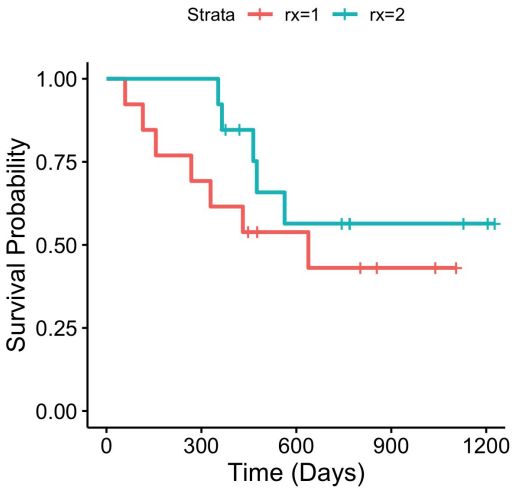
Chapter 22: Survival Trees

Modern Clinical Data Science
Bethany Percha, Instructor

Icahn
School of
Medicine at
Mount
Sinai

Survival trees



Statistical Methods in Medical Research 1995; **4**: 237–261

Trees and splines in survival analysis

Orna Intrator Department of Statistics, Hebrew University, Jerusalem, Israel and
Charles Kooperberg Department of Statistics, University of Washington, Seattle, USA

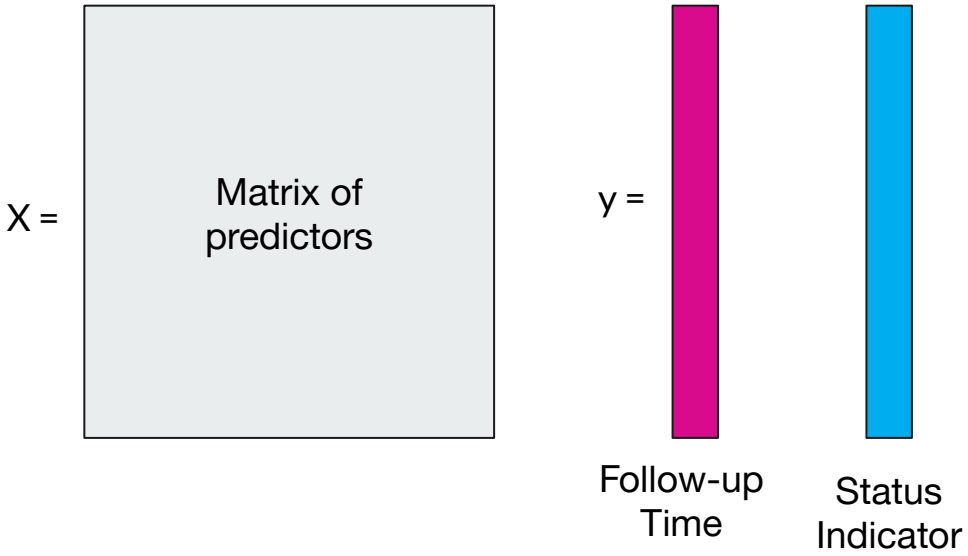
During the past few years several nonparametric alternatives to the Cox proportional hazards model have appeared in the literature. These methods extend techniques that are well known from regression analysis to the analysis of censored survival data. In this paper we discuss methods based on (partition) trees and (polynomial) splines, analyse two datasets using both Survival Trees and HARE, and compare the strengths and weaknesses of the two methods. One of the strengths of HARE is that its model fitting procedure has an implicit check for proportionality of the underlying hazards model. It also provides an explicit model for the conditional hazards function, which makes it very convenient to obtain graphical summaries. On the other hand, the tree-based methods automatically partition a dataset into groups of cases that are similar in survival history. Results obtained by survival trees and HARE are often complementary. Trees and splines in survival analysis should provide the data analyst with two useful tools when analysing survival data.

1 Introduction

In this paper we discuss and compare two groups of nonparametric methodologies for the analysis of censored survival data, methods based on recursive partitioning and those based on polynomial splines. Traditional methods for analysing survival data include exploratory methods such as the Kaplan–Meier estimate, the Nelson–Aalen estimate, and various types of tests that summarize differences between two or more survival distributions, and modelling methods such as the Cox proportional hazards model and the accelerated lifetime model. The nonparametric methods discussed in this paper can give insight into data that the traditional methods fail to provide.

Why trees?

- Can handle large number of predictors / more predictors than samples
- Built in feature selection
- Cox model makes strict proportionality assumption (interpretability nice, but could be oversimplified)



Penalized Cox models are another option when feature selection is primary concern

Regularized Cox Regression

Kenneth Tay Noah Simon Jerome Friedman Trevor Hastie
Rob Tibshirani Balasubramanian Narasimhan

February 17, 2021

Contents

Introduction

This vignette describes how one can use the `glmnet` package to fit regularized Cox models.

The Cox proportional hazards model is commonly used for the study of the relationship beteween predictor variables and survival time. In the usual survival analysis framework, we have data of the form $(y_1, x_1, \delta_1), \ldots, (y_n, x_n, \delta_n)$ where y_i, the observed time, is a time of failure if δ_i is 1 or a right-censored time if δ_i is 0. We also let $t_1 < t_2 < \ldots < t_m$ be the increasing list of unique failure times, and let $j(i)$ denote the index of the observation failing at time t_i.

The Cox model assumes a semi-parametric form for the hazard

$$h_i(t) = h_0(t)e^{x_i^T\beta},$$

where $h_i(t)$ is the hazard for patient i at time t, $h_0(t)$ is a shared baseline hazard, and β is a fixed, length p vector. In the classic setting $n \geq p$, inference is made via the partial likelihood

$$L(\beta) = \prod_{i=1}^{m} \frac{e^{x_{j(i)}^T\beta}}{\sum_{j \in R_i} e^{x_j^T\beta}},$$

Boosted regression models are another option

Package 'gbm'

July 15, 2020

Version 2.1.8

Title Generalized Boosted Regression Models

Depends R (>= 2.9.0)

Imports lattice, parallel, survival

Suggests covr, gridExtra, knitr, pdp, RUnit, splines, tinytest, vip, viridis

Description An implementation of extensions to Freund and Schapire's AdaBoost algorithm and Friedman's gradient boosting machine. Includes regression methods for least squares, absolute loss, t-distribution loss, quantile regression, logistic, multinomial logistic, Poisson, Cox proportional hazards partial likelihood, AdaBoost exponential loss, Huberized hinge loss, and Learning to Rank measures (LambdaMart). Originally developed by Greg Ridgeway.

License GPL (>= 2) | file LICENSE

URL https://github.com/gbm-developers/gbm

BugReports https://github.com/gbm-developers/gbm/issues

Encoding UTF-8

RoxygenNote 7.1.1

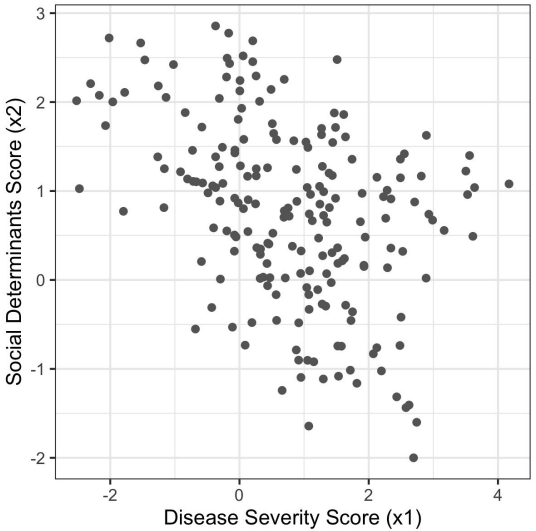
VignetteBuilder knitr

Some disadvantages of survival trees

- Many different splitting criteria
- Different packages all seem to “do their own thing” (e.g. parametric splits vs. log-rank)
- Less interpretable than Cox/Fine-Gray models; culturally less accepted; no standard hypothesis tests, etc.

Part I: Visualizing a Survival Tree

Revisiting Simulation



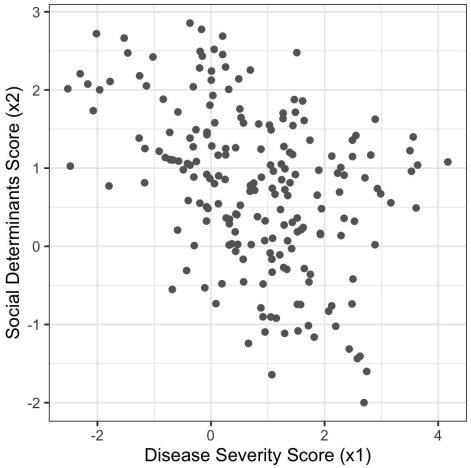
$$\text{discharge} \sim \text{Exponential}(\exp(2 - 1.5x_1 + 0.2x_2))$$

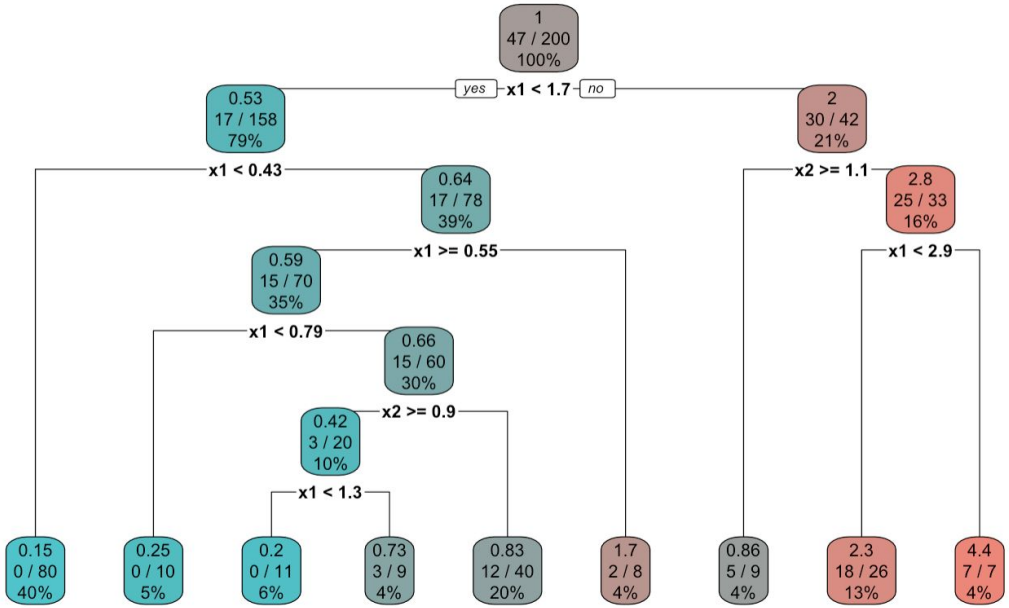
$$\text{death} \sim \text{Exponential}(\exp(-1 + 0.8x_1))$$

$$\text{censoring} \sim \text{Exponential}(1)$$

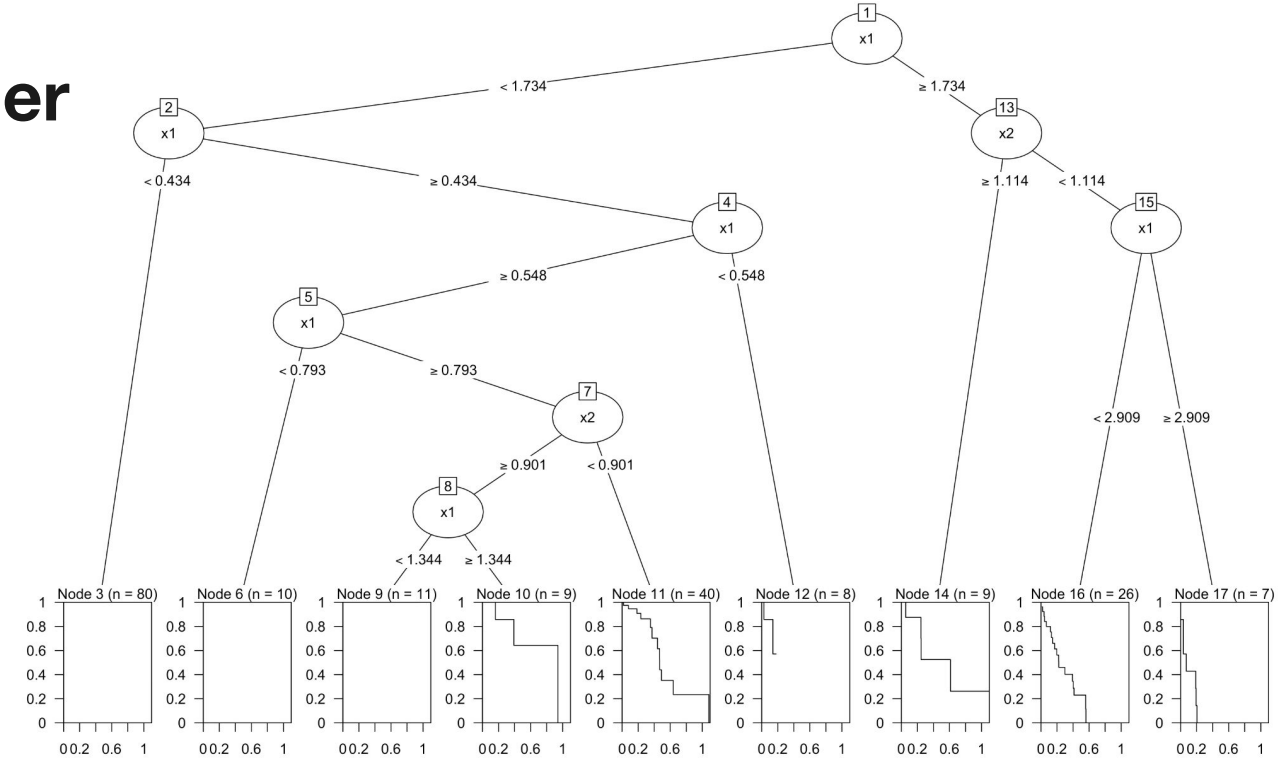
A single survival tree for death

```r
palette <- colorRampPalette(colors=c("#00BFC4", "#F8766D"))
tree_m <- rpart(Surv(ftime12, ctime12 == "death") ~ x1 + x2, data = df)
rpart.plot(tree_m, box.palette=palette(20))
```





Kaplan-Meier view



Comparison of splitting methods

DE GRUYTER Int. J. Biostat. 2015; 11(1): 175–188

Asanao Shimokawa*, Yohei Kawasaki and Etsuo Miyaoka

Comparison of Splitting Methods on Survival Tree

Abstract: We compare splitting methods for constructing survival trees that are used as a model of survival time based on covariates. A number of splitting criteria on the classification and regression tree (CART) have been proposed by various authors, and we compare nine criteria through simulations. Comparative studies have been restricted to criteria that suppose the survival model for each terminal node in the final tree as a non-parametric model. As the main results, the criteria using the exponential log-likelihood loss, log-rank test statistics, the deviance residual under the proportional hazard model, or square error of martingale residual are recommended when it appears that the data have constant hazard with the passage of time. On the other hand, when the data are thought to have decreasing hazard with passage of time, the criterion using the two-sample test statistic, or square error of deviance residual would be optimal. Moreover, when the data are thought to have increasing hazard with the passage of time, the criterion using the exponential log-likelihood loss, or impurity that combines observed times and the proportion of censored observations would be the best. We also present the results of an actual medical research to show the utility of survival trees.

Keywords: survival tree, CART, recursive partitioning

DOI 10.1515/ijb-2014-0029

1 Introduction

In the field of medical research, analysis of time-to-event data is an important subject. The estimation of a survival function using time-to-event data cannot be considered a simple regression problem owing to the presence of censored data. Censored data does not have the correct interval length between the start point (e.g. detection date of illness or surgery date) and end point (e.g. date of death or date of recurrence) as a response variable. In this paper, we deal with right censored cases because they are frequently encountered in medical data analyses. In order to handle a regression problem that includes censored data based on covariates, the Cox proportional hazard (PD) model [1] has been most widely used. In addition to the simpleness of inference, this semiparametric model has an advantage in that it can easily understand the covariate effects. However, this model requires PD assumptions, and certain assumptions about the relationship between covariates and response variables. Moreover, when this model includes many covari-

Part II: Tree Ensembles

A good paper comparing random survival forests to Cox models

QBS
Vol. 36, No. 2, pp. 85-96 (2017)
https://doi.org/10.22283/qbs.2017.36.2.85
pISSN 2288-1344 eISSN 2508-7185 **Invited Article**

Quantitative Bio-Science
Official Journal of the Institute of Natural Science at Keimyung University
Available Online at https://qbs.kmu.ac.kr:442/

A Selective Review on Random Survival Forests for High Dimensional Data

Hong Wang[1], Gang Li[2,*]

[1]School of Mathematics and Statistics, Central South University, Hunan 410083, China
[2]Department of Biostatistics and Biomathematics, School of Public Health, University of California at Los Angeles, CA 90095, USA

(Received October 19, 2017; Revised November 8, 2017; Accepted November 12, 2017)

ABSTRACT

Over the past decades, there has been considerable interest in applying statistical machine learning methods in survival analysis. Ensemble based approaches, especially random survival forests, have been developed in a variety of contexts due to their high precision and non-parametric nature. This article aims to provide a timely review on recent developments and applications of random survival forests for time-to-event data with high dimensional covariates. This selective review begins with an introduction to the random survival forest framework, followed by a survey of recent developments on splitting criteria, variable selection, and other advanced topics of random survival forests for time-to-event data in high dimensional settings. We also discuss potential research directions for future research.

Key words : Censoring, Random survival forest, Survival ensemble, Survival tree, Time-to-event data

1. Introduction

Survival analysis is an active area of research in biostatistics, which focuses on a time-to-event outcome that is typically censored [1-3]. Continuing advancement in data acquisition technology in recent years has made high dimensional or ultra-

event data are those based on the Cox PH model. Current approaches include regularized Cox PH models [5-10], partial least squares [11,12], statistical boosting using Cox-gradient descent or Cox likelihood [13-15]. However, the assumptions underlying these methods such as the proportional hazards assumption are often violated in high-dimensional time-to-

Description of Dataset

The dataset used here includes survival data for 137 patients with 9 censored observations from Veteran's Administration Lung Cancer Trial [33]. In the trial, patients were **randomized to receive either a standard chemotherapy or a test chemotherapy**, and the event of interest here is the survival time in days since the treatment.

A number of covariates which potentially affect survival time are provided: **trt** (type of lung cancer treatment: 1 for standard and 2 for test drug); **celltype** (type of cell involved: squamous, small cell, adeno); **karno** (Karnofsky score); **diagtime** (time between diagnosis and randomization); **age** (age in years); **prior** (any prior therapy, 0 for none and 10 for yes).

Model 1: Cox Proportional Hazards

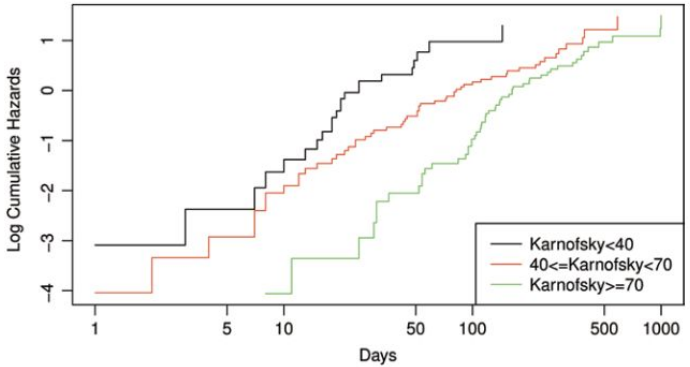


Fig. 1. Plot of the estimated log cumulative hazard functions for different Karnofsky scores.

Table 1. Cox proportional hazard model

Covariate	Coef	Exp (coef)	Se (coef)	z	p
trt	2.95E-01	1.34E+00	2.08E-01	1.42	0.1558
celltypesmallcell	8.62E-01	2.37E+00	2.75E-01	3.13	0.0017
celltypeadeno	1.20E+00	3.31E+00	3.01E-01	3.97	7.00E-05
celltypelarge	4.01E-01	1.49E+00	2.83E-01	1.42	0.1557
karno	−3.28E-02	9.68E-01	5.51E-03	−5.96	2.60E-09
diagtime	8.13E-05	1.00E+00	9.14E-03	0.01	0.9929
age	−8.71E-03	9.91E-01	9.30E-03	−0.94	0.3492
prior	7.16E-03	1.01E+00	2.32E-02	0.31	0.7579

Model 2: Survival Tree

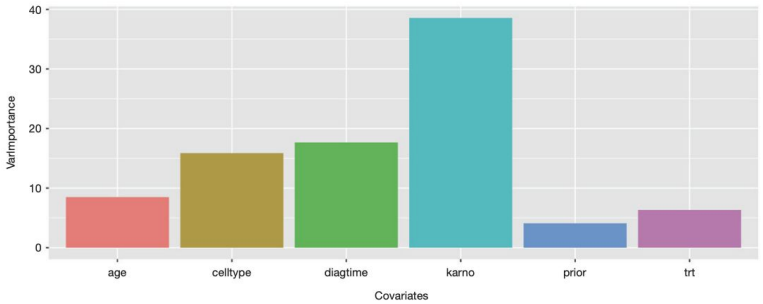
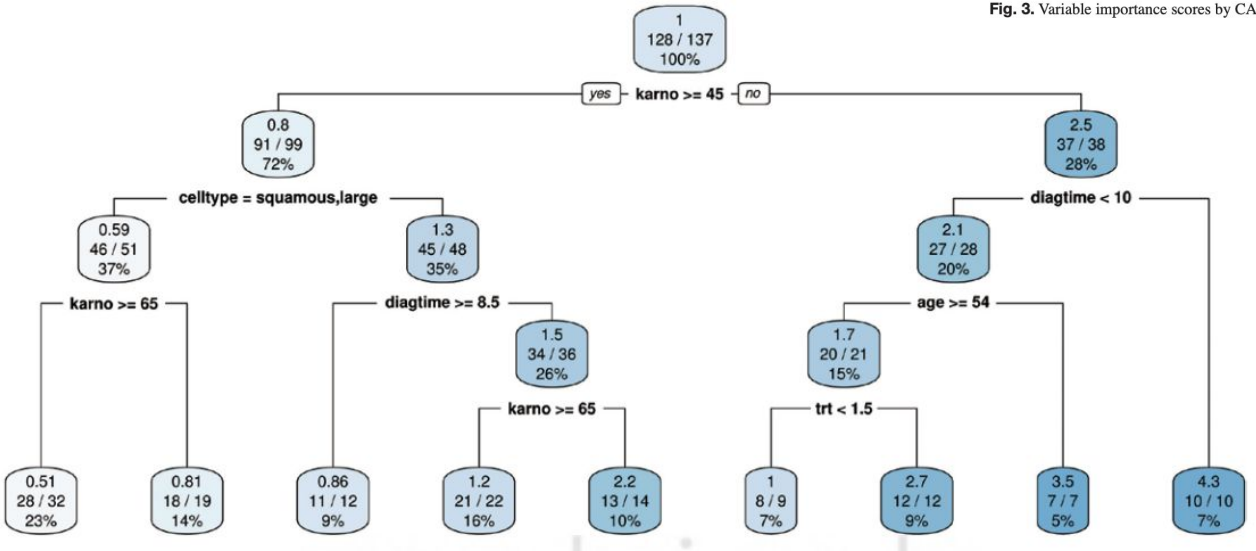


Fig. 3. Variable importance scores by CART.



Model 3: Random Survival Forest

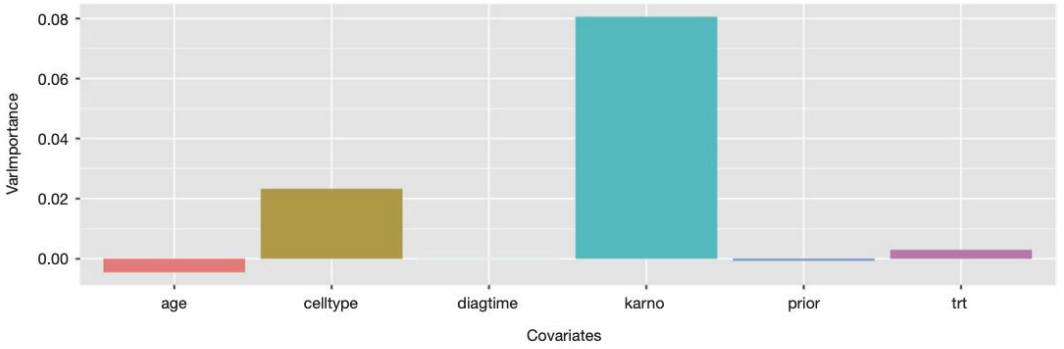


Fig. 4. Variable importance scores by random survival forest.

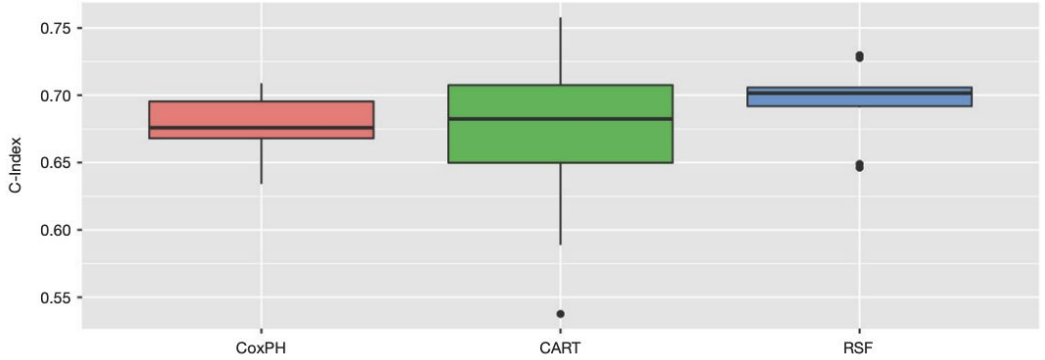
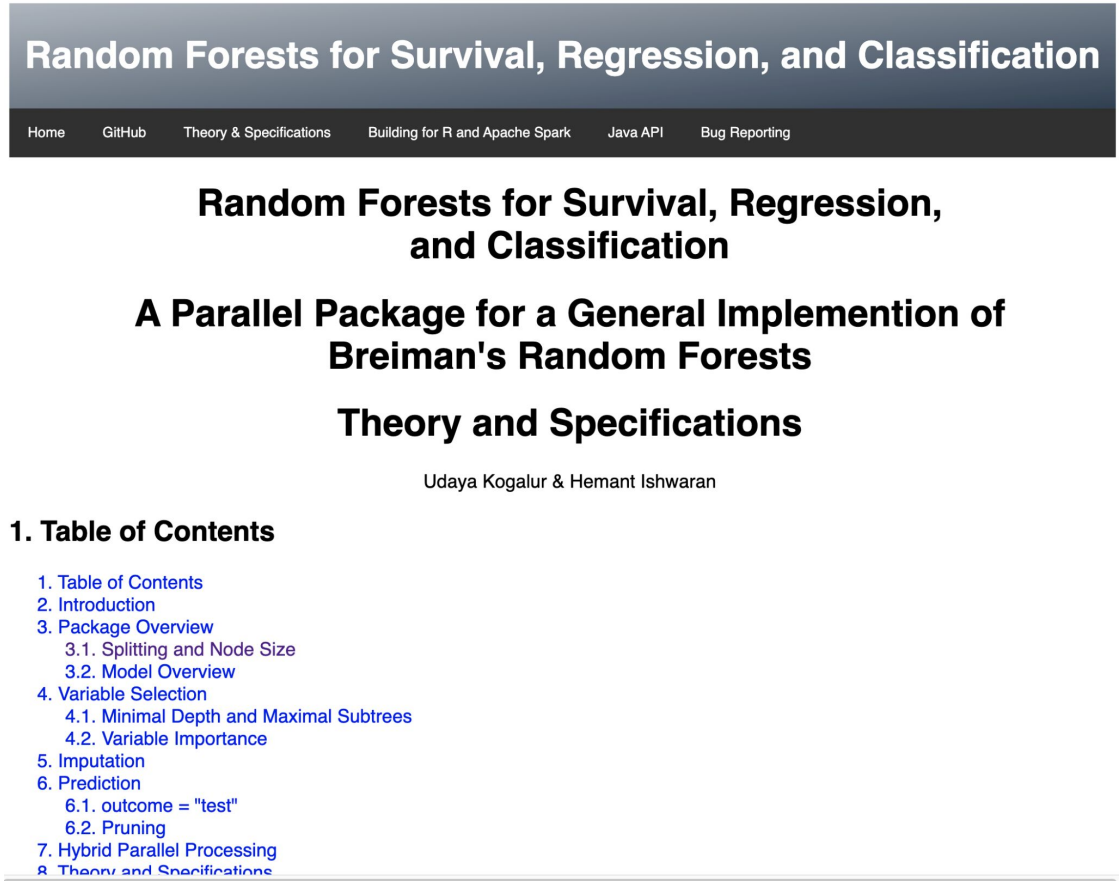


Fig. 5. Performance comparison between Cox, CART and RSF.

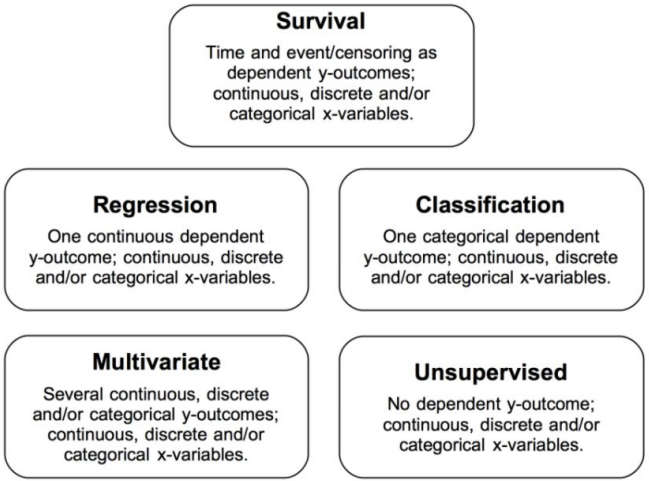
Part III: Trees for Everything

Tree-based approaches can work for many problem classes



Unified model framework

Family	Example Grow Call with Formula Specification
Survival	`rfsrc(Surv(time, status) ~ ., data=veteran)`
Competing Risk	`rfsrc(Surv(time, status) ~ ., data=wihs)`
Regression	`rfsrc(Ozone ~., data=airquality)`
Classification	`rfsrc(Species ~., data=iris)`
Multivariate	`rfsrc(Multivar(mpg, cyl) ~., data=mtcars)`
Unsupervised	`rfsrc(Unsupervised() ~., data=mtcars)`



Five Models Implemented in RF-SRC

Split rules

Family	Split Rules
Survival	*log-rank* (1)
	log-rank score (2)
Competing Risk	*log-rank modified weighted* (12) Gray's K-sample test
	log-rank (11)
Regression	*mean-squared error weighted* (16)
	mean-squared error unweighted (17)
	mean-squared error heavy weighted (18)
Classification	*Gini index weighted* (22)
	Gini index unweighted (23)
	Gini index heavy weighted (24)
Multivariate	Composite mean-squared error
	Composite Gini index
Unsupervised	Composite mean-squared error
	Composite Gini index

Terminal node outputs

Family	Terminal Node Statistics
Survival	Kaplan-Meier (3)
Competing Risk	cause-specific cumulative hazard (13)
	cause-specific incidence (14)
Regression	mean (19)
Classification	class proportions (25)
Multivariate	mean per response (19)
	class proportions per response (25)
Unsupervised	none

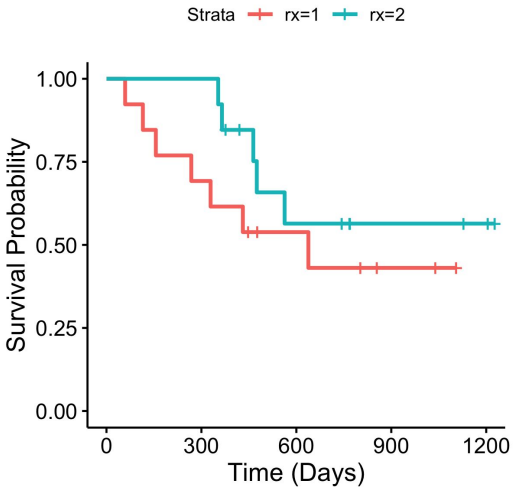
Prediction errors

Family	Prediction Error
Survival	Harrell's C-index using cum-haz (5) and (6)
Competing Risk	Harrell's C-index using cause-specific cum-haz (15) and (6)
Regression	mean-squared error (20)
Classification	conditional and over-all misclassification rate (26) and (27)
Multivariate	none
Unsupervised	none

Summary

- Survival trees and tree ensembles are still active areas of research and development.
- As we start to see larger and larger datasets used in healthcare and more focus on “real-world evidence”, they may start to shine.
- The popularity of different methods is often related to how good/well documented the available software is (e.g., TensorFlow, Torch).

Random survival forests



Package ‘randomForestSRC’

February 10, 2021

Version 2.10.1

Date 2021-02-09

Title Fast Unified Random Forests for Survival, Regression, and Classification (RF-SRC)

Author Hemant Ishwaran <hemant.ishwaran@gmail.com>, Udaya B. Kogalur <ubk@kogalur.com>

Maintainer Udaya B. Kogalur <ubk@kogalur.com>

BugReports https://github.com

Depends R (>= 3.6.0),

Imports parallel, data.tree, Diagran

Suggests survival, pec, prodlim, ml cluster

Description Fast OpenMP parallel ate, unsupervised, survival, co sion. Extreme random forests ing data. Fast random forests rors for variable importance. N ize trees on your Safari or Goo

License GPL (>= 3)

URL http://web.ccs.miami.edu https://github.com/kogal

NeedsCompilation yes

Repository CRAN

Date/Publication 2021-02-10 15:0

Package ‘ranger’

January 10, 2020

Type Package

Title A Fast Implementation of Random Forests

Version 0.12.1

Date 2020-01-10

Author Marvin N. Wright [aut, cre], Stefan Wager [ctb], Philipp Probst [ctb]

Maintainer Marvin N. Wright <cran@wrig.de>

Description A fast implementation of Random Forests, particularly suited for high dimensional data. Ensembles of classification, regression, survival and probability prediction trees are supported. Data from genome-wide association studies can be analyzed efficiently. In addition to data frames, datasets of class 'gwaa.data' (R package 'GenABEL') and 'dgCMatrix' (R package 'Matrix') can be directly analyzed.

License GPL-3

Imports Rcpp (>= 0.11.2), Matrix

LinkingTo Rcpp, RcppEigen

Depends R (>= 3.1)

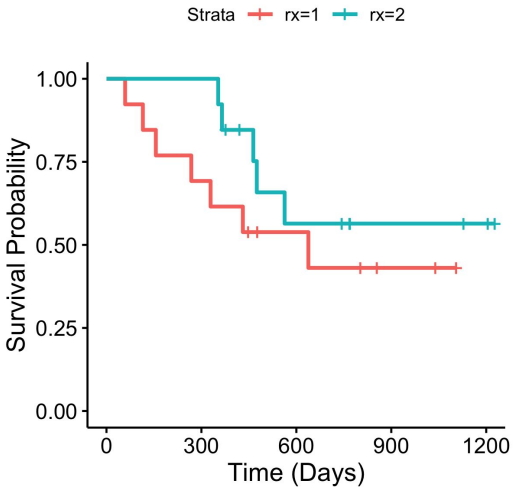
Suggests survival, testthat

Encoding UTF-8

RoxygenNote 7.0.2

URL https://github.com/imbs-hl/ranger

Boosted survival trees



Boosted Trees for Risk Prognosis

Alexis Bellot ALEXIS.BELLOT@ENG.OX.AC.UK
Department of Engineering Science
University of Oxford
Oxford, United Kingdom

Mihaela van der Schaar MIHAELA.VANDERSCHAAR@ENG.OX.AC.UK
Department of Engineering Science
University of Oxford
Oxford, United Kingdom

Editor: Editor's name

Abstract

We present a new approach to ensemble learning for risk prognosis in heterogeneous medical populations. Our aim is to improve overall prognosis by focusing on under-represented patient subgroups with an atypical disease presentation; with current prognostic tools, these subgroups are being consistently mis-estimated. Our method proceeds sequentially by learning nonparametric survival estimators which iteratively learn to improve predictions of previously misdiagnosed patients - a process called *boosting*. This results in fully nonparametric survival estimates, that is, constrained neither by assumptions regarding the baseline hazard nor assumptions regarding the underlying covariate interactions - and thus differentiating our approach from existing boosting methods for survival analysis. In addition, our approach yields a measure of the relative covariate importance that accurately identifies relevant covariates within complex survival dynamics, thereby informing further medical understanding of disease interactions. We study the properties of our approach on a variety of heterogeneous medical datasets, demonstrating significant performance improvements over existing survival and ensemble methods.

1. Introduction