorphic trees with the same splits in a different order will be identified as dissimilar by this metric. Furthermore, Shannon and Banks' approach assumes a unimodal probability distribution on the ensemble of classification trees, parameterized by a central tree structure representing the true model and a precision or concentration coefficient representing the variability around the central tree. The authors propose the maximum likelihood estimate of the central tree as the best tree to represent the set. Therefore, in contrast to our approach, Shannon and Banks' final tree may not necessarily be one of the observed trees in the ensemble.

Our approach is closer to that of Chipman *et al.* [21] in that trees are clustered on the basis of distance metrics. In fact, our d_1 and d_2 metrics are equivalent to their partition and fit metrics, respectively. Chipman *et al.* [21] used these to explore the set of plausible models on the basis of a multidimensional scaling (MDS) plot of the distance matrix. Using the MDS plot, they identify groups of similar tree models. In contrast, we focus on selecting representative trees for the entire ensemble on the basis of ranking the trees according to a distance metric. This allows us to pick the single 'most representative' tree, as well as other trees in the neighborhood. We argue that understanding the most representative tree as well as the neighborhood are useful. If a single model is to be selected, then one would choose the tree with the smallest value of the distance metric. On the other hand, by selecting a set of representative trees (i.e., neighborhood), and averaging over them, rather than averaging over all trees in the ensemble, yields a compromise between selection and averaging.

We computed error rates on the basis of the OOB ensemble estimate. Because ultimately our goal was to assign error rates to the representative trees, we computed the OOB estimate using neighborhoods of trees that are deemed similar based on $D(T)$. The error rates obtained from averaging over representative tree neighborhoods were smaller in comparison with the error rates obtained from averaging over all trees in the ensemble. With our simulations and data examples, averaging over neighborhoods of representative trees yielded gains in predictive accuracy for binary data scenarios. However, the difference in error rates obtained using all trees versus representative trees in the ensemble was relatively small in the censored data scenarios.

Our proposed approach can be applied only on ensembles in which the trees are independent samples from some distribution. The trees we generate by bootstrapping are 'independent' in that the structure in one member of the ensemble is not influenced by the structure of another member. Sampling from

a posterior on the space of trees would be another example of an ensemble with independent elements. However, an ensemble arising from boosting is quite different; data are iteratively reweighted, with more weight given at successive iterations to observations that are predicted poorly. This introduces a stochastic element to the search and successive trees attempt to improve upon trees identified at earlier iterations. One may not necessarily want to select a 'representative' tree from such an ensemble because individually, each boosted tree contributes a small amount to the prediction and the output of a boosted ensemble would require all trees.

A natural question that arises is how many representative trees do we ultimately present to the clinician who is interested in a clinical decision making tool. It would seem most logical to present the single most representative tree that has the smallest value of $D(T)$ based on a specific metric. Although our simulations would favor choosing D_{d_2} in terms of stability of the sets of representative trees selected, as well as predictive accuracy, we argue that the choice of 'which metric to use' is somewhat contextual and depends on the goal of the study. Thus, barring D_{d_0} (which appears to be rather coarse), any of the other three metrics could be appropriate choices depending on the context. Yet another option would be to combine different aspects of tree similarity by taking the average of $D_{d_1^*}$ and D_{d_2} to yield an overall score. The tree with the smallest value of the overall score could be chosen as the single most representative tree. Depending on the size of the ensemble, we recommend using a neighborhood of the top 5%–10% most representative trees for obtaining error rate and predictions for this tree. This approach allows us to extract the most representative tree from the plethora of tree models while still borrowing the strength from its 'likes' for prediction purposes. Lastly, there may be situations where alternative competing models are of interest, and in such scenarios, examining the top 10 or 20 most representative trees may be useful towards elucidation of such models.

Acknowledgement

Dr. Banerjee's research was supported by grant P30-CA46592-05 from the NCI.

References

1. Breiman L, Friedman JH, Olshen RA, Stone CJ. *Classification and Regression Trees*. Wadsworth: Belmont, California, 1984.
2. Zhang H, Singer B. *Recursive Partitioning in the Health Sciences*. Springer: New York, 1999.
3. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning*. Springer: New York, 2001.
4. Cutler A, Cutler DR, Stevens JR. Tree-based methods. In *High-Dimensional Data Analysis in Cancer Research*, Li X, Xu R (eds). Springer, Inc: New York, 2009; 83–101.
5. LeBlanc M, Crowley J. Relative risk trees for censored survival data. *Biometrics* 1992; **48**:411–425.
6. LeBlanc M. Tree-based methods for prognostic stratification. In *Handbook of Statistics in Clinical Oncology*, Crowley J (ed.). Marcel Dekker, Inc: New York, 2001; 457–472.
7. Keles S, Segal MR. Residual-based tree-structured survival analysis. *Statistics in Medicine* 2002; **21**:313–326.
8. Banerjee M, Noone A. Tree-based methods for survival data. In *Statistical Advances in the Biomedical Sciences*, Biswas A, Datta S, Fine JP, Segal MR (eds). John Wiley & Sons: New Jersey, 2008; 265–285.
9. Zhang H, Yu CY, Singer B, Xiong MM. Recursive partitioning for tumor classification with gene expression microarray data. *Proceedings of the National Academy of Sciences* 2001; **98**:6730–6735.
10. Banerjee M, Biswas D, Sakr W, Wood DP, Jr. Recursive partitioning for prognostic grouping of patients with clinically localized prostate carcinoma. *Cancer* 2000; **89**:404–411.
11. Katz A, Buchholz TA, Thames H, Smith CD, McNeese MD, Theriault R, Singletary SE, Strom EA. Recursive partitioning analysis of locoregional recurrence patterns following mastectomy: implications for adjuvant irradiation. *International Journal of Radiation Oncology, Biology, Physics* 2001; **50**:397–403.
12. Freedman GM, Hanlon AL, Fowble BL, Anderson PR, Nicoloau N. Recursive partitioning identifies patients at high and low risk for ipsilateral tumor recurrence after breast-conserving surgery and radiation. *Journal of Clinical Oncology* 2002; **20**:4015–4021.
13. Banerjee M, George J, Song EY, Roy A, Hryniuk W. Tree-based model for breast cancer prognostication. *Journal of Clinical Oncology* 2004; **22**:2567–2575.
14. Segal M, Barbour JD, Grant RM. Relating HIV-1 sequence variation to replication capacity via trees and forests. *Statistical Applications in Genetics and Molecular Biology* 2004; **3**(Iss. 1). Article 2.
15. Breiman L. Bagging predictors. *Machine Learning* 1996; **24**:123–140.
16. Breiman L. Random forests. *Machine Learning* 2001; **45**:5–32.
17. Ishwaran H, Blackstone EH, Pothier CE, Lauer MS. Relative risk forests for exercise heart rate recovery as a predictor of mortality. *Journal of the American Statistical Association* 2004; **99**:591–600.
18. Ishwaran H, Kogalur UB. Random survival forests for R. *Rnews* 2007; **7**(2):25–31.
19. Quinlan J. Bagging, boosting, and C4.5. In *Proceedings Thirteenth American Association for Artificial Intelligence National Conference on Artificial Intelligence*. CA. AAAI Press: Menlo Park, 1996; 725–730.

20. Freund Y, Schapire RE. Experiments with a new boosting algorithm. *In Proceedings of the Thirteenth International Conference on Machine Learning*, Bari, Italy, 1996; 148–156.
21. Chipman HA, George EI, McCulloch RE. Managing multiple models. In *Artificial Intelligence and Statistics*, Jaakkola T, Richardson T (eds). Morgan Kaufmann, 2001; 11–18.
22. Shannon WD, Banks D. Combining classification trees using MLE. *Statistics in Medicine* 1999; **18**:727–740.
23. Harrell FE, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. *The Journal of the American Medical Association* 1982; **18**:2543–2546.
24. May M, Royston P, Egger M, Justice AC, Sterne J. Development and validation of a prognostic model for survival time data: application to prognosis of HIV positive patients treated with antiretroviral therapy. *Statistics in Medicine* 2004; **23**:2375–2398.
25. Therneau TM, Atkinson B. An introduction to recursive partitioning using the RPART routines. *Technical report*, Mayo Foundation, 1997.
26. Therneau TM, Atkinson B. RPART: recursive partitioning, 2007. R port by Brian Ripley <ripley@stats.ox.ac.uk>. R package version 3.1-36.
27. Graf E, Schmoor C, Sauerbrei W, Schumacher M. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine* 1999; **18**:2529–2545.