Wiley StatsRef:
Statistics Reference Online
DOI: 10.1002/ISBN.stat00999.pub9

Random Survival Forests

Hemant Ishwaran[a]*and Min Lu[a]

Keywords: ensembles; Kaplan-Meier estimator; machine learning; Nelson-Aalen CHF; survival

Abstract: Random survival forests (RSF) is a flexible nonparametric tree-ensemble method for the analysis of right-censored survival data. In this article we provide a short overview of RSF. We review survival splitting rules for growing random survival trees, in-bag and out-of-bag (OOB) ensemble estimators, prediction performance, variable importance, and partial plots. We also briefly describe the extension of RSF to competing risks. Copyright © 2018 John Wiley & Sons, Ltd.

1. Introduction

Random forests (RF) (stat06520) is a popular tree-ensemble method introduced by Leo Breiman [3] with broad applications to machine learning and statistics. It is well known that constructing ensembles by averaging base learners, such as trees, can substantially improve prediction performance. RF builds on this concept by injecting further randomization into the base learning process. Specifically, randomization is introduced in two forms. First, a randomly drawn bootstrap sample of the data is used to grow a tree. Second, at each node of the tree, a randomly selected subset of variables is chosen as candidates for splitting. The purpose of this two-step randomization is to decorrelate trees, which encourages low variance for ensemble due to the property of bagging [2]. Furthermore, RF trees are typically grown very deeply; in fact, Breiman's original RF classifier called for growing a classification tree to purity (one observation per terminal node). The use of deep trees, a bias reduction technique, when combined with reduced variance due to averaging and randomization, enables RF to approximate rich classes of functions while maintaining low generalization error.

Early applications of RF focused on regression and classification problems. Random survival forests [22] (RSF) was introduced to extend RF to the setting of right-censored survival data. Implementation of RSF follows the same general principles as RF: (a) Survival trees are grown using bootstrapped data; (b) Random feature selection is used when splitting tree nodes; (c) Trees are generally grown deeply, and (d) The survival forest ensemble is calculated by averaging tree survival predictors.

The RSF algorithm can be broadly described as follows:

1. Draw `ntree` bootstrap samples from the original data.

[a]**University of Miami Medical School, Department of Public Health Sciences, Division of Biostatistics**
*__Email: hemant.ishwaran@gmail.com__

Prepared using WileySTAT.cls [Version: v1.00]

2. Grow a survival tree for each bootstrapped dataset. At each node of the tree randomly select `mtry` variables for splitting on. Split on the variable that optimizes a chosen survival splitting criterion.
3. Grow the tree to full size under the constraint that a terminal node should have no less than `nodesize` unique cases. Calculate the tree predictor.
4. Calculate in-bag and out-of-bag (OOB) ensemble estimators by averaging the tree predictors.
5. Use the OOB estimator to estimate out-of-sample prediction performance.
6. Use OOB estimation to calculate variable importance.

Although Cox's proportional hazard regression method [7] is very popular for time-to-event data analysis, RSF has become attractive as a nonparametric method with less restrictive model assumptions. Some of RSF's important properties are that: (a) It is fully nonparametric and can identify survival risk factors without assuming parametric relationship (linear or nonlinear) or prior knowledge of interactions among variables; (b) It is robust to outliers and does not suffer from convergence problem; (c) It can be used for high dimensional data; (d) It offers out-of-bag (cross-validated) prediction that does not overfit the data and therefore can be used for reliable inference of the training data; and (e) It provides a fully nonparametric variable importance measure of a variables' contribution to predicting survival.

In the following sections we outline the steps in applying the RSF algorithm and illustrate its use with examples. The extension of RSF to competing risks is also briefly coverered. All examples are illustrated using the R-package, `randomForestSRC` [21]. Finally we note that this article focuses only on RSF as defined in [22] which strictly adheres to the RF approach described by Breiman. For other ensemble survival approaches see [16,17,32].

2. Splitting Rules for Growing a RSF Tree

The presence of censoring is a unique feature of survival data that complicates certain aspects of implementing RSF. In right-censored survival data the observed data is (T, δ) where T is time and δ is the censoring indicator. The observed time T is defined as the minimum of the true (potentially unobserved) survival event time T^o and the true (potentially unobserved) censoring time C^o; thus $T = \min(T^o, C^o)$ and the actual event time might not be observed. The censoring indicator is defined as $\delta = 1\{T^o \leq C^o\}$. When $\delta = 1$, an event has occurred (i.e., death has occurred) and we observe the true event time, $T = T^o$. Otherwise when $\delta = 0$, the observation is censored and we only observe the censoring time $T = C^o$: thus we know that the subject has survived to time C^o, but not when the subject actually dies. The true event time being subject to censoring must be dealt with when growing a RSF tree. In particular, the splitting rule for growing the tree must specifically account for this censoring. In this section we discuss survival splitting used to grow a RSF tree. We begin by discussing RF trees.

2.1. RF trees

The basic unit of RF (the so-called base learner) is a binary tree (stat07514.pub2) constructed using recursive partitioning (RPART). The RF tree base learner is grown using the approach of CART [5] (classification and regression tree), a method in which binary splits recursively partition the tree into homogeneous or near-homogeneous terminal nodes (the ends of the tree). A good binary split pushes data from a parent tree-node to its two daughter nodes so that the ensuing homogeneity in the daughter nodes is improved from the parent node. Common splitting rules used for CART are the within node sum of squares for regression and the Gini splitting criterion for classification. Note that while RF trees are grown using CART's approach, RF trees differ as they are grown nondeterministically using bootstrapping and random feature selection. As discussed, this randomization decorrelates trees and reduces variance.

Futhermore, while CART typically grows shallow trees to avoid overfitting, RF trees are generally grown deeply in order to reduce bias.

2.2. RSF trees

RSF trees are also grown by applying RPART. Because RSF deals with event history (survival) data, the goal is to split the tree node into left and right daughters with dissimilar event history (survival) behavior. This is accomplished by using an appropriate splitting rule.

2.2.1. Log-rank splitting

One of the most popular splitting rules is the log-rank test statistic. Traditonally the log-rank test is used for two-sample testing with survival data, but it can be employed for survival splitting as a means for maximizing between-node survival differences[6,29,30,25,26].

To explain log-rank splitting, let h denote the tree node to be split. Without loss of generality let h be the root node (top of the tree). For simplicity assume the data is not bootstrapped, and denote the data by $(T_1, \mathbf{X}_1, \delta_1), \ldots, (T_n, \mathbf{X}_n, \delta_n)$ where $\mathbf{X}_i$ is $i's$ feature vector (covariate) and T_i and δ_i are the observed time and censoring indicators for i. Let X denote a specific variable (i.e., coordinate of the feature vector $\mathbf{X}$). A proposed split using X is of the form $X \leq c$ and $X > c$ (for simplicity we assume X is nominal) and splits h into left and right daughters, L and R, respectively. Let

$$t_1 < t_2 < \cdots < t_m$$

be the distinct death times and let $d_{j,L}, d_{j,R}$ and $Y_{j,L}, Y_{j,R}$ equal the number of deaths and individuals at risk at time t_j in daughter nodes L, R. At risk means the number of individuals in a daughter who are alive at time t_j, or who have an event (death) at time t_j:

$$Y_{j,L} = \#\{T_i \geq t_j,\ X_i \leq c\}, \qquad Y_{j,R} = \#\{T_i \geq t_j,\ X_i > c\}.$$

Define

$$Y_j = Y_{j,L} + Y_{j,R}, \qquad d_j = d_{j,L} + d_{j,R}.$$

The log-rank split-statistic value for the split $L = \{X_i \leq c\}$ and $R = \{X_i > c\}$ is

$$L(X, c) = \frac{\displaystyle\sum_{j=1}^{m} \left(d_{j,L} - Y_{j,L} \frac{d_j}{Y_j} \right)}{\sqrt{\displaystyle\sum_{j=1}^{m} \frac{Y_{j,L}}{Y_j} \left(1 - \frac{Y_{j,L}}{Y_j} \right) \left(\frac{Y_j - d_j}{Y_j - 1} \right) d_j}}.$$

The value $|L(X, c)|$ is a measure of node separation. The larger the value, the greater the survival difference between L and R, and the better the split is. The best split is determined by finding the feature X^* and split-value c^* such that $|L(X^*, c^*)| \geq |L(X, c)|$ for all X and c.

2.2.2. Other splitting rules (deterministic and randomized)

Other splitting approaches for survival trees include rules based on measures of impurity for survival data[8,12]. Another promising approach are split-statistics based on the Brier score[13]. This may be prefered to log-rank splitting when

censoring depends strongly on $\mathbf{X}$. Doubly robust split-statistics[31] have also been proposed for further robustification in complex survival settings. However, with these latter statistics there is a heavier computational cost in their implementation. In general, split-statistics for survival trees are far more computationally demanding than those used for RF regression and classification. This is true even for the log-rank split-statistic. One successful method for reducing computational expense is to employ randomized splitting rules[22,23,19]. Rather than splitting the node by considering all possible split-values for a variable, instead a fixed number of randomly selected split-points $c_1, \ldots, c_{\texttt{nsplit}}$ are chosen. For example, the best randomized split using log-rank splitting is the maximal value of

$$|L(X, c_1)|, \ldots, |L(X, c_{\texttt{nsplit}})|.$$

For each variable X, this reduces n split-statistic evaluations (worst case scenario) to $\texttt{nsplit} \ll n$ evaluations. Not only does randomized splitting greatly reduce computations, it also mitigates the well known tree bias of favoring splits on variables with a large number of split-points, such as continuous variables or factors with a large number of categorical labels[27]. Related work includes[11] who investigated extremely randomized trees. Here a single random split-point is chosen for each variable (i.e., $\texttt{nsplit} = 1$).

3. Survival Tree Predictor

RSF estimates the survival function, $S(t|\mathbf{X}) = \mathbb{P}\{T^o \geq t|\mathbf{X}\}$, and the cumulative hazard function (CHF),

$$H(t|\mathbf{X}) = \int_{(0,t]} \frac{F(du|\mathbf{X})}{S(u|\mathbf{X})}, \qquad F(u|\mathbf{X}) = \mathbb{P}\{T^o \leq u|\mathbf{X}\}.$$

In this section we review how these two quantities are estimated using a survival tree.

3.1. In sample (in-bag) estimators

Once the survival tree is grown, the ends of the tree are called the terminal nodes. The survival tree predictor is defined in terms of the predictor within each terminal nodel. Let h be a terminal node of the tree and let

$$t_{1,h} < t_{2,h} < \cdots < t_{m(h),h}$$

be the unique death times in h and let $d^*_{j,h}$ and $Y^*_{j,h}$ equal the number of deaths and individuals at risk at time $t_{j,h}$ (we use the superscript * here to emphasize these values are bootstrapped due to the survival tree being constructed from bootstrapped data). The CHF and survival functions for h are estimated using the bootstrapped Nelson-Aalen and Kaplan-Meier estimators (stat06004.pub2):

$$H^*_h(t) = \sum_{t_{j,h} \leq t} \frac{d^*_{j,h}}{Y^*_{j,h}}, \qquad S^*_h(t) = \prod_{t_{j,h} \leq t} \left(1 - \frac{d^*_{j,h}}{Y^*_{j,h}}\right).$$

The survival tree predictor is defined by assigning all cases within h the same CHF and survival estimate. This makes sense because the purpose of the survival tree is to partition the data into homogeneous groups (i.e., terminal nodes) of individuals with similar survival behavior. To estimate $H(t|\mathbf{X})$ and $S(t|\mathbf{X})$ for a given feature $\mathbf{X}$, drop $\mathbf{X}$ down the tree. Because of the binary nature of a tree, $\mathbf{X}$ will fall into a unique terminal node h. The CHF and survival estimator for $\mathbf{X}$ is the bootstrapped Nelson-Aalen and Kaplan-Meier estimator for $\mathbf{X}$'s terminal node:

$$H^*(t|\mathbf{X}) = H^*_h(t), \quad S^*(t|\mathbf{X}) = S^*_h(t), \text{ if } \mathbf{X} \in h.$$

Because the above estimators are based on bootstrap data, we refer to them as in-sample or in-bag estimators.

3.2. Out-of-bag (OOB) estimators

Each survival tree is calculated using a bootstrap sample of the original data. On average a bootstrap leaves out 36.8% of the data. This data is not used to grow the tree and represents out-of-sample data that can be used for cross-validation purposes. This data is called out-of-bag (OOB).

To define the OOB estimator it will be convenient to define an indicator I_i which indicates whether case i is in-bag or OOB. Let $I_i = 1$ if i is OOB, otherwise set $I_i = 0$ if i is in-bag. The OOB CHF and survival estimators for an OOB case is determined by the cases' terminal node membership. Drop i down the tree and let h denote i's terminal node. The OOB CHF and survival estimators for i are

$$H^{**}(t|\mathbf{X}_i) = H^*_h(t), \quad S^{**}(t|\mathbf{X}_i) = S^*_h(t), \text{ if } \mathbf{X}_i \in h \text{ and } I_i = 1.$$

In the above we use the superscript ** to emphasize that estimators are OOB.

4. Ensemble CHF and Survival Function

The ensemble CHF and survival function are determined by averaging the tree estimator. Let $H^*_b(t|\mathbf{X})$ and $S^*_b(t|\mathbf{X})$ be the in-bag CHF and survival estimator for the bth survival tree. The in-bag ensemble estimators are

$$\bar{H}^*(t|\mathbf{X}) = \frac{1}{\texttt{ntree}} \sum_{b=1}^{\texttt{ntree}} H^*_b(t|\mathbf{X}), \quad \bar{S}^*(t|\mathbf{X}) = \frac{1}{\texttt{ntree}} \sum_{b=1}^{\texttt{ntree}} S^*_b(t|\mathbf{X}).$$

Likewise, the OOB ensemble is calculated by averaging the OOB tree estimators. Let $O_i = O(\mathbf{X}_i)$ record trees where case i is OOB. The OOB ensemble estimators are

$$\bar{H}^{**}(t|\mathbf{X}_i) = \frac{1}{|O_i|} \sum_{b \in O_i} H^*_b(t|\mathbf{X}_i), \quad \bar{S}^{**}(t|\mathbf{X}_i) = \frac{1}{|O_i|} \sum_{b \in O_i} S^*_b(t|\mathbf{X}_i).$$

An important distinction between the two sets of estimators is that OOB estimators are used for inference on the training data and for estimating prediction error and only apply to features $\mathbf{X} = \mathbf{X}_i$ in the original data. In-bag estimators on the other hand are used for prediction and can be used for any feature $\mathbf{X}$.

To illustrate, we used the survival data from [18] consisting of 2231 adult patients with systolic heart failure. All patients underwent cardiopulmonary stress testing. During a mean follow-up of 5 years (maximum for survivors, 11 years), 742 patients died. The outcome is all-cause mortality and a total of $p = 39$ covariates were measured for each patient including demographic, cardiac and noncardiac comorbidity, and stress testing information. Figure 1 displays the (in-bag) predicted survival functions for two hypothetical individuals, where all p features of the two individuals are set to the median level except for the variable peak VO_2. For one of the individuals this is set at the 25th quantile for peak VO_2 (peak $VO_2 = 12.8$ mL/kg per min) and shown using a solid black line. For the other individual this was set to the 75th quantile (peak $VO_2 = 19.3$ mL/kg per min) and shown using a dashed red line.

5. Prediction Performance

Prediction error can be evaluated using Harrell's concordance index [15]. The C-index (concordance index) is related to the area under the ROC curve (stat05255). Through all permissible pairs of individuals over the data, it estimates the probability that the individual who experienced the event first had a worse predicted outcome. Here, we compare the

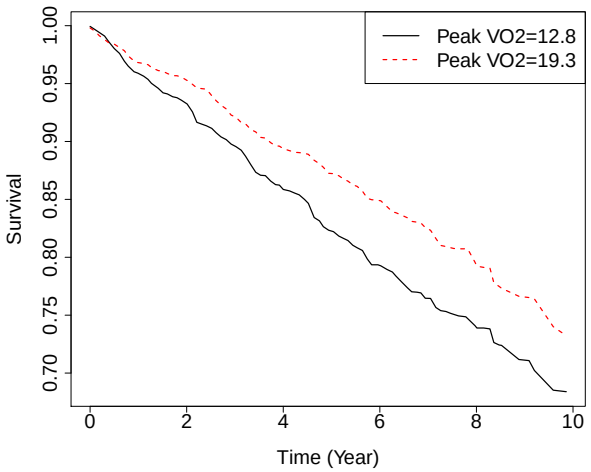

Figure 1. Predicted survival functions for two hypothetical individuals from RSF analysis of systolic heart failure data. Solid black line represents individual with peak $VO_2 = 12.8$ mL/kg per min. Red dash line represents individual with Peak $VO_2 = 19.3$ mL/kg per min. All other variables for both individuals are set to the median value.

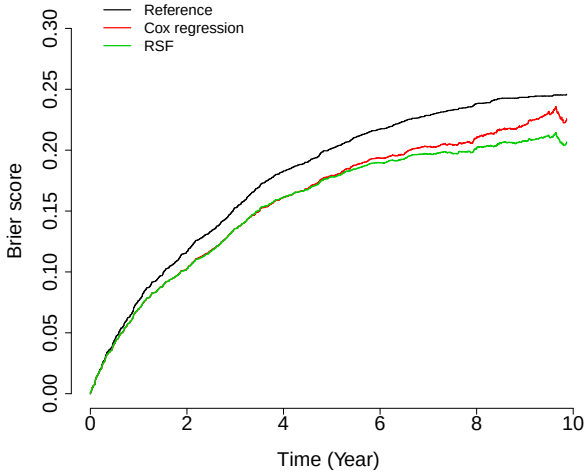

Figure 2. Brier score averaged through 10 runs of 5-fold cross-validation for the systolic heart failure data[18]. RSF is compared to Cox regression.

OOB predicted outcome. To rank two cases i and j with features $\mathbf{X}_i$ and $\mathbf{X}_j$, we say i has a worse predicted outcome than j if[22]

$$\sum_{l=1}^{m} \bar{H}^{**}(t_l|\mathbf{X}_i) > \sum_{l=1}^{m} \bar{H}^{**}(t_l|\mathbf{X}_j),$$

where $t_1 < t_2 < \cdots < t_m$ are the unique event times. The left- and right-hand sides denote the OOB mortality for i and j which are measures reflecting number of expected deaths if all data points had the same features as $\mathbf{X}_i$ or $\mathbf{X}_j$ (see[22] for more details about mortality). The OOB prediction error (PE) is defined as 1 minus the C-index. A value of 0.5 indicates prediction no better than random guessing.

The Brier score, $\mathrm{BS}(t)$, is another popular measure used to assess prediction peformance. Let $\hat{S}(t|\mathbf{X})$ be some estimator of the survival function. To estimate the prediction performance of $\hat{S}$, let $\hat{G}(t|\mathbf{X})$ be a prechosen estimator of the censoring survival function, $G(t|\mathbf{X}) = \mathbb{P}\{C^o \geq t|\mathbf{X}\}$. The Brier score for $\hat{S}(t|\mathbf{X})$ can be estimated by[10]

$$\widehat{\mathrm{BS}}(t) = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\hat{S}^2(t|\mathbf{X}_i)I\{T_i \leq t\}\delta_i}{\hat{G}(T_i - |\mathbf{X}_i)} + \frac{\left(1-\hat{S}(t|\mathbf{X}_i)\right)^2 I\{T_i > t\}}{\hat{G}(t|\mathbf{X}_i)}\right\}.$$

The integrated Brier score at time τ is defined as

$$\mathrm{IBS}(\tau) = \frac{1}{\tau}\int_0^{\tau} \mathrm{BS}(t)dt$$

which can be estimated by substituting $\widehat{\mathrm{BS}}(t)$ for $\mathrm{BS}(t)$. Lower values for the Brier score indicate better prediction performance. Figure 2 displays $\widehat{\mathrm{BS}}(t)$ for the systolic heart failure data[18] for Cox regression and RSF. The reverse Kaplan-Meier estimator was used to estimate the censoring distribution. RSF outperforms Cox regression in this case because it yields an overall lower integrated Brier score.

6. Variable Importance

RSF provides a fully nonparametric measure of variable importance (VIMP). The most common measure is Breiman-Cutler VIMP [4] and is called permutation importance. VIMP calculated using permutation importance adopts a prediction based approach by measuring prediction error attributable to the variable. A clever feature is that rather than using cross-validation, which can be computationally expensive, permutation importance makes use of OOB estimation. Specifically, to calculate the VIMP for a variable X, we randomly permute the OOB values of X in a tree (the remaining coordinates of $\mathbf{X}$ are not altered). The perturbed OOB data is dropped down the tree and the OOB error for the resulting tree predictor determined. The amount by which this new error exceeds the original OOB error for the tree equals the tree importance for X. Averaging over trees yields permutation importance for X.

Large positive VIMP indicates high predictive ability while zero or negative values identify noise variables. Subsampling [24] can be used to estimate the standard error and to approximate the confidence intervals for VIMP. Figure 3 displays delete-d jackknife 99% asymptotic normal confidence intervals for the $p = 39$ variables from the systolic heart failure RSF analysis. Prediction error was calculated using the C-index.

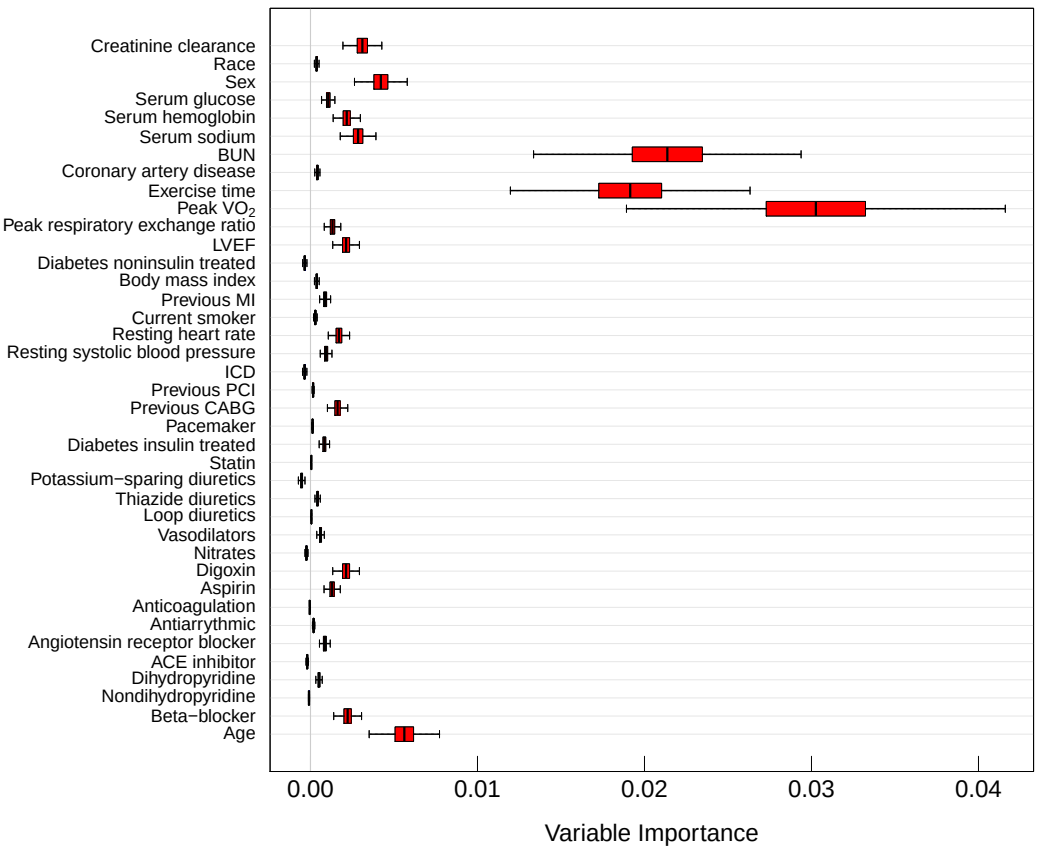

Figure 3. Delete-d jackknife 99% asymptotic normal confidence intervals of VIMP from RSF analysis of systolic heart failure data. Prediction error is defined using Harrell's concordance index.

7. Partial Plots

Another useful tool for interpreting the results from a RSF analysis is the partial dependence plot [9]. Figure 4 displays the partial dependence plot for the most important variable, peak VO_2 from our previous analysis (i.e., the variable with the largest VIMP from Figure 3). The figure displays 5 year OOB survival as a function of peak VO_2; in particular observe that survival depends strongly on its value, increasing in value with increasing peak VO_2 capacity.

An important feature of the partial dependence plot is that it displays the dependence of survival on the target variable while adjusting for all other variables. This is accomplished by integrating out the effect of the other variables. Specifically, let $\bar{S}^{**}(t|\mathbf{X}_i, X_i = x)$ be the OOB ensemble survival function for $\mathbf{X}_i$ where X_i represents the peak VO_2 value and the observed value of X_i is replaced by some prechosen value x. In other words, patient i's variables are set to their observed values except peak VO_2 which is fixed at x. The OOB partial predicted survival function for peak VO_2 at x equals

$$\bar{S}^{**}_X(t|x) = \frac{1}{n}\sum_{i=1}^{n} \bar{S}^{**}(t|\mathbf{X}_i, X_i = x).$$

The value $\bar{S}^{**}_X(t|x)$ is what is displayed on the vertical axis of Figure 4 for $t = 5$ years as x is varied.

Partial dependence plots can be defined for more than one variable. For example, if the target variables are $X^{(k)}$ and $X^{(l)}$, the OOB partial predicted survival function at $X^{(k)} = a$ and $X^{(l)} = b$ equals

$$\bar{S}^{**}_{X^{(k)},X^{(l)}}(t|a, b) = \frac{1}{n}\sum_{i=1}^{n} \bar{S}^{**}(t|\mathbf{X}_i, X_i^{(k)} = a, X_i^{(l)} = b).$$

Figure 5 displays the partial dependence plot of peak VO_2 and BUN which are the top two variables from our previous analysis. The contour plot shows how 5 year OOB survival depends jointly on these two variables. In particular, low peak VO_2 combined with high BUN yields poor survival, wheras high peak VO_2 combined with low BUN yields improved survival.

8. Competing Risks

In this section we briefly review the extension of RSF to competing risks developed in [20]. In competing risks, unlike survival where there is only one event type, the individual is subject to $J > 1$ competing risks (stat03948). As in survival data, a complication is that the individual can be right-censored. Formally, let T^o be the true event time and let $\delta^o \in \{1, \ldots, J\}$ record the event type. Let C^o denote the true censoring time. Under the presence of right-censoring we only observe $T = \min(T^o, C^o)$ and the censoring indicator $\delta = \delta^o \cdot I\{T^o \leq C^o\}$. Thus for each individual one either observes the time an event occurs $T = T^o$ and the type of event which occured $\delta = \delta^o \in \{1, \ldots, J\}$. Otherwise if the individual is right-censored, we observe the censoring time $T = C^o$ and the censoring indicator is $\delta = 0$.

8.1. Competing risk splitting rules

There are three splitting rules used by RSF to grow a competing risk tree [20]:

(1) Generalized log-rank test. This tests for equality of the event-specific hazard functions and is most appropriate when the analysis focuses on determining risk factors for an event-specific hazard. The generalized log-rank test is based on the weighted difference of the Nelson-Aalen event-specific CHF estimates in the daughter nodes.

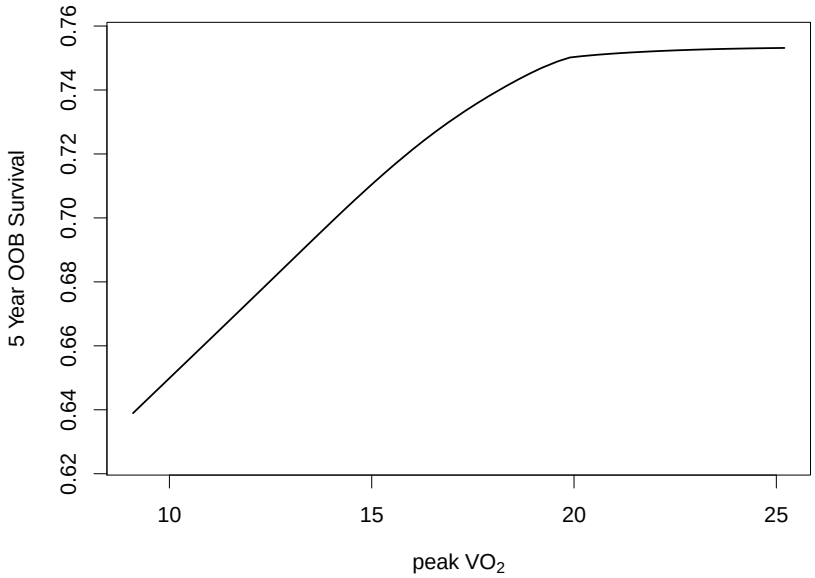

Figure 4. *Partial dependence plot displaying 5 year OOB survival as a function of peak VO_2.*

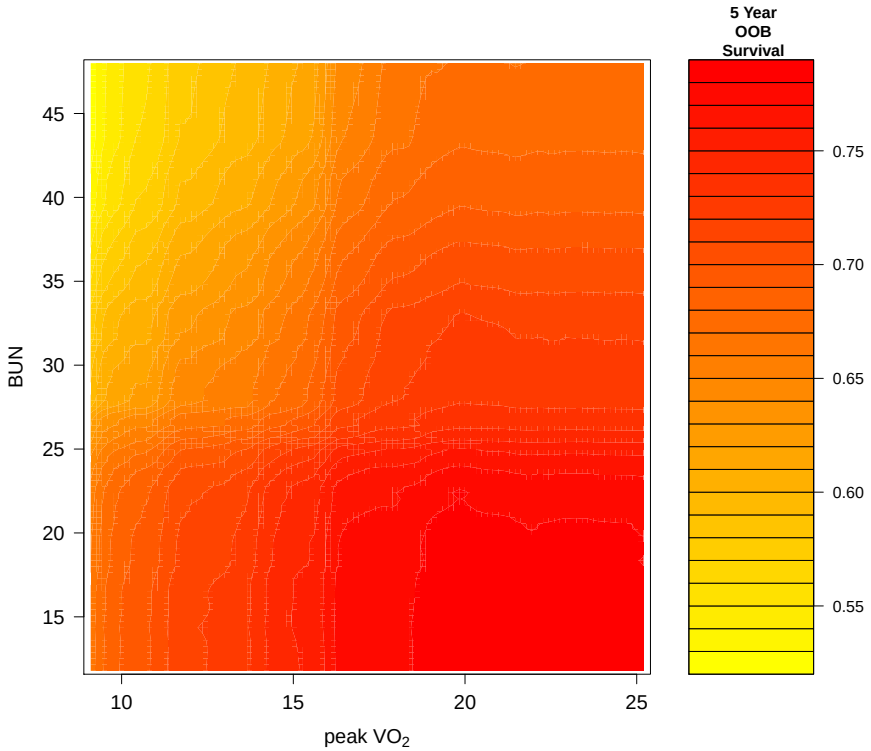

Figure 5. *Partial dependence plot displaying 5 year OOB survival as a function of peak VO_2 and BUN.*

(2) Gray's test. This is a modification of Gray's test [14] and tests for the equality of the cause-specific cumulative incidence functions (CIF). This is most appropriate when the goal is long term probability prediction.
(3) Composite (weighted) splitting rule. This averages the generalized log-rank test or Gray's test across the J event types. This is used if the aim is to predict the CIF of all events simultaneously or if interest is in identifying variables important for any of the J events.

8.2. Event-specific ensembles

Let $(T_i, \delta_i, \mathbf{X}_i)_{1\le i\le n}$ denote the data where T_i is the observed time, $\delta_i \in \{0, 1, \ldots, J\}$ is the observed censoring indicator, and $\mathbf{X}_i$ is the feature. Let $c_{i,b}$ be the number of times case i occurs in bootstrap sample of the bth tree. Let $h_b(\mathbf{X})$ be the terminal node of bth tree containing $\mathbf{X}$. Denote in-bag node-specific event counts by $N^*_{j,b}(t|\mathbf{X}) = \sum_{i\in h_b(\mathbf{X})} c_{i,b} I\{T_i \le t, \delta_i = j\}$ and in-bag number at risk by $Y^*_b(t|\mathbf{X}) = \sum_{i\in h_b(\mathbf{X})} c_{i,b} I\{T_i \ge t\}$. The tree estimator for the event-specific CIF, $F_j(t|\mathbf{X}) = \mathbb{P}\{T^o \le t, \delta^o = j|\mathbf{X}\}$, is the bootstrapped Aalen-Johansen estimator [1]:

$$F^*_{j,b}(t|\mathbf{X}) = \int_{(0,t]} S^*_b(u - |\mathbf{X})Y^*_b(u|\mathbf{X})^{-1} N^*_{j,b}(du|\mathbf{X}), \quad j = 1, \ldots, J,$$

where $S^*_b(t|\mathbf{X}) = \prod_{u\le t}\left[1 - \sum_{j=1}^J N^*_{j,b}(du|\mathbf{X})/Y^*_b(u|\mathbf{X})\right]$ is $\mathbf{X}$'s bootstrapped Kaplan-Meier estimate of event-free survival.

Averaging $F^*_{j,b}(t|\mathbf{X})$ over trees yields the in-bag ensemble estimate for the event-specific CIF, $\bar{F}^*_j(t|\mathbf{X})$. An OOB estimator is constructed using OOB data. Let $O_i = O(\mathbf{X}_i)$ record trees where case i is OOB. The OOB ensemble estimate is

$$\bar{F}^{**}_j(t|\mathbf{X}_i) = \frac{1}{|O_i|}\sum_{b\in O_i} F^*_{j,b}(t|\mathbf{X}_i).$$

To illustrate, we use the follicular cell lymphoma data from [28]. The subset of 541 patients includes all patients identified as having follicular type lymphoma. Patients were treated with radiation alone or with radiation and chemotherapy. The two types of events are relapse and death. Figure 6 displays the averaged OOB ensemble CIF for the two events from a RSF competing risk analysis using the composite Gray splitting rule. Figure 7 displays VIMP where prediction error was measured using the truncated Harrell's C-index [20]. Here VIMP was calculated using the generalized log-rank splitting rule, in which separate RSF analyses were run for each event type. This type of analysis is most appropriate where the goal is to identify risk factors specific to an event. We see that age of the individual is highly predictive of death but not the competing risk of relapse.

9. Related Articles

stat06520
stat08010
stat07514.pub2
stat06004.pub2
stat03948
stat06529
stat07516
stat07466
stat05255

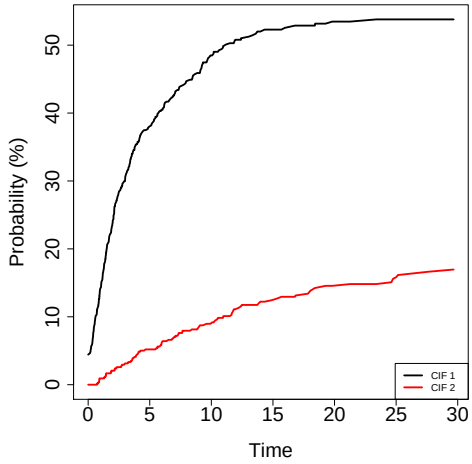

Figure 6. Averaged OOB ensemble cumulative incidence function (CIF) from RSF competing risk analysis of follicular cell lymphoma data using the composite Gray split-statistic. Black and red lines represent event type relapse and death respectively (1=relapse, 2=death).

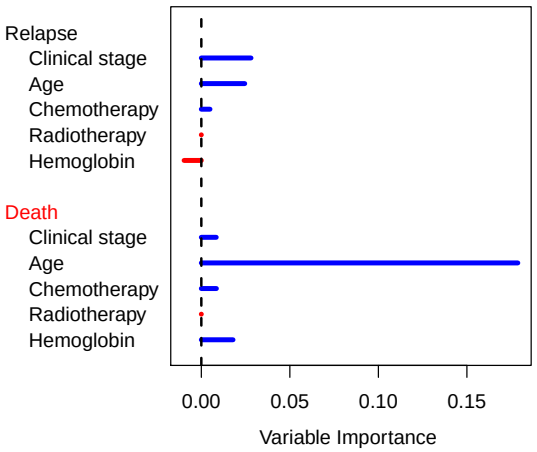

Figure 7. VIMP from two separate RSF competing risk analyses using generalized log-rank splitting.

References

[1] Aalen, OO & Johansen, S (1978), 'An empirical transition matrix for non-homogeneous Markov chains based on censored observations,' *Scandinavian Journal of Statistics*, pp. 141–150.

[2] Breiman, L (1996), 'Bagging predictors,' *Machine Learning*, **24**(2), pp. 123–140.

[3] Breiman, L (2001), 'Random forests,' *Machine Learning*, **45**, pp. 5–32.

[4] Breiman, L (2002), 'Manual on setting up, using, and understanding random forests v3. 1,' *Statistics Department University of California Berkeley, CA, USA*, **1**.

[5] Breiman, L, Friedman, J, Stone, CJ & Olshen, RA (1984), *Classification and Regression Trees*, CRC press.

[6] Ciampi, A, Hogg, SA, McKinney, S & Thiffault, J (1988), 'RECPAM: a computer program for recursive partition and amalgamation for censored survival data,' *Comp. Methods Programs Biomed.*, **26**(3), pp. 239–256.

[7] Cox, DR (1972), 'Models and life-tables regression,' *J. Royal Stat. Soc. Ser. B*, **34**, pp. 187–220.

[8] Davis, RB & Anderson, JR (1989), 'Exponential survival trees,' *Statistics in Medicine*, **8**(8), pp. 947–961.

[9] Friedman, JH (2001), 'Greedy function approximation: a gradient boosting machine,' *Annals of Statistics*, pp. 1189–1232.

[10] Gerds, TA & Schumacher, M (2006), 'Consistent estimation of the expected brier score in general survival models with right-censored event times,' *Biometrical Journal*, **48**(6), pp. 1029–1040.

[11] Geurts, P, Ernst, D & Wehenkel, L (2006), 'Extremely randomized trees,' *Machine Learning*, **63**(1), pp. 3–42.

[12] Gordon, L & Olshen, RA (1985), 'Tree-structured survival analysis.' *Cancer Treatment Reports*, **69**(10), pp. 1065–1069.

[13] Graf, E, Schmoor, C, Sauerbrei, W & Schumacher, M (1999), 'Assessment and comparison of prognostic classification schemes for survival data,' *Statistics in Medicine*, **18**(17-18), pp. 2529–2545.

[14] Gray, RJ (1988), 'A class of k-sample tests for comparing the cumulative incidence of a competing risk,' *The Annals of Statistics*, pp. 1141–1154.

[15] Harrell Jr, FE, Califf, RM, Pryor, DB, Lee, KL, Rosati, RA et al. (1982), 'Evaluating the yield of medical tests,' *JAMA*, **247**(18), pp. 2543–2546.

[16] Hothorn, T, Bühlmann, P, Dudoit, S, Molinaro, A & Van Der Laan, MJ (2005), 'Survival ensembles,' *Biostatistics*, **7**(3), pp. 355–373.

[17] Hothorn, T, Hornik, K & Zeileis, A (2006), 'Unbiased recursive partitioning: A conditional inference framework,' *Journal of Computational and Graphical statistics*, **15**(3), pp. 651–674.

[18] Hsich, E, Gorodeski, EZ, Blackstone, EH, Ishwaran, H & Lauer, MS (2011), 'Identifying important risk factors for survival in patient with systolic heart failure using random survival forests,' *Circulation: Cardiovascular Quality and Outcomes*, **4**(1), pp. 39–45.

[19] Ishwaran, H (2015), 'The effect of splitting on random forests,' *Machine Learning*, **99**(1), pp. 75–118.

[20] Ishwaran, H, Gerds, TA, Kogalur, UB, Moore, RD, Gange, SJ & Lau, BM (2014), 'Random survival forests for competing risks,' *Biostatistics*, **15**(4), pp. 757–773.

[21] Ishwaran, H & Kogalur, UB (2017), 'Random Forests for Survival, Regression, and Classification (RF-SRC),' https://cran.r-project.org/web/packages/randomForestSRC, R package version 2.5.0.

[22] Ishwaran, H, Kogalur, UB, Blackstone, EH & Lauer, MS (2008), 'Random survival forests,' *The Annals of Applied Statistics*, **2**(3), pp. 841–860.

[23] Ishwaran, H, Kogalur, UB, Gorodeski, EZ, Minn, AJ & Lauer, MS (2010), 'High-dimensional variable selection for survival data,' *Journal of the American Statistical Association*, **105**(489), pp. 205–217.

[24] Ishwaran, H & Lu, M (2018), 'Standard errors and confidence intervals for variable importance in random forest regression, classification, and survival,' *Statistics in Medicine*.

[25] LeBlanc, M & Crowley, J (1992), 'Relative risk trees for censored survival data,' *Biometrics*, pp. 411–425.

[26] LeBlanc, M & Crowley, J (1993), 'Survival trees by goodness of split,' *Journal of the American Statistical Association*, **88**(422), pp. 457–467.

[27] Loh, WY & Shih, YS (1997), 'Split selection methods for classification trees,' *Statistica Sinica*, pp. 815–840.

[28] Pintilie, M (2006), *Competing risks: a practical perspective*, vol. 58, John Wiley & Sons.

[29] Segal, MR (1988), 'Regression trees for censored data,' *Biometrics*, pp. 35–47.

[30] Segal, MR (1995), 'Extending the elements of tree-structured regression,' *Statistical Methods in Medical Research*, **4**(3), pp. 219–236.

[31] Steingrimsson, JA, Diao, L, Molinaro, AM & Strawderman, RL (2016), 'Doubly robust survival trees,' *Statistics in Medicine*, **35**(20), pp. 3595–3612.

[32] Zhu, R & Kosorok, MR (2012), 'Recursively imputed survival trees,' *Journal of the American Statistical Association*, **107**(497), pp. 331–340.