ggRandomForests: Visually Exploring Random Forests in R

John Ehrlinger and **Jeevanantham Rajeswaran**
Cleveland Clinic

Hemant Ishwaran
University of Miami

Liang Li
MD Anderson Cancer Center

Udaya B. Kogalur and **Eugene H. Blackstone**
Cleveland Clinic

Abstract

Random Forests (Breiman 2001) (RF) are a fully non-parametric statistical method requiring no distributional assumptions on covariate relation to the response. RF are a robust, nonlinear technique that optimizes predictive accuracy by fitting an ensemble of trees to stabilize model estimates. Random Forests for survival (Ishwaran and Kogalur 2007; Ishwaran, Kogalur, Blackstone, and Lauer 2008) (RF-S) are an extension of Breiman's RF techniques to survival settings, allowing efficient non-parametric analysis of time to event data. The **randomForestSRC** package (Ishwaran and Kogalur 2014) is a unified treatment of Breiman's random forests for survival, regression and classification problems.

Predictive accuracy make RF an attractive alternative to parametric models, though complexity and interpretability of the forest hinder wider application of the method. We introduce the **ggRandomForests** package, tools for creating and plotting data structures to visually understand random forest models grown in R with the **randomForestSRC** package. The **ggRandomForests** package is structured to extract intermediate data objects from **randomForestSRC** objects and generate figures using the **ggplot2** (Wickham 2009) graphics package.

Using both classification (RF-C) and survival (RF-S) examples from our research at the Cleveland Clinic, we will demonstrate the **randomForestSRC** package. We use Variable Importance measure (VIMP) (Breiman 2001) as well as Minimal Depth (Ishwaran, Kogalur, Gorodeski, Minn, and Lauer 2010), a property derived from the construction of each tree within the forest, to assess the impact of variables on forest prediction. We will also demonstrate the use of variable dependence plots (Friedman 2000) to aid interpretation RF results in different response settings.

Keywords: random forest, survival, classification, regression, VIMP, minimal depth.

1. About this document

This document is an introduction to the R package

2. Introduction

2.1. Random Forests

Random Forests (Breiman 2001) (RF) are a robust statistical method that utilizes all variables in predicting the specified outcome. It does not require prior knowledge of the relation of variables (linearity or non-linearity) to the response, or of interactions between variables. RF does not eliminate any risk factor, but rather chooses the most important variables by assessing variable impact on the predictive ability of the forest of trees. Each tree is created with a random sub-group of patients. Each tree is designed to be independent of the others within the forest. This is achieved within the tree growing process by evaluating a split variable candidates from a random subset of covariates.

A Random Forest is built up by bagging (Breiman 1996a) a collection of classification and regression trees (Breiman, Friedman, Olshen, and Stone 1984) (CART). The method uses a set of B bootstrap (Efron and Tibshirani 1994) samples, growing a single tree on each sample. The strength of RF is in ensuring that each tree is independent by subsampling a set of $m \leq p$ input variables for evaluation at each node split of the tree growing process. This independence property ensures that RF minimizes the variance of the aggregated tree estimates. Each node is split into two groups by maximizing the separation of observations according to a measure of the response variable until reaching a stopping criteria of terminal node purity or member size. Each observation is uniquely sorted into only one terminal node per tree. The Random Forest estimate for each observation is obtained by simple averaging (regression) or vote aggregating (classification) the terminal node results across the collection of trees.

Random Forests also have a built in estimate of prediction error. Each bootstrap sample selects approximately 63.2% of data set ibservations on average. The remaining 36.8% observations are left Out-of-Bag (Breiman 1996b) (OOB) which can be used as a hold out test set for each tree. The OOB error rate is calculated for each observation by predicting a response for each observation over the set of trees NOT trained on that particular observation. The OOB prediction error estimates are nearly identical to k–fold cross validation estimates. This RF feature allows us to obtain both model fit and validation in one pass of the algorithm.

RF utilizes all variables in predicting the specified outcome, effectively weighting the most important covariates by assessing their impact on separating dissimilar groups of observations.

2.2. Random Forests for Survival

Random Forests for survival (Ishwaran 2007; Ishwaran *et al.* 2008) (RF-S) are an extension of Random Forests (Breiman 2001) to survival settings, creating a forest of survival trees. Similar to RF, RF-S is a collection of B survival trees. At each node of the tree, we randomly selected a subset of $m \leq p$ candidate variables. The node is then split into two groups by constructing Kaplan–Meier survival curves and choosing the variable that maximizing the log-rank statistic. We split categorical variables according to their natural categories and continuous variables by comparing 10 randomly selected cut points. For each subsequent node of the tree, we repeated the same process: random selection of candidate variables, splitting of each variable with construction of survival plots and calculation of log-rank statistic, and selection of the best splitting variable. The process continued down each branch of the tree

until we reached a unique subset of observations that contain no fewer than 3 deaths within a terminal node. This approach produces extensively grown trees where each terminal node includes a group of observations having similar characteristics and survival outcomes.

We use the randomForestSRC package (**?**) (RF-SRC) for all randomForest methods. This package combines all forest types (survival, regression and classification) in one package with consistent function calls and interfaces between all three types.

3. Example Datasets

3.1. Classification: IRIS

3.2. Regression: Air Quality

3.3. Regression: Cars

3.4. Survival: Veteran

3.5. Survival: PBC

4. Growing a Random Forest

4.1. Forest Imputation

The randomForests package (Liaw and Wiener 2002) include a forest imputation method within the randomForest package.

We impute missing data (both x and y-variables) using a modification of the missing data algorithm of Ishwaran *et al.* (2008). Prior to splitting a node, missing data for a variable is imputed by randomly drawing values from non-missing in-bag data. The purpose of the imputed data is to make it possible to assign cases to daughter nodes in the event the node is split on a variable with missing data. Imputed data is however not used to calculate the split-statistic which uses non-missing data only. Following a node split, imputed data are reset to missing and the process is repeated until terminal nodes are reached. Missing data is then imputed using OOB non-missing terminal node data. For integer valued variables and censoring indicators, imputation uses a maximal class rule, whereas continuous variables and survival time use a mean rule.

The proximity matrix from the randomForest is used to update the imputation of the NAs. For continuous predictors, the imputed value is the weighted average of the non-missing obervations, where the weights are the proximities. For categorical predictors, the imputed value is the category with the largest average proximity. This process is iterated iter times.

Regardless of what method is used, records in which all outcome and x-variable information are missing are removed from the forest analysis. Variables having all missing values are also

removed.

5. Identification of Predictive Variables

Unlike the linear model settings, RF does not explicity specify the functional form of the covariate response relation. We use two seperate methods in order to determine how the RF has built the response prediction, the standard method of Variable Importance (VIMP) as well as the forest minimal depth.

5.1. Variable Importance

Variable importance (VIMP) was originally defined in CART using a measure involving surrogate variables (see Chapter 5 of Breiman *et al.* (1984)). Definitions in terms of mean overall improvement in node impurity for a tree have also been proposed. In regression trees, node impurity is measured by mean squared error, whereas in classification problems, the Gini index is used (**?**). The most popular VIMP method to date, however, adopts a prediction error approach involving "noising-up" a variable. In random forests, for example, VIMP for a variable x_v is the difference between prediction error when x_v is noised up by permuting its value randomly, compared to prediction error under the original predictor (Breiman 2001; Liaw and Wiener 2002; Ishwaran 2007; Ishwaran *et al.* 2008).

Given that VIMP is the absolute difference between prediction errors before and after permutation, a large VIMP value indicates that misspecification of that variable detracts from the predictive accuracy of the forest. VIMP close to zero indicates the variable has no impact on predictive accuracy, and negative values indicate the predictive accuracy improves when the variable is mispecified. In the later case, we assume noise is more informative than the variable. As such, we ignore variables with negative and zero values of VIMP, relying on large positive values to indicate that the predictive power of the forest is dependent on those variables.

5.2. Minimal Depth

In VIMP, prognostic risk factors are determined by inspection of the forest, ranking the most important variables according to impact on predictive ability of the forest. An alternative method recognizes that most important variables for prediction are those that most frequently split nodes nearest to the trunks of the trees (ie, at the root node). Node levels are numbered based on their relative distance to the trunk of the tree (ie. 0, 1, 2). A measure of important risk factors is determined by averaging the depth of first split for each variable over all trees within the forest. Lower values of this measure indicate those variables that split larger groups of patients. This method has been shown to successfully identify the strongest predictors, with no loss of overall model accuracy because of excessive parsimony.

The maximal subtree for a variable x is the largest subtree whose root node splits on x. Thus, all parent nodes of x's maximal subtree have nodes that split on variables other than x. The largest maximal subtree possible is the root node. In general, however, there can be more than one maximal subtree for a variable. A maximal subtree may also not exist if there are no splits on the variable.

The minimal depth of a maximal subtree (the first order depth) measures predictiveness of a

variable x. It equals the shortest distance (the depth) from the root node to the parent node of the maximal subtree (zero is the smallest value possible). The smaller the minimal depth, the more impact x has on prediction. The mean of the minimal depth distribution is used as the threshold value for deciding whether a variable's minimal depth value is small enough for the variable to be classified as strong.

5.3. Variable Importance and Minimal Depth

6. Variable Dependence

Once we have an idea of which variables contribute to the predictive accuracy of the forest, it is useful to get some idea of form of this contribution. We use graphical methods to show the predicted response given dependence on covariates. We can plot the marginal effect of an covariate on the class probability (classification), response (regression), mortality (survival), or the expected years lost (competing risk) for a RF analysis. We plot the ensemble predicted value on the vertical axis and covariates along the horizontal axis.

6.1. Marginal Dependence

Marginal variable dependence plots the predicted response as a function of the covariate, showing each subject as a point on the plot. For classification and regression, this is straight forward predicting the response. In survival settings, we must account for the additional dimension of time. In this case, we plot the response at a specific time point of interest, for example survival at one year. Each predicted point is dependent on the full combination of all covariates, not only on the covariate displayed in the dependence plot. So interpretation can only be in general terms.

6.2. Partial Dependence

An alternative to marginal variable dependence is to integrate out the effects of varaibles beside the covariate of interest. For these *partial variable dependence* plots, the y-value for a variable X, evaluated at X=x, is

$$\tilde{f}(x) = \frac{1}{n}\sum_{i=1}^{n} \hat{f}(x, x_{i,o}),$$

where $\hat{f}$ is the predicted response and $x_{i,o}$ represents the value for all other covariates other than X for the observation i (**?**). The partial plot can be viewed as a risk adjusted estimate of the response as a function of the covariate X.

6.3. Variable Interactions and Conditional Plots

Using the different variable dependence measures, we can calculate pairwise interactions for any pair of variables.

Using minimal depth, we calculate the maximal subtree using the normalized minimal depth of variable i relative to the root node (normalized wrt the size of the tree) the maximal

subtree interaction measure is the normalized minimal depth of a variable j wrt the maximal subtree for variable i (normalized wrt the size of i's maximal subtree). Smaller diagonal entries indicate predictive variables. Small interaction entries having small diagonal entries are a sign of an interaction between variable i and j (**??**)

A joint-VIMP approach is also available where two variables are paired and their paired VIMP calculated (refered to as 'Paired' importance). The VIMP for each separate variable is also calculated. The sum of these two values is refered to as 'Additive' importance. A large positive or negative difference between 'Paired' and 'Additive' indicates an association worth pursuing if the univariate VIMP for each of the paired-variables is reasonably large (Ishwaran 2007).

By plotting the resulting interaction measures for each variable, we can detect the "most interactive" pairs, and develop conditonal plots . These plots are similar to stratifed results, arranged in a set of panels by the interactive variable of interest.

6.4. Variable Interactions

6.5. Conditional Dependence

7. Conclusions

8. Acknowledgments

References

Breiman L (1996a). "Bagging predictors." *Machine Learning*, **26**, 123–140.

Breiman L (1996b). "Out–Of–Bag Estimation." *Technical report*, Statistics Department, University of California,Berkeley, CA. 94708. URL ftp://ftp.stat.berkeley.edu/pub/users/breiman/OOBestimation.ps.Z.

Breiman L (2001). "Random Forests." *Machine Learning*, **45**(1), 5–32.

Breiman L, Friedman JH, Olshen R, Stone C (1984). *Classification and Regression Trees*. Wadsworth and Brooks, Monterey, CA.

Efron B, Tibshirani R (1994). *An Introduction to the Bootstrap*. Chapman & Hall/CRC. ISBN 0412042312.

Friedman JH (2000). "Greedy Function Approximation: A Gradient Boosting Machine." *Annals of Statistics*, **29**, 1189–1232.

Ishwaran H (2007). "Variable importance in binary regression trees and forests." *Electronic Journal of Statistics*, **1**, 519–537.

Ishwaran H, Kogalur UB (2007). "Random survival forests for R." *R News*, **7**, 25–31.

Ishwaran H, Kogalur UB (2014). "Random Forests for Survival, Regression and Classification (RF-SRC), R package version 1.5.2."

Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS (2008). "Random survival forests." *The Annals of Applied Statistics*, **2**(3), 841–860.

Ishwaran H, Kogalur UB, Gorodeski EZ, Minn AJ, Lauer MS (2010). "High-dimensional variable selection for survival data." *J. Amer. Statist. Assoc.*, **105**, 205–217.

Liaw A, Wiener M (2002). "Classification and Regression by randomForest." *R News*, **2**(3), 18–22.

Wickham H (2009). *ggplot2: elegant graphics for data analysis*. Springer New York. ISBN 978-0-387-98140-6.

Affiliation:

John Ehrlinger
Quantitative Health Sciences
Lerner Research Institute
Cleveland Clinic
9500 Euclid Ave
Cleveland, Ohio 44195
E-mail: john.ehrlinger@gmail.com
URL: http://www.biostat.uzh.ch/aboutus/people/rufibach.html