Optimal survival trees ensemble

Naz Gul

A Thesis Submitted for the Degree of

Master of Science by Dissertation

Department of Mathematical Sciences

University of Essex

January 2018

Dedicated to

My affectionate parents (late) and my sons; Sadeed Ali Khan and Shaheer Ali Khan.

Abstract

Selection of accurate and diverse trees based on individual and collective performance in an ensemble has recently been studied for classification and regression problems. Following this notion, the possibility of selecting optimal survival trees is considered in this work. Initially, a large set of survival trees are grown by the method of random survival forest. Using out-of-bag observations for each corresponding survival tree, the trees grown are ranked in ascending order with respect to their prediction errors. A certain number of the top ranked survival trees are selected to be assessed for their collective performance in an ensemble. An ensemble is initiated from the top ranked selected survival tree and further trees are tested one by one by adding them to the ensemble. A survival tree is selected for the final ensemble if it improves the performance by assessing on an independent training data. This ensemble is called optimal survival trees ensemble (OSTE). The proposed method is checked on 17 benchmark datasets and the results are compared with those of random survival forest, conditional inference forest, bagging and Cox proportional hazard model. In addition to improved predictive performance, the proposed method also reduces the number of survival trees in the ensemble as compared to the other tree based methods. Furthermore, the method is implemented in an *R* package called "OSTE".

Acknowledgements

First of all I thank God Almighty for all His blessings!

I would like to take this opportunity to acknowledge many people who have assisted me during my studies and stay at Essex:

I gratefully acknowledge the support and generosity of my supervisor Professor Berthold Lausen, his patience, time, guidance and special support in this research.

I also thankfully acknowledge Abdul Wali Khan University, Mardan, Pakistan for funding this research. To the Founder Vice Chancellor Professor Ihsan Ali for his sincere efforts to develop the University's faculty.

To my Supervisory Board member Dr Spyridon Vrontos who gave me his valuable suggestions and inputs.

To all the staff at the Department of Mathematical Sciences, University of Essex for their kindness and assistance.

To my husband, sisters and brothers who are continually praying for my fortune.

To my friend, Nosheen, for her well wishes, great support and nice company.

To my lovely sons Sadeed Ali Khan & Shaheer Ali Khan for they are the greatest sources of joy in my life.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Survival analysis is one of the subfields of statistics that analyses the measurements of time from a specific point of time until the occurrence of an event of interest. The event could be marriage, death, recovery from a certain disease, censored, i .e. observations with track lost, etc. The main purpose of survival analysis is the estimation and comparison of survival and hazard functions where the former is the survival probability of an individual from some time point to the point of interest and the later is the ratio of the probability density function to the survival function. Survival analysis also assesses how the predictor variables are related to the survival time.

During analysis of survival data, there are some individuals for whom the information about their survival times are incomplete hence, happening of events for such cases remains unobserved. Such type of observations/individuals are called censored. In other words, observations are said to be censored when the event of interest happens before the start or after the end of the study. Here the issue of removing those observations for whom the event happened after the given time of interest is the reduction in sample size. Many attempts

are made in different ways to find the solution of such problems. Methods in survival analysis are divided into parametric, non-parametric and semi-parametric methods. Many techniques are developed in each division to solve the issue.

In today's developed world, the success of machine learning techniques, especially tree based techniques, is remarkable. How to use tree structure introduced by Breiman et al. [1] for survival data was under discussion for several years. Ishwaran et al. [2] grown the survival tree using the approach of maximizing within nodes homogeneity and between nodes heterogeneity. The interpretation of such tree is very easy because of its flexibility and requirement of few assumptions. However, too big or too small tree creates generalization problems for the population of interest. At this point the researchers turns from single learners to group of learners called ensemble methods.

Hothorn et al. [3] bagged survival trees by averaging the predications of survival trees instead of using majority voting, extending the concept of Breiman's [4] bootstrap aggregation. Random survival forests (RSF) are introduced for more refinement of the bagging idea [5]. For further improving random survival forest, the idea of selecting accurate and diverse survival trees from a large initial ensemble of survival trees is introduced in this thesis. This idea was originally introduced by Khan et al. [6, 7] for classification and regression by selecting classification and regression trees from a large pool. The aim here is to achieve promising predictive performance by using an ensemble of a small number of survival trees.

This thesis consists of a total of 5 chapters where the rest of the 4 chapters are arranged as follows.

Chapter 2 gives a short overview of survival analysis. Moreover, tree based techniques

for survival data starting from a general introduction to ensemble methods and the associated literature are also given briefly. A detailed description of some stat-of-the-art methods such as bagging survival trees, random survival forest and conditional inference forest are given in the chapter as a background for this work. A semi-parametric method i.e. Cox proportional hazard model is also discussed in the chapter for comparison purpose.

Chapter 3 discusses the research methodology followed in this thesis. This describes benchmarking for evaluating the methods considered and the datasets used for this purpose along with their sources. The software packages used for benchmark analysis throughout the thesis are also mentioned in this chapter. Furthermore, evaluation metrics used in the thesis are also described in this chapter.

Chapter 4 carries the original work of this thesis. Ensemble of optimal trees for survival data, OSTE, is introduced in this chapter. Benchmarking is done on a number of benchmark problems. The results are compared with those of Cox proportional hazard, bootstrap aggregation of survival trees, random survival forest and conditional inference forest.

The overall conclusion based on the results from the previous chapters is given in Chapter 5. Advantages and shortcomings of the proposed method and some possible directions for future work based on the contribution made in the thesis are also given in this chapter.

The last part of this thesis is the appendix where the pdf manual of the R package `OSTE` is given. The package implements the proposed method in the R language.

Chapter 2

Review of survival analysis with machine learning methods and relevant literature

2.1 Introduction

Due to developed technologies, many disciplines have the capability to collect and monitor observations in various experiments over long-term periods. The primary objective of monitoring is to acquire a better estimate of the time of occurrence of a particular event of interest. This predication/estimation process faces many challenges. One of the main challenges is the presence of censored instances e.g. unobservable event outcomes after certain time or losing track during the monitoring period. In the literature, traditional statistical methods are developed to solve these complications. Survival analysis, an important subfield of statistics, is the most commonly used technique to handle such issues. In addition to these challenges, predictive modelling of survival data has some issues and hence, recently, researchers have developed some machine learning methods and new computational algorithms to tackle such complex problems of survival analysis.

The study of time until some endpoint of interest is called survival analysis [8]. The endpoint is often a combination of different event types. For example, the combination of

different causes of death or the occurance of a disease and the occurance of death without prior disease in the form of one single endpoint [9]. In other words, survival analysis is a statistical procedure used for the analysis of data where time until the occurrence of an event is the outcome variable of interest [10, 11].

The origin of survival analysis may be traced back to 1600's where the main goal was the construction of better life tables and long term extensions of non-parametric estimators of data. It is one of the most commonly used methods for analysing data in many fields of study ranging from environmental health and medicine to astronomy and marketing especially after World War II [12, 13].

Survival analysis has three main purposes [14]:

1. Estimation of survival and hazard functions from survival data.

2. Comparison of hazard and survival functions between groups.

3. Assessment of the relationship between predictor variables and survival time.

There are three main divisions of survival analysis based on the assumptions and how the parameters are used in the model. These are parametric, semi-parametric and non-parametric. The main focus here is on the machine learning algorithms, such as survival trees and ensemble learning like bagging survival trees and random survival forest which have gained much attention in the recent years and come under a separate branch. As a starting argument, one of the semi-parametric methods i.e. Cox proportional hazard model is also discussed in this chapter.

Survival methods estimate important model parameters by assimilating the information from both censored and uncensored data unlike ordinary regression model which can not

handle censored observations. Censored observations are those for which information about their survival time is incomplete.

2.2 Basics of survival analysis

A brief description of the fundamentals of survival analysis and the preliminaries used are given in the following sub-sections.

2.2.1 Censoring

Censoring, a distinguishing factor of survival analysis that differentiate it from other statistical analyses, is a common issue in survival analysis. It arises when there is some information available about an individual's survival time but the exact survival time is unknown [15]. Censoring might be of three types; interval, left and right. If the event of interest occurs after some time of the recorded follow-up time then it is called right censored. A subject is said to be left censored if the event of interest occurs some time before the recorded follow up time. Lack of required information make this type of censoring difficult to handle. For interval censoring, the exact time of failure is unknown but it is known that the event of interest will occur within a given range.

For example a group of those cancer patients from whom the primary tumour is removed during surgery is followed in a survey examining them after every 3 months period. The time to recurrence is investigated in each examination. In the first period of survey, just after 3 months of the surgery, if the investigated/examined patient had a recurrence then his/her survival time is said to be left censored because the recurrence occurred in

less than 3 months period. The survival time of a patient will be right censored if the recurrence is not observed yet but it is unknown that the tumour will re-occur until next investigation period of survey or not. Patients whose examination results at 3 months are disease free but they are lost to follow-up until 6 months investigation or in other words if the information about them is missing in next examination are considered to be interval censored [16]. Right censoring is the most popular among all.

The presence of censoring differentiate survival analysis from other statistical analyses.

2.2.2 Survival time and event of interest

Time starting from a specific point to the occurrence of an event, for example treatment period till recovery of a certain disease is called survival time denoted by T. Time is assumed to be continuous and positive (except in case of discrete-time models examination) therefore, the probability of an event at a single point of a continuous distribution is zero. Survival analysis problems are familiar with survival time of only those instances for which the event has occurred during the study period. Therefore, this can be affected if there are some subjects in the study who may not experience the event or they do not remain in the study. For these instances, only the censored time denoted by C is available. For an individual i we can observe the minimum of T_i and C_i which is represented as:

$$y_i = min(T_i, C_i). \tag{2.1}$$

The censoring can thus be defined by the following indicator function

$$\delta_i = \begin{cases} 0, & \text{if uncensored i.e. } T_i \leq C_i, \\ 1, & \text{if censored i.e. } T_i > C_i. \end{cases} \tag{2.2}$$

As given before, the definition of the event is another important concept in survival analysis. It is an important consideration to specify the start and end dates of an event when designing the study [16].

2.2.3 The cumulative density function and probability density function

To describe the continuous probability distribution of a random variable in survival analysis, for example time the cumulative density function, *cdf* denoted by $F_T(t)$ is usually used. It is defined as

$$F_T(t) = P(T < t) = \int_0^t f_T(t)dt. \tag{2.3}$$

$F_T(t)$ is a non decreasing function of T ranging from 0 to t (the actual value of T that we select). It shows the probability that the event of interest occurs earlier than t.

In terms of survival function it can be represented as

$$F_T(t) = 1 - S(t), \tag{2.4}$$

where $S(t)$ is the survival function (defined later).

The Probability density function, *pdf* denoted by $f_T(t)$ is the slope of *cdf* defined as

$$f_T(t) = \frac{d}{dt}F_T(t). \tag{2.5}$$

2.2.4 Survival function

Survival function describes and models survival data. The word survival or survive has its roots in medicine where death is the event of interest. It can be estimated by the Kaplan Meier method (see Section 2.2.7) representing the probability of an individual surviving from the time of origin to some given time t. It is also known as survivor or reliability function. Suppose T is a continues and positive random variable representing the time until some event of interest then $S(t)$ represent the probability of surviving beyond time t, given as

$$S(t) = P(T \geq t) = 1 - F_T(t),$$

where $F_T(t)$ is the density function of T.

Taking integration on both sides of Equation 2.5,

$$F_T(t) = \int_0^t f(t)dt.$$

Thus

$$S(t) = 1 - \int_0^t f(t)dt. \tag{2.6}$$

The initial value of survival function i.e. S(t) is 1 when $t = 0$ showing that 100% of observed subjects have survived in the beginning or an event of interest has not occurred yet. For $t = \infty$, $S(t) = 0$. The survival curve (i.e. graph of $S(t)$) begins at $S(0) = 1$ and as t increases to infinity it decreases to 0.

2.2.5 Hazard function

Another commonly used function is the hazard function that gives the ratio of the probability density function to the survival function. It is used for determining a mathematical model for survival analysis shown as

$$\begin{aligned} h(t) &= \lim_{\triangle t \to 0} \frac{P(t \leq T < t + \triangle t | T \geq t)}{\triangle t} \\ &= \lim_{\triangle t \to 0} \frac{F(t + \triangle t) - F(t)}{\triangle t \, S(t)}. \end{aligned} \tag{2.7}$$

The numerator of the above equation shows the probability of the occurrence of an event in the interval $(t, t + \triangle t)$ given that no event has occurred before time t [17]. It is clear from the $\lim_{\triangle t \to 0}$ that the probability is within very small intervals and thus the hazard function also got the name of instantaneous risk [18]. Stated differently, the hazard function measures the immediate risk, that is for subject survived till time t an event at time t will happen. It thus seems more spontaneous to be used in survival analysis [16]. If $h(t)$ is known, $S(t)$ can be calculated and vice-versa. The hazard function can also be represented as

$$h(t) = \frac{f_T(t)}{S(t)} = \frac{f_T(t)}{1 - F_T(t)}, \tag{2.8}$$

where $F_T(t)$, distribution function, is the probability that a variate T takes on a value less than or equal to a number t. Thus

$$\begin{aligned} h(t) &= -\frac{\partial}{\partial t} log[1 - F_T(t)], \\ &= -\frac{\partial}{\partial t} log[S(t)]. \end{aligned}$$

The estimation process of $h(t)$ is not simple. Alternatively the cumulative hazard function (*CHF*) [16] is used. It is the integration of hazard function and can be interpreted as the probability of failure at time t given survival until time t, i.e.,

$$CHF(t) = \int_0^t h(t)dt. \tag{2.9}$$

Most of the time for right censored data the survival distribution function of an individual that survives at time t can be calculated using the following relation [19]

$$S(t) = exp[-CHF(t)].$$

Taking log on both sides of the above equation

$$\begin{aligned} logS(t) &= log(exp[-CHF(t)]) \\ &= [-CHF(t)] \\ &= \frac{1}{CHF(t)}, \end{aligned}$$

and thus

$$CHF(t) = \frac{1}{logS(t)} = -log[S(t)]. \tag{2.10}$$

CHF is also used in the calculation of Nelson-Aalen estimator [16].

2.2.6 Test statistics for survival data

There are several statistics used for survival data. Some of the most commonly used are briefly mentioned here. The emphasis will be to describe them in the context of a survival tree.

Log-rank test is a hypothesis test usually used to compare the survival distribution of two samples/groups. It is applicable to the data consisting of progressive/continues censoring and where the early and late events of interest have the same probabilities. It is a non-parametric test that targets on hazard function. It records the time at which the failure event occurs and creates a two-by-two contingency table. This table shows the total number of subjects under study and the total deaths occurred. To yield a χ^2 statistic with 1 degree of freedom, quantities e.g. observed deaths, expected deaths i.e. sum of the expected number of events at the time of each event and variance of the expected number of deaths in each group are summed [16]. It is calculated as follows:

$$\chi^2 = \sum_{l=1}^{g} \frac{(O_l - E_l)^2}{E_l}, \tag{2.11}$$

which is a χ^2 distribution with $g - 1$ degree of freedom. In the above statistic O_l and E_l represents the observed number of events and the total expected number of events in each group l respectively. Moreover, the expected number of events at the time that each event happens is obtained by multiplying the risk of death at that time (number of deaths divided by number alive), with the number alive in each group.

In the context of survival trees based methods the log-rank statistic and its extensions such as a conservation of events splitting rule, a log-rank score rule and a fast approximation

to the log-rank splitting rule, might be used to measure the distance between the two daughter nodes of a tree (see section 2.4.1.1) [2, 20]. The log-rank splitting rule has been used in this work.

Let assume that parent node h consisting of N individuals with their survival times and censoring information denoted by $(T_1, \delta_1), \ldots, (T_N, \delta_N)$ is required to be split in daughter nodes. Let the death times in the node h are $t_1 < t_2 < \cdots < t_N$. The number of deaths and individuals at risk at time t_i in the daughter nodes $j = 1, 2$ are represented by $d_{i,j}$ and $Y_{i,j}$. The test statistic is calculated using the following formula:

$$L(x, c) = \frac{\sum_{i=1}^{N}(d_{i,1} - Y_{i,1}\frac{d_i}{Y_i})}{\sqrt{\sum_{i=1}^{N}\frac{Y_{i,1}}{Y_i}(1 - \frac{Y_{i,1}}{Y_i})(\frac{Y_i - d_i}{Y_i - 1})d_i}},$$

where Y_i and d_i are the total number of individuals at risk and total number of deaths in under study node. Node separation is based on the value of a statistic obtained from the split with the largest value of $|L(x, c)|$. It is clear that the larger the test statistic value the better is the split. The log-rank test encourages the daughter nodes in the tree to be far apart. Leblanc and Crowley [21] also used the log-rank statistic as a splitting rule for nodes. It's calculation also produces the ratio of observed events to the expected events of interest [16]. To test the differences between survival curves for groups, log-rank test may be used [16]. This test allows the comparison of two Kaplan-Meier survival curves. The log-rank test shows that the survival between two groups is significantly different, but it does not provide informations that how different they are [22]. To solve this issue, one method is to show the survival in each group at comparable times. Another method is the comparison of observed and expected numbers of events in each group. It is known

that for two groups in which the survivals are equal, the number of observed events will be proportional to the number of expected events and will be the same in each group. For example if $\frac{O_1}{E_1}$ shows the ratio of observed events to the expected events in the first group under the true null hypothesis then the ratio, also called the hazard ratio

$$R = \frac{O_1/E_1}{O_2/E_2},$$

is an estimate of the relative event rates in two groups. The hazard ratio is also known as failure rate [23].

It is defined as a measure of deaths for the items involved in a study. It play a useful role in the analysis of survival data in that it counts the rate at which the failures occur. The relative hazard in the two groups is based on the complete study period and thus it does not stay the same throughout the whole period. The word "proportional" is used as a keyword associated with this test. This means that the chance of the hazard occurring in one group is proportional in comparison to the other group or the survival curves for two groups must have hazard functions that are proportional over time. In other words, it is an assumption of the Cox proportional hazards model which shows that a linear increase in the predictor will have a uniform multiplicative relationship with the hazard. Moreover, it states that the hazard ratio for two subjects characterized by different sets of covariates is independent of time and depends only on the values of these covariates. In other words at all time points the effect of a given covariate on a hazard level is the same, meaning that the hazard ratio is constant over time [24]. The following equation determines the hazard

rate at any time point

$$h(t) = \frac{f(t)}{1 - F(t)} = \frac{f(t)}{S(t)}, \tag{2.12}$$

where $f(t)$ is the probability that failure will occur in a stated interval while $S(t)$ is the survival function. The hazard rate is always positive.

2.2.7 Kaplan-Meier estimator

Kaplan and Meier [25] developed the Kaplan-Meier (KM) or the product-limit (PL) estimator to estimate the survival function using the actual length of the observed time. It is defined as the probability of surviving in a specified time while observing the time in small intervals. The product limit estimator, combine the informations from all censored and uncensored observations as a series of steps defined by the observed censored and survival times at any point in time. It is the simplest way to compute the successive probabilities of occurrence of an event at a certain point of time. In other words, it is a non-parametric estimator that estimate the survival function from long term data. To get the resultant cumulative survival probability, the survival probabilities from one interval to the next may be multiplied together by assuming the independent occurrence of events [22].

In the underlying analysis three assumptions are used. First it is assumed that the subjects whether censored or still continue to be followed have the same prospects of survival at any time in a study. Second assumption is that the probabilities for the subjects entered in the study either early or late are same. It means that the conditional survival probabilities of the observations, with time limits greater or less than t, are same. Happening of an event at specified time comes under third assumption [25].

Total survival probability till the specified time interval is the multiplication of all probabilities of survivals at all intervals of time preceding that time interval.

Let $T_1 < T_2 < \cdots < T_k$ be a set of distinct event times observed for n ($k \leq n$) instances. The number of observed events is $o_j \geq 1$ for a specific event time T_j, where ($j = 1, 2, \ldots, K$). r_j are censored instances with a censored time equal or greater than T_j. Due to censoring, r_j can not be considered as the difference between r_{j-1} and d_{j-1}. The correct way is to subtract the number of instances censored during the time period between T_{j-1} and T_j i.e. c_{j-1} and o_{j-1} from r_j as:

$$r_j = r_{j-1} - o_{j-1} - c_{j-1}.$$

The conditional probability of those subjects surviving beyond time T_j can be defined as:

$$p(T_j) = \frac{r_j - o_j}{r_j},$$

and thus the KM or PL estimate of $S(t) = P(T \geq t)$ can be written as:

$$\hat{S}(t) = \prod_{j:T_j<t} p(T_j) = \prod_{\text{j: decreases to } 0_j<t} (1 - \frac{o_j}{r_j}). \tag{2.13}$$

Kaplan-Meier is one of the widely used estimators. To summarize the given data in the best way, KM survival curve, plot of survival probabilities estimated by KM estimator against time t is used [22]. Similarly, KM estimator has a small sampling variance and fewer discontinuities and hence there is no need to know about the observation limits for the dead/lost items [25]. However, if the data consist of a large number of subjects or they are grouped into some interval periods according to the time, or if a very large population is

under discussion then the results of life table are more optimal [26]. For sampled population the PL estimates may fall out side the range of true values unlike reduce sample estimates which are weighted averages [25]. Similarly, to estimate the cumulative hazard function for censored data Nelson-Aalen estimator [19] is used.

Nelson-Aalen is a non-parametric estimator suggested by Nelson for the graphical representation of how the parametric models fit to the data. Later on, Aalen studied the properties of this estimator taking large and small samples of a given data. For this purpose martingale methods are used. This work prepared the ground for the extension of the estimator for survival data [19]. Let $(t_1, t_2, \dots)$ be the times observed for the event of interest then the Nelson-Aalen estimator is given as:

$$\hat{A}(t) = \sum_{t_j \leq t} \frac{o_j}{r_j}, \tag{2.14}$$

where o_j represents observed events of interest and r_j the individuals at risk.

As survival data are incomplete due to the presence of censoring hence, the variance is calculated as [19]

$$\hat{\sigma}^2(t) = \sum_{t_j \leq t} \frac{(r_j - o_j)o_j}{(r_j - 1)r_j^2}. \tag{2.15}$$

Here r_j is the number of subjects/individuals entered in the study earlier than time t_j and are still there instead of individuals at risk.

There are numerous approaches used for analyzing survival data, some state-of-the-art methods are briefly described in the following sections.

2.3 Cox proportional hazard model

Under the semi-parametric category, the most commonly used approach for survival data are Cox models [27]. Cox [27] mainly focused on hazard function and introduced a large family of models. A broadly applicable and the most widely used representative member of this family is Cox proportional hazard (Coxph) model. The core idea of this model is to define a hazard level as a dependent variable explained by the baseline hazard a time-related component and the covariates-related component. The model is based on several assumptions one of which is the proportional hazard assumption. The Cox proportional hazard model is a method used to investigate that how the variables effect the time that a specified event happens. These effects on survival are assumed to be constant over time and additive in one scale, showing the semi-parametric nature of the model. It leaves the baseline hazard undefined while the partial likelihood is maximized to address the problem of censoring [28]. Researchers need to discuss and interpret only one assumption that the hazards are proportional over time. No assumption about the parametric distribution of the survival times is required, which makes the method more robust. The Cox model gives the relative risk type measure of association which explains the risk of the events for certain categories of covariates. Moreover, the use of partial likelihood function gives flexibility to the model to introduce explanatory variables dependent on time and handle censored survival time [27]. In this model, the measure of effect is hazard rate, which is the risk of

failure. The model can be written as

$$h(t, \mathbf{x}_i) = h_0(t) \exp(\beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_2 +, \ldots, +\beta_i \mathbf{x}_i), \quad (2.16)$$

$$= h_0(t) \exp(\beta' \mathbf{x}_i), \quad (2.17)$$

where $h_0(t)$ represents the baseline hazard, $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_i)$ is a vector of covariates and $(\beta_1, \beta_2, \ldots, \beta_i)$ is a coefficient vector that measures the size of effect made by the covariates.

The baseline hazard is independent of $\mathbf{x}$. It depends on t only. The exponential involves $\mathbf{x}$ where $\mathbf{x}$ is independent of time. $h(t, \mathbf{x}_i)$ is the product of $h_0(t)$ and the exponential function [18]. The Cox model is a flexible model which allows to leave the baseline hazard function completely unspecified. The above equation can be written in a linear form by taking natural logarithm of both sides as

$$\log h(t, \mathbf{x}_i) = \log(h_0(t) \exp(\beta' \mathbf{x}_i)), \quad (2.18)$$

$$= \log h_0(t) + \beta' \mathbf{x}_i, \quad (2.19)$$

where β' is a regression coefficient vector, which can also be expressed as "relative risk" an attractive feature of the under discussion model [18]. The parameters of the Cox model can be interpreted by coding the single covariate $\mathbf{x}$ as 0 or 1. From Equation 2.16 we obtain

$$h(t, \mathbf{x} = 0) = h_0(t), \quad (2.20)$$

$$h(t, \mathbf{x} = 1) = \exp(\beta) h_0(t), \quad (2.21)$$

where $\exp(\beta)$ is referred to as relative risk. The positive β shows high hazard of group with $\mathbf{x} = 1$ as compared to a group with $\mathbf{x} = 0$, while negative β shows higher hazard for $\mathbf{x} = 0$ among the groups. The hazard will change by a factor of $\exp(\beta)$ for continuous $\mathbf{x}$ if a 1 unit change occurs in $\mathbf{x}$ [18].

Cox proportional hazard model interprets the effect of covariate variables in a simple manner and can be used easily for inference [5]. The most interesting attribute is that for estimating regression coefficient only failure time ranking is needed. However, Cox proportional hazard models forces a particular link (i.e. the effect of one-unit increase in a covariate is a constant multiple) between the responses and covariates.

In some cases the model may not fulfil the proportionality assumption and hence, this will lead to the misspecification of the model [29].

As mentioned above, the advantage of the semi-parametric methods, for example Cox proportional hazard model is that they requires no information about the distribution of survival time. This, however, on the other hand makes model interpretation difficult [30].

Non-parametric methods show higher efficiency when no suitable theoretical distribution is known. Survival function and its graphical presentation is the most widely used non-parametric function among all in survival analysis. Different approaches such as Kaplan-Meier, Nelson-Aalen and Life-Tables are used by researches to estimate survival function. Kaplan-Meier and Nelson-Aalen, the most popular and commonly used estimators in non-parametric technique are discussed in the previous section.

2.4 Machine learning methods

Due to their ability of modeling the non-linear relationships and quality of prediction, machine learning methods achieved a remarkable success in the near past [31]. Dealing with the issues of censored information and time estimation of the model in survival analysis are the main challenges for machine learning methods [32].

A comprehensive review of most commonly used machine learning methods in survival analysis is given in the following sections.

2.4.1 The basics of tree methods

Tree-based methods have several advantages over common regression methods, used for the process of estimating and testing the effects of covariates and for predicting the outcomes [33]. One of the popular machine learning methods for classifying given observations on the basis of recursive binary partitioning into small subgroups is called classification and regression tree (CART) introduced by Breiman et al. [1].

CARTs are receiving great attention in all research areas. The main idea behind a tree based model is to partition the data according to some splitting criterion recursively and place similar objects based on the event of interest in one or the same node. For the development of tree methods two approaches are used. Growing a tree to maximize homogeneity of the outcomes within-node is the first approach that aims to identify set/groups of observations with similar outcomes. To maximize the between-node heterogeneity, there is a second approach of growing tree in which the splits that produce the largest difference between the subgroups is chosen. An example of such approaches could be found

in [34], where p-values on all available factors and all possible splits within the factors are computed. Observations are divided into two subgroups based on the factor and the corresponding split point within the factor with the smallest p-value. This process is repeated recursively until there is no allowable split or the p-value becomes greater than some specified threshold. This method is also adjusted for variables measured on different scales [34]. However, due to characteristics of censored data, standard CART algorithms are not directly interchangeable to the context of survival analysis. Ciampi et al. [35] first attempted to use tree structure for survival data. However, the creation of survival trees was first discussed by Gordon and Olshen [20].

2.4.1.1 Survival tree

Survival trees were augmented from the mid–1980s up to the mid–1990s, where the main goal was to extend existing tree methods to the survival data with censoring [5].

Trees for survival data use both the approaches of CARTs as mentioned in the previous section [33]. The only difference is in the choice of splitting criterion. CART technique can be used for both univariate and multivariate survival data. Thus the most popular tool of survival analysis is a CART-style decision tree called survival tree.

Let $(y_i, \delta_i, \mathbf{x_i})$ be the available survival data for N subjects where $y_i = min(T_i, C_i)$, δ_i is a censoring indicator that takes values 0 and 1 for uncensored and censored subjects respectively, $\mathbf{x_i}$ is a vector of covariates and subscript $i = 1, 2, 3, \ldots, N$. Survival tree consists of three types of nodes called root/parent node, internal/daughter/child nodes and terminal nodes. A tree is grown by taking all given observations in the root node. Then split the root node into two daughter nodes using predetermined survival criterion (see Section

2.2.6). To achieve a maximum difference between daughter nodes all possible cut-offs of all independent variables are needed to be tried. Split the consequent daughter nodes into left and right daughters again using the same strategy. For each subsequent node the same steps are repeated recursively [2].

To further illustrate, consider the following example. Suppose we want to predict a response from a given data, a binary survival tree needs to be grow by recursively splitting the given data at parent node h on a single predictor. The two inequalities $x \leq c$ and $x > c$ are used as criteria for splitting the parent node h in a child nodes v and w respectively. A value c for a given predictor x that maximize the survival difference between two nodes is needed. For this, first of all we chose x from given predictor variables, determining a split value c and place an individuals in either right or left child/daughter node based on the above inequalities. Then we calculate the survival difference between the two child nodes using a predetermined splitting rule. The process is repeated with another c until we find the value of c which maximizes survival differences. In other words, the best split for node h is the one for which predictor x and split value c maximize the survival difference between the two daughter nodes. To each of the two daughter nodes the same process is applied recursively until the sample size of a node is sufficiently small.

The splitting rule used in growing survival tree must involve survival time and information about censoring. The best split is that which maximize survival difference between daughter nodes. The maximum survival difference not only increases the number of nodes in a tree but also makes nodes homogeneous and populated with similar survival cases [2]. Here in this thesis, the log-rank statistic (see section 2.2.6) is used.

Survival trees are used for the construction of models that predicts the outcomes by us-

ing covariates like regression techniques. However, survival trees are more advantageous than simple regression techniques. Interpretation of survival trees is very easy as they are based on a range of non-hazard measures of risk (i.e. assumptions of Cox proportional hazard model) unlike regression techniques. In general they require fewer assumptions than regression techniques and is free from proportional hazard assumption. There are several extensions of the Cox model that relax the proportional hazard assumption. Examples of such models can be found in [24, 36]. These methods, however, do not come under the scope of this thesis. They provide straightforward rule for the classification of outcomes which is the common objective of survival analysis. Survival trees can easily handle high-level interactions and those covariates whose values change over time, i.e. time dependent covariates and hence their graphical representation in the form of binary trees is very easy. It is one of the non-parametric alternatives to (semi) parametric models. Survival trees are very flexible and can identify many types of interaction with out the need of beforehand specification [5]. Selection of single tree as a final model and deciding about splitting criterion, that when to stop growing the tree is an important aspect. The population of interest will not be generalized properly if the tree overfits the data. Similarly important attributes of the relationship between covariates and the outcome might be missing if the tree is too small.

Two approaches are used to select best final tree. Backward selection method, which selects relevant subtree by pruning some branches of already grown large tree and forward selection method, in which splitting a node further is stopped by using a built–in stopping rule. Although, there are several advancements made by using ensemble of trees instead of using a single tree, a stand alone tree is still of great importance to know about the data

in details [5]. Some ensemble methods are described in the following sections.

2.4.2 Ensemble methods

The word ensemble means a group. Ensemble methods are actually the group of predictive models to achieve cohesive and accurate models. Previous research shows that an ensemble is more accurate than any of the single learner in the ensemble [37]. Ensemble methods are supervised learning algorithms where the main idea is to generate a set of models (base learners) from training data for the solution of the same problem instead of using a single learner as is done in ordinary learning approaches. The main aim of ensemble methods is to bunch up the estimations of many base learners produced by given learning algorithms to upgrade strength or generalization over one learner. The values for the response variable for the new/test data points are then predicted by considering the prediction results from all generated learners.

In the last two decades these methods have enjoyed growing attention in machine learning community [38]. The reasons of using ensembles are numerous, a few of them are listed here.

First and foremost reason to use ensemble is the statistical reason, for example the randomly chosen base model from many that have similar test errors or training might be really poor, however, averaging them may prevent us from making a weak decision. Secondly, computation via ensembles makes it easy to find close to global minimum solution by combining different results obtained from learning algorithm run many times [39]. Thirdly, these methods are used for the solution of representational problems, for instance, the combination of linear models can solve non-linearly separable problems [39].

The main focus of this work will be on survival tree based ensemble methods, it is, therefore, considered important to discuss the more advanced machine learning methods that have been developed to deal with and predict from censored data.

To study survival data with censoring Cox proportional hazard regression model, discussed in earlier section, is widely used. These models on one hand, allows simple interpretations of the covariate effects and thus can be used for inference. However, on the other hand, they force a specific link between the covariates and the response which must be specified by the analyst. Moreover, the statistical properties of inference made after many models have been tried are still largely unknown. Furthermore, transformations or expansion of the design matrix becomes necessary for modeling non-linear effects of variables. Similarly identification of multiple variables interactions is also a great problem [2].

One of the alternative solution of these problems are the ensemble methods e.g., bagging and forests.

2.4.2.1 Bagging for survival data

Instability, small change in learning sample inducing large change in the resulting predictor, in learning algorithms was first examined by Breiman and reviewed different ways for eliminating or reducing these instabilities. Bootstrap aggregation/bagging is one of the oldest and most commonly used ensemble method which typically reduces the variance of the base models that are used and improves classification accuracy by manipulating instability [4].

Well known problems of tree based methods are instability and variable selection bias. In many applications bagging stabilize the predictors. Instability of tree in survival appli-

cation was examined by Dannegger [40].

The general methodology of bagging is drawing bootstrap samples from the original data and for each of these samples growing maximal survival tree without pruning, a technique that reduces the size of over-fitted decision trees by removing some of its sections. Then calculating prediction for each individual survival tree and finally averaging the individual survival trees prediction for final prediction. Hothorn et al. [3] introduced the averaging of single tree predications for the aggregation of survival function instead of majority voting.

Suppose that $\mathcal{L} = (\mathbf{X}, \mathbf{Y})$ be the given dataset. The $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_d\}$ is the set of d features where each feature is a record of n observations. The response $\mathbf{Y} = \{y_1, y_2, \ldots, y_n\}$ is a vector of n values where each y_i takes on the value as defined in Equation 2.1. The steps of bagging procedure are [3]:

- From the given training data draw a total of B bootstrap samples, where approximately 37% of the original data will be missing from each sample due to sampling with replacement.

- Grow a survival tree for each bootstrap sample by calculating the corresponding splitting criterion for each candidate variable. Repeat the same steps for all candidate variables ensuring that, the number of events in all terminal nodes to be homogeneous and equal to the pre-defined minimum node size.

- For each and every terminal node of each survival tree, calculate the required hazard function estimator.

- Calculate the prediction estimate of the hazard function by dropping the new ob-

servation down in the tree and taking weighted aggregate of the estimated hazard function over all trees.

A general work-flow of the bagging procedure is given in Figure 2.1.

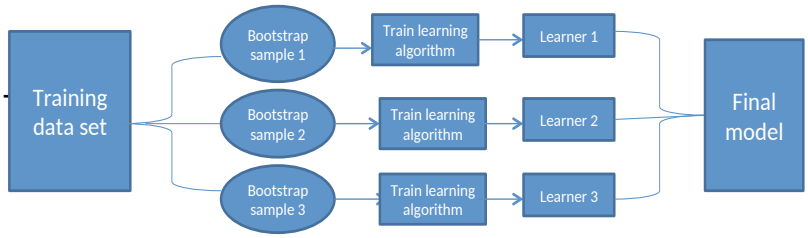

Figure 2.1: *A general workflow of bagging procedure.*

Survival trees bagging achieved great importance in variable selection and prediction in genomic study, where the main goal is to detect the ratio of non-susceptible individuals to susceptible individuals [41]. On the other hand, bagging shows less efficiency in situations where single predicator is extremely stable [40]. Aggregation increases the predictive accuracy but still it is considered as a black box of multiple trees [3].

2.4.2.2 Random survival forest (RSF)

Survival trees, ideal candidates for combination by means of an ensemble method can be converted into very robust predictive tools, such as survival forests [5]. Actually, ensemble is constructed from base learners such as trees to improve the prediction performance. This ensemble learning can further be improved by introducing randomization into base learning process in two forms. First, growing trees on bootstrap samples drawn randomly from given data. Secondly, selecting subset of variables(covariates) randomly as candidate

variables for splitting, at each node of the tree. This approach is called random forest [42]. Random survival forest (RSF) is actually the extension of random forest and bagging for survival data. The use of a random subset of features for the selection of attributes in a certain node based on a given splitting criterion is the only property of random forest that differentiate it from bagging.

In RSF, each tree is grown on independent bootstrap samples from original data excluding 37% of the given data. The data not included in bootstrapping are called out-of-bag observations which are used for the calculation of prediction error at the end. After bootstrapping, splitting is required that is split each node in the grown tree by selecting a candidate variable (from the random subset of features) which maximize survival difference between daughter nodes. RSF splits trees continuously until terminal nodes reach the same type of cases or further division of daughter/child nodes is not possible. Ensemble cumulative hazard function (CHF) is obtained by averaging the individual cumulative hazard functions calculated for each tree [2].

- Draw B bootstrap samples from the given original training data.
- On each bootstrap sample grow a survival tree by randomly selecting a subset of $p < d$ features at each node while splitting the nodes of each tree. Split on each node is decided on the candidate variable that maximizes survival difference between daughter nodes.
- Continue the above process until the terminal node has greater than or equal to a specific number of unique deaths.
- Average the CHF calculated for individual trees using the required estimator to obtain

CHF for ensemble.

Pseudo code of random survival forest is given in Algorithm 1.

Take $\mathcal{L} = (\mathbf{X}, \mathbf{Y})$ as the training data; B as the number of bootstrap samples,

RSF.Ensemble = **NULL**;

for $b = 1 \rightarrow B$ **do**

  Take a bootstrap sample S_b from $\mathcal{L}$;

  Call **GrowTree**(S_i) and save result as $\hat{S}_b$;

  if $b == 1$ **then**

    RSF.Ensemble = $\hat{S}_1$;

  else

    RSF.Ensemble = *Combine*(*RSF.Ensemble*, $\hat{S}_b$);

  end

end

GrowTree(S_b);

Qroot = S_b or a subset of S_b;

{

if *Qroot has greater than or equal to a specific number of unique deaths* **then**

  return (*Qroot*);

else

  Randomly select p candidate variable from all d features;

  Select the feature that best split *Qroot* to create child nodes of *Qroot*;

  for $j = 1 \rightarrow |child\ nodes|$ **do**

    GrowTree(child node);

  end

end

}

Algorithm 1: Pseudocode for random survival forest algorithm

2.4.2.3 Merits and demerits of RSF

RSF can handle the problem of multicollinearity and can identify predictive variables. Similarly RSF reduces over fitting of training data by using bootstrap samples [2]. RSF is independent of hypothesis testing and uses raw data directly for the computation of RSF models. Only few parameters need to be specified because of largely automated process.

RSF prefer continues variables to split node from data consisting of continuous and categorical variables [2]. RSFs are able to model complex interactions and non-linear effects which make it an attractive tool for the complex survival data analysis. RSF puts equal weights on all terminal nodes [43]. Some of the observations are assigned high weights and thus appears in every bootstrap sample due to which the computation of out-of-bag error rate estimates becomes impossible. Thus out-of-bag prediction in such cases is the main drawback of random survival forest [28]. RSF uses log-rank test statistic as the splitting criterion and selects only those variables that have several split points. Thus RSF favours splits for covariates with many possible split points and hence, induces bias in other parameters estimates e.g variable importance, which in turn effect the prediction accuracy [44].

2.4.2.4 Conditional inference forest

Conditional inference forests (CIF) that uses linear-rank statistic by default as splitting criterion is an approach to reduce split points bias in RSF by using two step approach. In the first step, a log-rank transformation test is performed to check if any independent variables are associated with the given response variable. The variable is considered to be best for splitting if the association is found to be significant with a small p-value of covariate

otherwise no splitting is conducted. The p-values are obtained through permutation. Permutation means the repeated rearrangement of labels on the observed data points and computation of the relevant test statistic for each rearranged data point. This is a way of obtaining the distribution of the test statistic under the null hypothesis of no difference, no association, etc. The best split point is found in the second step where the algorithm makes a binary split in the selected best variable, dividing the dataset into two sub-sets. For all possible partitions of the split variable two-sample linear statistics are calculated and compared to find optimal split point. For each subset the above discussed two steps are repeated until no variable is found to be associated with the outcome at the pre-defined level of statistical significance. The algorithm is thus called recursive. Ensemble of conditional inference trees have several advantages over traditional approaches. First, pruning (i.e. simplifying) the resulting tree to avoid over fitting is not required. Second, the algorithm also returns the p-values that give confidence to the programmer about every split. The ensemble of trees grown using conditional inference results into a conditional inference (CIF) model. The CIF model put more weight on terminal nodes where there is a large number of subjects at risk. This is because of the use of a weighted Kaplan-Meier estimate which is based on all subjects from the training dataset [43]. CIF perform linear rank statistic (the statistics that are used as tests in survival analysis as generalised non-parametric methods for testing the null hypothesis of equal survival distributions among groups. Such tests are commonly used when making comparisons among two or more than two groups or when comparing a single group with a known or hypothesized group [45]. The logrank test is an example of such tests.) between split variables and responses and finds optimal split points by comparing two-sample linear statistics for all

possible partitions [46].

2.5 Other related work

Several other machine learning methods are used for the analysis of survival data. Conditional inference forests (CIF), a machine learning method, has been suggested for survival analysis like random survival forest (RSF) and bagging survival trees discussed in previous sections. CIF uses different statistical approaches for the selection of split variables and split points. Therefore, to avoid this change an other method called maximally selected rank statistic [29] is used. The proposed method deals with the non-linearity in the covariates and hence reduce bias [29].

Maximally selected rank statistic is a statistic that allows the evaluation of cut-points on continuous or ordinal predictor variable which leads to the classification of observations into two classes. This test does not need to transform the time-dependent end point and calculates an exact cut-off point. The discrimination power of the cut-off point is also evaluated and estimated with a p-value. The statistic is applied to obtain an exact p-value for classification and to asses the effect of selected cut-points in case of binary class problems instead of log-rank statistic [47]. Generalized maximally selected rank statistic is used to evaluate large number of cut points and to analyze the asymptotic distribution of these maximally selected statistics [48]. Moreover, an algorithm for the exact distribution of a linear rank statistics is extended to calculate a lower bound in order to get the exact distribution of maximally selected rank statistics [49].

2.6 Chapter summary

This chapter gave a general introduction to survival analysis and the basic preliminaries used in the field. Some widely used methods for survival analysis are also discussed in the chapter. These methods include the Cox proportional hazard model, survival tree, random survival forest, bagging for survival trees and some other related methods. Advantages and disadvantages of the methods are also given in this chapter.

Chapter 3

Research methodology

3.1 Introduction

This chapter gives the objectives of the work done and the basic methodology used in carrying out the research in this thesis. To this end benchmarking, benchmark datasets and packages used for the methods in the thesis are described. Methods used for the purpose of comparison are also given.

3.2 Objectives of the work

This research work is done aiming at the following main objectives:

1. Extending the notion of optimal tree selection for survival analysis.

2. Reducing the number of survival trees in the forest.

3. Achieving better/comparable results by using the reduced size ensemble.

3.3 Benchmarking

In any research field learning and its generalization are the main issues. In machine learning approach the main concern is learning based on training data and generalizing it to unseen/testing data. For this purpose machine learning methods are desired to go through some standard solutions. These standards are called benchmarks. To inform programmers of the best choice of algorithms for their task at hand is a primary goal of benchmarking. It establishes the strengths and weaknesses of different machine learning implementations when applied to distinct types of data. According to Bache and Lichman [50] standard benchmark problems are given in the repository of machine learning or some other sources that are used to compare newly proposed methods with some other state-of-the-art methods.

3.3.1 Benchmark datasets

The datasets used for the purpose of benchmarking are called benchmark datasets. In this work, 17 such datasets are considered for the assessment of the performance of the proposed method in comparison to other state-of-the-art methods. The datasets are open problems from various sources used to evaluate and compare different learning algorithms. These datasets are summarized briefly in Table 3.1. Against each dataset number of observations, number of features and type of features whether real, integer or nominal is given.

The datasets used here are all of survival type. The description and sources from where these datasets have been taken is given in the following section.

Table 3.1: *Datasets description: Number of observations, number of features, type of features whether integer, real, or nominal (I/N/F) and data source are given against each dataset.*

Datasets	No. of observations	No. of features	Features type	Censored observations	Source
kidney	119	3	(2/1/0)	26	[51]
twins	24	4	(4/0/0)	8	[51]
kidtran	863	5	(5/0/0)	140	[51]
channing	462	5	(5/0/0)	176	[51]
Hodg	43	6	(6/0/0)	26	[51]
myeloid	646	6	(5/0/1)	320	[52]
veteran	137	8	(0/7/1)	128	[52]
retinopathy	394	9	(5/1/3)	155	[52]
bfeed	927	10	(10/0/0)	892	[51]
GBSG2	686	10	(7/0/3)	299	[53]
NKI	295	14	(0/8/6)	79	`https://www.ncbi.nlm.nih.gov/gap/?term=phs000547.v1.p1`
cgd	203	15	(6/4/5)	76	[52]
colon	1858	15	(0/15/1)	920	[52]
cost	518	15	(4/1/10)	404	[53]
burn	154	17	(17/0/0)	99	[51]
Pbc	418	19	(11/7/1)	347	[52]
BMT	137	22	(22/0/0)	81	[51]

3.3.2 Veteran

The dataset Veteran consists of the data from the Veteran's Administration Lung Cancer Trial (Kalbfleisch and Prentice) [54]. It is a standard survival analysis dataset and is available in `survival` [52] R package. It is the randomized trial of two treatment procedures for lung cancer. The dataset consists of a total of 137 observations with 8 variables. The variables consist of the type of lung cancer treatment (i.e. 1-standard 2-test drug), cell type, Status that denotes the status of the patient as 1 if dead or 0 if alive, survival time in days since the treatment, Diag represents the time since diagnosis in months, age in years, the Karnofsky score, therapy that denotes any prior therapy (0=none, 1=yes).

3.3.3 kidtran

The times to infection of kidney dialysis patients (kidtran) dataset has 863 observations and 5 variables. The original source of the dataset is "Survival Analysis Techniques for Censored and Truncated Data" [10] and is freely available in *R* package `KMsurv` [51]. The variables are gender (1-male, 2-female), race i.e. 1 if white and 2 if black, age in years, time which shows period of study, death indicator delta as 0 if alive otherwise 1(dead). The objective of the study is to assess the time to first clinically apparent infection in a group of patients suffering from renal insufficiency.

3.3.4 Twins

This dataset consists of a total of 24 observations on 4 features. The dataset is about the patients died from coronary heart disease. The features are named as id, age, death and

gender. The response variable is shown in months as age followed by an indicator death (i.e. 1 if died from CHD, 0 otherwise). The other two variables show the identification number and gender whether male or female of the subject under study. The dataset is used in "Survival Analysis Techniques for Censored and Truncated Data" [10] and is readily available in R package `KMsurve` [51].

3.3.5 Hodg

The hodg dataset has 43 observations and 6 variables. This dataset contains the variables gtype (graft type 1 for allogenic and 2 for autologous), dtype (disease type: 1 and 2 for Non Hodgkin lymphoma and Hodgkins disease respectively), Time to death, delta (death/relapse indicator (0 if alive otherwise 1), Karnofsky score and Waiting time in months to transplant. The dataset is used in "Survival Analysis Techniques for Censored and Truncated Data" [10] and readily available in R package `KMsurve` [51].

3.3.6 bfeed

bfeed (breast feeding) data is taken from the survey asking female about any pregnancies that have occurred in 1983 since they were last surveyed, conducted by "National Longitudinal Survey of Youth" that begun in 1979. This dataset consists of the breast feeding related information, one of the main section of survey questioner. The information is taken from 927 mothers (became mothers for the first time) who select breast feeding and gave answers for all the variables of interest. The children born after 1978 and having $20 <$ gestational age < 45 weeks are included in the survey. Duration is the response variable in the dataset followed by an indicator whether the weaning to infant is completed

or not. Race of mother is represented as 1, 2, 3 whether white, black or other, poverty status of a mother is indicated by 1 if mother is in poverty, the status of mother whether smoking and drinking at child birth are recorded as separate variables, that is one if yes or 0 otherwise, age and education of mother and the year in which the child born are other information recorded in this dataset. Prenatal care in first three months (first trimester) of pregnancy and after that are taken as explanatory variables. The dataset is originally from "Survival Analysis Techniques for Censored and Truncated Data" [10].

3.3.7 kidney

This dataset is the record of the time assessment to first clinically apparent infection at the exit site in patients who has a defect in renal i.e. inability to clear waste products from body. About the placement of a catheter two choices, whether surgical placement or percutaneous placement, is given to each patient. There are 43 patients who select surgery while 76 patients utilized a percutaneous placement of their catheter. Thus a total of 119 persons are observed on 3 variables. Delta is used as censored indicator. This dataset is available free in `KMsurv` R package [51].

3.3.8 cgd

This dataset has been taken from a placebo controlled trial of gamma interferon in a chronic granulotomous disease (CGD). The dataset consists of a records on time to serious infections observed in patients through the end of the study. There are a total of 15 variables in this dataset. These variables are enrolling centre, treatment whether placebo or gamma interferon, sex, age, entry at study, height, weight (in kg), inheritance pattern,

use of steroids, use of prophylactic antibiotics, 4 groups of centres, days to last follow-up, start and end of each time interval and observation number within subject with the status 1 if the interval ends with an infection otherwise 0 as a censoring indicator. The original source of the data is "Counting Processes and Survival Analysis" [55] and the data is freely available in package `survival` [52].

3.3.9 Burn

This dataset consists the 18-months treatment records of burned patients.

The dataset provide information about the care methods used for the burned patients, infections in burn wound and other medical cares. During study period the time in days until staphylococcus infection was recorded. The information that whether an infection had occurred or not is recorded as an indicator variable. Gender, race, burn site, severity and type of burn are other fixed covariates recorded in the under discussion study. Excision time and time to prophylactic antibiotic treatment administered are also recorded during study. Two other covariates dependent on time, namely, whether the patient's wound had been excised or not and whether the patient sometime during the course of the study had been treated with an antibiotic or not are recorded as indicator variables.

A total of 154 patients are observed in which 84 patients received chlorhexidine, the new bathing solution, while other 70 patients served as the historical control group by giving them povidone-iodine, the routine bathing care. The original source of the data is the "Survival Analysis Techniques for Censored and Truncated Data" [10] while in the `KMsurv` R package [51] the dataset is available readily.

3.3.10 GBSG2

This dataset contains observations on 686 women who suffered from breast cancer. The original source of the data is "Building multivariable prognostic and diagnostic models: transformation of the predictors by using fractional polynomials" [56]. The features in the data are age of the patients in years, time of recurrence free survival time (in days), hormonal therapy, a factor at two levels yes and no, menopausal status, a factor at two levels pre (premenopausal) and post (postmenopausal) as horTh and menostat respectively. Variables tsize and tgrade denote the size of tumour in mm and grade of tumour, an ordered factor at levels $I < II < III$ respectively while, number of positive nodes and progesterone receptor are shown by variables names pnodes and progrec respectively. Censoring indicator is represented by cens i.e 0 if censored and 1 if event happens. The estrogen receptor is showed by variable named estrec. The dataset is freely available in the `pec` R package [53].

3.3.11 Channing

This dataset consists the information of individuals ages at death who were in residence in Palo Alto, California from January 1964 to July 1975. A health care program provided by the centre observed a total of 462 individuals, whose allowed for easy access to medical care without any additional financial burden on the resident.

The original source of the data is the "Survival Analysis Techniques for Censored and Truncated Data" [10] while in the `KMsurv` *R* package [51] the dataset is available for free. There are a total of 5 variables in the dataset i.e. status of death (1:dead otherwise 0), age (in months) of entry into retirement home, age of death or left retirement home, time as a

difference between the above two ages and gender i.e. 1 for male and 2 for female.

3.3.12 retinopathy

The dataset is based on a trial to delay diabetic retinopathy through laser coagulation treatment. A data frame consists of 394 observations on 9 variables i.e, type of laser used, treated eye whether right or left, age of a person at diagnosis time, type of diabetes, trt that is 0 for control eye and 1 for treated eye, time to loss of vision and eye risk score. A status variable 0 if censored and 1 otherwise is recorded as censoring indicator where censoring is caused by death, remove from study or end of the study. For each patient there are two observations in the dataset, one for the eye received laser treatment and the other for the untreated eye. The time when treatment starts to the time when visual acuity dropped below 5/200 is considered as the event of interest for each eye. Survival times in this dataset are the difference between actual time when vision lost and minimum possible time to event.

The dataset is available in `survival` R package [52].

3.3.13 myeloid

This dataset consists of 646 observations on 6 variables. These variables are, treatment arm (A or B), time to death represented by futime 1 for those who died and 0 for censored or lost, time to hematropetic stem cell transplant, complete response time and time to recurrence of disease. In `survival` R package [52] the dataset is available free of cost.

3.3.14 Pbc (Primary Biliary Cirrhosis)

This dataset has been taken from the trial conducted by Mayo Clinic between 1974 and 1984. The information about the primary biliary cirrhosis (PBC) of the liver were collected during the trial. The eligibility criteria for the randomized placebo controlled trial of the drug D-penicillamine was met by a total of 424 PBC patients in which first 312 cases participated in the randomized trial and contain largely complete data. The remaining 112 cases just consented to have basic measurements recorded and to be followed for survival.

The variables are arranged in the data as follows: age in years, serum albumin (g/dl), alkaline phosphotase as alk.phos, presence of ascites, aspartate aminotransferase, once called SGOT as ast, bili (serum bilirunbin), serum cholesterol as chol, urine copper, edema as 0 no edema, 0.5 untreated or successfully treated, 1 edema despite diuretic therapy and h.epato (presence of hepatomegaly or enlarged liver).

3.3.15 Colon

Colon cancer is actually a disease occurs due to the growth of out-of-control cell. It originate from small, noncancerous tumours called adenomatous polyps that form on the inner walls of the large intestine and damage healthy tissue that is near the tumour causing many complications. This dataset is about one of the first successful adjuvant chemotherapy trials for colon cancer. A low-toxicity compound Levamisole that is used previously for the treatment of worm infestations in animals while 5-FU is a moderate toxic chemotherapy agent. Per person two records, one for recurrence and one for death are collected in the under discussion study. There are total of 1858 subjects observed on

15 variables. The variable are study i.e. 1 for all patients, (Treatment - Obs(ervation), Lev(amisole), Lev(amisole)+5-FU) namely rx, sex and age of patient, colon obstruction and perforation, adherence to nearby organs, number of cancer detectable lymph nodes, days until event or censoring, differentiation of tumour as 1, 2, 3 whether well, moderate or poor), Extent of local spread i.e. 1:submucosa, 2:muscle, 3:serosa, 4:contiguous structures, time from surgery to registration, positive lymph nodes that is more than 4, type of event and status as a censoring status.

3.3.16 BMT

This dataset consists of the information about the process of recovery from a bone marrow transplantation. A total of 137 patients were treated during study for maximum follow-up of 7 years. During follow up time it is observed that 42 patients are relapsed and 41 died, 26 patients had an episode of acute GVHD. The number of patients whose either relapsed or died in remission without their platelets returning to normal levels, recorded is 17. Several potential risk factors were measured at transplantation time. For each disease, patients on the bases of their status at transplantation time were grouped into risk categories. These categories are Disease Group 1, 2 and 3 for ALL, AML Low Risk and for AML High Risk respectively. Other risk factors denoted from z1 to z10 consists of recipient and donor age and gender, cytomegalovirus immune status, waiting time from diagnosis until transplantation, French-American-British (FAB) classification based on standard morphological criteria, Hospital i.e 1, 2, 3 and 4 for the Ohio State University, Alferd, St. Vincent and Hahnemann respectively. MTX is recorded as a graft-versus-host-Prophylactic that is 1 if Yes and 0 otherwise. t1 represents time to death or on study time

while t2 is a survival time free from any disease. d1, d2 and d3 show death indicator (i.e 1 for those who died and 0 for others), relapse indicator and survival indicator that is completely disease free while da, dc and dp variables shows acute GVHD indicator and chronic GVHD indicator and platelet recovery indicator respectively. Variables ta, tc and tp show time to acute and chronic Graft-Versus-Host disease.

3.3.17 NKI

In this dataset the gene expression measurements of 337 lymph node positive breast cancer patients are recorded. Originally, the dataset included 3322 patients, 17 clinical variables and 693,543 SNPs. The benchmarking is performed on 2781 individuals and 331,195 SNPs after application of a standard quality control and linkage disequilibrium pruning ($r2 > 0.7$) [29]. Finally, the computational burden is reduced and missing data is eliminated by excluding all SNPs with a call fraction below 100% keeping 151,346 SNPs. The endpoint relapse-free survival are analysed. This dataset is available at dbGaP and has ID `phs000547.v1.p1` (`https://www.ncbi.nlm.nih.gov/gap/?term=phs000547.v1.p1`).

3.3.18 cost

This dataset is actually a subset of the data from the Copenhagen stroke study. A total of 518 stroke patients are observed. There are a total of 14 features i.e. age and sex, Hypertension, History of ischemic heart disease at the time of admission, history of previous strokes before admission, history of other disabilities (e.g. severe dementia), daily alcohol consumption, mellitus status of diabetes that indicates the glucose level higher than 11 mmol/L, daily smoking status, atrial fibrillation, hemorrhage (stroke subtype), strokeScore, cholesterol

level are recorded with the survival time and status (0: censored, 1: event). The dataset is freely available in *R* package `pec` [53] while the original source is "Evaluating random forests for survival analysis using prediction error curves" [57].

3.3.19 Software and packages

R is a computer programming language that provides a friendly environment for every body to participate. This is a free of cost language with a huge repository of utilities for solving various problems that are regularly developing. Therefore, for execution of algorithms, obtaining the results for the methods on the benchmark dataset in this thesis, *R* [58] programming language is used.

The corresponding R packages for the methods considered in this thesis are described as follows.

For random survival forest ensemble the package `ranger` [59] is used. This package is considered to be a fast implementation of the "randomForest" [60] R package originally developed for classification and regression. The hyper-parameters tuned, by 10-fold cross validation, are total number of trees, number of features selected at each node for splitting and terminal node size. These are denoted by `num.trees`, `mtry` and `min.node.size`, respectively in the `ranger` *R* package.

For bagging survival trees the *R* package `ipred` [61] is used. The only hyper-parameter that is tuned is the number of trees denoted by `nbagg` in the *R* package. Similarly for conditional inference forest, free available package `party` [37] is used. The total number of trees and number of features that are selected at each node for splitting the nodes are the only hyper-parameters denoted by `ntree` and `mtry` defined via *cforest_control* are tuned.

The package `survival` [62] is used for Cox proportional hazard model using default values for all parameters. For the variable initial iteration values, the default is zero for all features, i.e `init` is set to zero. Iterations continues until the relative change in the log partial likelihood is smaller than 1e−09, i.e. `eps`< 1e−09. The maximum iteration attempts for convergence is 20 by default. For tie handling, the Efron [63] approximation is used as the default i.e. `ties=''efron''`.

3.4 Error estimation

In survival analysis numerous methods are used to estimate the predictive performance of the model. Here, we use integrated Brier score as the performance metric for comparing the methods considered in this work. The following section describe how to calculate Brier score and its integration in detail.

3.4.1 Integrated Brier score (IBS)

IBS is calculated simply by taking the integration of Brier score [64]. Brier score (BS) is actually a measure of the mean squared difference between the actual outcome y_i and the predicted probability of the possible outcome for the ith observation at given time t. The advantage of BS over other common methods for prediction assessment is its correct classification in different outcome groups and agreement of the predictions with the true risk. In other words, these are called discrimination and calibration (see section 4.2.1) [65].

In survival analysis the Brier score is the squared difference between the survival function indicator and the predicted survival probability at given time t_0. In free censored sur-

vival data BS can easily be estimated by taking average of the squared distances between the indicator of survival function for the subjects and the survival probability predicted by the model for that subjects [65]. However, for right censored data which is the main concern here in this thesis, the integrated brier score technique is used.

Let the time of the event of interest of subject i be T_i, the outcome is $\delta_i = 1(T_i > t_0)$ and $\hat{S}(t_0|\mathbf{x}_i)$ is the probability estimate of a subject at risk observing at t_0, where t_0 is the exact follow-up time and $\mathbf{x}_i$ are the given covariates then the BS is given as [65]

$$\begin{aligned} BS(t_0) &= E(1(T_i > t_0) - \hat{S}(t_0|\mathbf{x}_i))^2, \\ &= E(1(T_i > t_0) - S(t_0|\mathbf{x}) - \hat{S}(t_0|\mathbf{x}_i) + S(t_0|\mathbf{x}))^2, \\ &= E(1(T_i > t_0) - S(t_0|\mathbf{x}))^2 + E(\hat{S}(t_0|\mathbf{x}_i) + S(t_0|\mathbf{x}))^2. \end{aligned}$$

The censored indicator δ_i is not always easy to calculate. If an individual i survived at least until time t_0 i.e $1(T_i > t_0)$ then $\delta_i = 1$ otherwise 0. The indicator is consider to be unknown if the individual is censored before t_0 i.e $T_i < t_0$. Inverse probability of censoring weighting (IPCW) [66] is used to overcome this issue. Hence, the Brier score can be written as:

$$\hat{BS}(t_0) = \frac{1}{n}\sum_{i=1}^{n}(1(T_i > t_0) - \hat{S}(t_0|x_i))^2 w_i,$$

where

$$w_i = \begin{cases} 0, & \text{if } T_i > t_0 \text{ and } o_i = 0, \\ \frac{1}{\hat{G}(t_0)}, & \text{if } T_i > t_0, \\ \frac{1}{\hat{G}(T_i)}, & \text{if } T_i < t_0 \text{ and } o_i = 1, \end{cases} \tag{3.1}$$

where $\hat{G}(t_0)$ is the Kaplan-Meier estimate of the probability of being uncensored at time t_0. The predictive performance of the model is said to be better for the lower values of the Brier score. 0 indicates perfect predictions, however, in practice it is very rare or may be impossible to get 0 value for the brier score.

The Brier score defined above is a function of time t_0 where in right censored data measure of predictive accuracy over a range of time points is required. Therefore, an integrated Brier score (IBS) with respect to some weight function can be estimated by integrating the BS calculated through the aforementioned method [67].

The integrated Brier score (IBS) for $t \in (0, t^*)$ is calculated as follows:

$$IBS(t_0) = \int_0^{t^*} \hat{BS}(t_0) dt \; \hat{w}_{(t_0)},$$

where $\hat{w}_{(t_0)}$ is a function that weight the Brier scores at individual time points. It is a straightforward trapezoidal rule that integrate the area under the prediction curve [67].

Chapter 4

Optimal survival trees ensemble

4.1 Introduction

The main objective of monitoring and analyzing survival data for long period of time is to estimate the time of occurrence of a particular event of interest in the best way. To increase the predictive performance of the survival model tree based approaches might be used. Intuitively, a survival forest of accurate and diverse survival trees may perform better than a forest consisting of simple survival trees. This intuition also hold for the Breiman's [1] forest of regression and classification trees, which has led to the method of optimal trees ensemble (OTE) for regression and classification [6].

Ensemble of optimal trees refines the idea of bagging and random forest based on the Breiman's upper bound for the overall prediction of random forest given as:

$$(PE)^* \leq \bar{\rho} PE_j,$$

where PE^* denotes the overall prediction error of random forest, $\bar{\rho}$ is the weighted correlation between residuals from two independent classification or regression trees and $PE_j, j = (1, \ldots, B)$ is the prediction error of the jth classification or regression tree in the forest. B is the total number of trees in the forest.

OTE selects the best tree from a large number of trees initially grown by random forest on the basis of their individual and collective performance. On the bases of individual performance on out-of-bag observations, a proportion of top ranked trees are selected and assessed on an independent training data for their collective performance using Brier score. A tree is said to be suitable for the final ensemble if its addition to the previously added trees increase its predictive performance [6]. OTE method select trees for final ensemble one by one starting from the tree with the highest prediction accuracy. The method has been shown to give comparable results using fewer trees than some other state-of-the-art methods considered.

Here it is aimed to extend OTE to survival data and to select the best survival trees from an initial ensemble in terms of their individual predictive accuracy as well as their contribution to the ensemble and integrate them together to develop a new forest ensemble. This ensemble will be called as optimal survival trees ensemble (OSTE). Using a total of 17 benchmark datasets, the results from OSTE are compared with those of bagging survival trees, random survival forest, conditional inference forest and Cox proportional hazard model.

4.2 Optimal ensemble of survival tree (OSTE)

Optimal survival trees (OSTE) is an attempt to refine the idea of random survival forest by assessing survival trees not only on their collective performance but also on their individual performance. To obtain the ensemble, divide the given training data $\mathcal{L} = (\mathbf{X}, \mathbf{Y})$ into two parts $\mathcal{L}_B = (\mathbf{X}_B, \mathbf{Y}_B)$ and $\mathcal{L}_V = (\mathbf{X}_V, \mathbf{Y}_V)$. $\mathcal{L}_B$ and $\mathcal{L}_V$ are random and non-overlapping partitions of the training data. From $\mathcal{L}_B = (\mathbf{X}_B, \mathbf{Y}_B)$ draw B bootstrap samples and grow survival tree on each sample. To induce more randomness select a subset of $p < d$ features at each node of the tree. During bootstrapping some observations are left out of samples which are called out-of-bag (OOB) observations. In training the corresponding model, OOB observations play no role, these observations, however, could be used as test data from each of the corresponding bootstrap sample for prediction error of individual survival trees. The grown survival trees are arranged in ascending order according to their C-index (introduced in Section) and the top M trees are selected. To check the diversity of the selected trees, they are tested one by one as follows:

- To get the final ensemble of survival trees, trees collective performance is assessed on the independent training data $\mathcal{L}_V = (\mathbf{X}_V, \mathbf{Y}_V)$. The second best survival tree is combined with the best survival tree and the resultant ensemble is assessed by using the training data $\mathcal{L}_V = (\mathbf{X}_V, \mathbf{Y}_V)$. The third best survival tree is added to the ensemble of size two and the performance is measured similarly. A survival tree is selected for the final ensemble if its addition decreases the prediction error of the ensemble. This is done for all the M survival trees.

- A survival tree, $\hat{L}_k$ where $k = (1, 2, \ldots, M)$ is chosen for the final ensemble of optimal

survival trees if its addition to the ensemble with out k^{th} survival tree fulfills the following criterion

$$IBS^{(k-)} > IBS^{(k+)},$$

where $IBS^{(k+)}$ is the integrated brier score (IBS) of the ensemble including the k^{th} tree and $IBS^{(k-)}$ is the integrated brier score of the ensemble in which the k^{th} tree is not included yet.

4.2.1 Concordance index

Discrimination and Calibration [68] are two measures used to check the predictive performance of mathematical model having binary outcomes. Discrimination check that how much ability the model has to classify the subject into relevant class while calibration describe that how closely the actual outcomes relate to the numerical values of the predicted probabilities. Discrimination is the most preferred one due to the fact that re-discrimination is not possible unlike calibration. On the other hand implementation of discrimination does not affect the calibration.

In survival analysis, for each subject we have survival times and predictions about them unlike logistic regression where each subject has to fall into one of two possible categories. This makes discrimination in survival analysis difficult. Survival model can be evaluated by considering the relative risk of an event for different subjects instead of the absolute survival times for each subject. The concordance index (C-index) [69] is one of the suggested measures. It is the extension of the concept of the receiver operating characteristic (ROC) curve area (the area under the receiver operating characteristic curve) where ROC measure the discriminative ability of the under discussion biomarker at each time point [70].

Concordance index is one of the most reported metrics to predict biomarkers in survival setting. It is also used for the comparison of models derived in different statistical cultures, such as a Cox regression model and a random survival forest model [42]. A pair is said to be concordant if an individual predicted risk is high while the survival time is short. Among all pairs of an individual the relative frequency of the concordant pair is called C-index when the data under study is not censored. Briefly, the C-index can be described as the probability that a subject under study with a small survival time is associated with a high value of an indicator (biomarker) and vice versa. In other words C-index measures the ability of a biomarker to discriminate between subjects with small survival times and subjects with large survival times. This strategy is very helpful in the field of biomedical research where patients are needed to be subdivided into groups with good or poor prognosis [70]. It is also applicable to continuous, ordinal and dichotomous outcomes. For survival outcomes, the C-index is defined as:

$$C = P(\delta_1 > \delta_2 | T_1 < T_2), \tag{4.1}$$

where T_1, T_2 and δ_1, δ_2 are the event times and the predicted biomarker values [70], respectively. The biomarker value shows closeness to a perfect discriminatory power for $C = 1$ while for $C = 0.5$ a marker does not perform well.

4.2.2 OSTE Algorithm

The algorithm of the proposed method OSTE consists of the following steps:

- Partition the training data into two non-overlapping parts $\mathcal{L} = (\mathbf{X}, \mathbf{Y}) = \mathcal{L}_B$ and $\mathcal{L}_V$.

- Draw B bootstrap samples from the data $\mathcal{L}_B = (\mathbf{X}_B, \mathbf{Y}_B)$.
- On the bootstrap samples grow survival trees in such a way that the splitting variable is chosen from $p < d$ features at each node.
- Arrange the grown trees on the bases of their individual prediction error on OOB data in ascending order and select the top M trees. The prediction error is estimated via concordance index given in Section 4.2.
- Add the M selected trees one by one and calculate integrated Brier score. Check the performance on validation data $\mathcal{L}_V = (\mathbf{X}_V, \mathbf{Y}_V)$. Select the survival tree if the results are improved otherwise discard.
- New data are predicted by combining the results of the selected trees in the final ensemble.

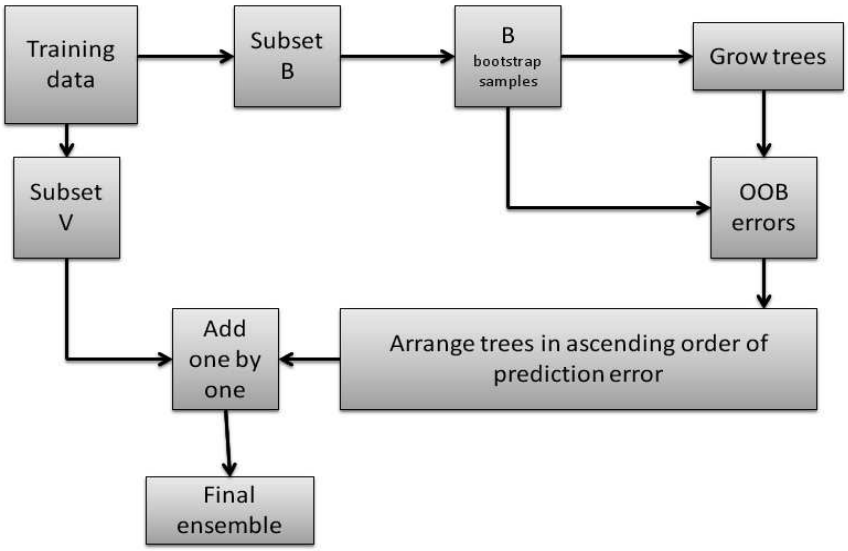

Figure 4.1: *Work flow of the OSTE method.*

Take $\mathcal{L}_B = (\mathbf{X}_B, \mathbf{Y}_B)$ as the training data; B as the number of bootstrap sample, ;
for $i = 1 \rightarrow B$ **do**
    Take a bootstrap sample s_i from $\mathcal{L}_B$;
    Store the observation left out from the i^{th} sample as OOB(i);
    Make a root node Q_{root} containing s_i;
    Call **GrowTree**(s_i) to grow tree ith tree ;
    Estimate prediction error for ith tree using OOB(i)
end
GrowTree (s_i);
$Qroot = s_i$ or a subset of s_i;
{
if *Qroot has greater than or equal to a specific number of unique deaths* **then**
    return (*Qroot*);
else
    Randomly select p candidate variable from all d features;
    Select the feature that best split *Qroot* to create child nodes of *Qroot*;
    for $j = 1 \rightarrow |child\ nodes|$ **do**
        GrowTree(child node);
    end
end
}
Select the best trees based on individual performance;
{
for $j = 1 \rightarrow B$ **do**
    if *prediction error of tree j is* $< q$ **then**
        select S_b;
        (q is a quantile point in the error distribution of B trees.)
    else
        Drop;
    end
end
}
Select the best survival trees based on ensemble performance ;
Initialize the ensemble from the first top ranked tree;
for $k = 2 \rightarrow M$ **do**
    if $\hat{BS}^{(k-)} < \hat{BS}^{(k+)}$ **then**
        Select the kth tree, where $\hat{BS}^{(k+)}$ is the Brier score of the ensemble having the kth tree and $\hat{BS}^{(k-)}$ is the Brier score of the ensemble with out the kth tree after applying data, $\mathcal{L}_V = (\mathbf{X}_V, \mathbf{Y}_V)$;
    else
        Do not select the kth tree
    end
end

4.3 Experiment and results

4.3.1 Experimental setup

Each data set is divided into training and testing parts. A random set of 70% of the total dataset is taken as the training while the remaining 30% of the dataset is left for testing purposes. The same training and testing parts are used for all the methods considered in the analysis for valid comparison.

In the case of OSTE, a total of 1000 independent survival trees are grown on bootstrap samples selected from 95% of the training data as the initial ensemble. For splitting the nodes of the trees, p features are randomly selected at each node from the total of d features while growing the trees. The remaining 5% of the training data is used for diversity check. For all datasets the number p of feature is taken as the square root of the total features i.e $p = \sqrt{d}$, which is also the default value in the standard random survival forest. Terminal node size is fixed at 3. Based on individual accuracy, $M = 20\%$ of total trees grown are selected. A total of 1000 runs are performed for each data set and the final results are the average from all these runs using the 30% test data. Log-rank statistic is used while growing the forest.

In the case of RSF, number of trees is tuned by using 10-fold cross validation on values from the set $\{500, 1000, 1500, 2000\}$. Tree are grown unproned with terminal node size equal to 3. The value of `mtry` is tuned by using 10-fold cross validation considering all possible values of the number of features for all the datasets on the corresponding training part. Log-rank statistic is used while growing the forest.

In the case of bagging for survival trees, number of trees is tuned by using 10-fold cross

validation on {500, 1000, 1500, 2000}. Trees are grown unproned. Log-rank statistic is used while growing the forest.

For conditional inference forest, number of trees are fine tuned using values from the set {500, 1000, 1500, 2000}. The `party` *R* package is used for getting the result with the rest of the parameters on their default values.

For Cox-proportional hazard model, the default setting is used as implemented in the R package described in Section 3.3.19 of Chapter 3.

4.4 Results and discussion

Using the experimental settings given in Section 4.3.1, integrated Brier scores for all the methods are calculated on the datasets introduced in Chapter 3. The results are given in Table 4.1 The results are the average integrated Brier scores from a total of 1000 repetitions of applying the methods each time randomly dividing the data into training and testing parts as described in the above section. The table show that OSTE is giving better average results than the others on 5 out of 17 data sets. RSF gives better results on 1 data set while bagging outperformed others on 5 datasets. The results of bagging and RSF were same on twins dataset. CIF gave better results on 4 and Cox on 3 datasets.

Figures 4.2-4.4 give the results of the methods for the 17 datasets in the form of box plots. The box plots for Cox, bagging, RSF, CIF and OSTE are shown by brown, gray, red, yellow and blue colors, receptively. Figure 4.2 gives the integrated Brier scores for all the methods on the datasets veteran, kidtran, bfeed, twins, GBSG2 and burn. OSTE shows better performance on kidtran and bfeed datasets while for other datasets the results are

Table 4.1: *Integrated Brier scores of the methods against each data set. The best score is highlighted in bold font.*

Datasets	n	d	Cox	bagging	RSF	CIF	OSTE
kidney	119	3	0.1272	0.1296	0.1296	**0.1257**	0.1291
twins	24	4	0.0144	**0.0132**	**0.0132**	0.0147	0.0139
kidtran	863	5	0.0341	0.0324	0.0144	0.0203	**0.0135**
channing	462	5	0.0584	**0.0512**	0.0554	0.0664	0.0550
Hodg	43	6	**0.1521**	0.1885	0.1836	0.1703	0.2067
myeloid	646	6	0.1393	**0.1348**	0.1349	0.1360	0.2474
veteran	137	8	0.2571	0.1707	0.1692	**0.1582**	0.1683
retinopathy	394	9	0.1757	0.1795	0.1765	**0.1714**	0.1762
bfeed	927	10	0.1925	0.2397	0.1942	0.1941	**0.1478**
GBSG2	686	10	**0.0148**	0.0151	0.0149	0.0182	0.0170
NKI	295	14	0.1510	0.1154	0.1113	**0.1077**	0.1110
cgd	203	15	0.2831	**0.0819**	0.0862	0.0831	0.0844
colon	1858	15	0.1737	**0.1534**	0.1605	0.1735	0.1897
cost	518	15	**0.1764**	0.1825	0.1807	0.1851	0.1789
burn	154	17	0.1661	0.1477	0.1474	0.1527	**0.1469**
Pbc	418	19	0.0669	0.0669	0.0504	0.0523	**0.0082**
BMT	137	22	0.0799	0.0450	0.0560	0.0511	**0.0299**

similar to the rest of the methods.

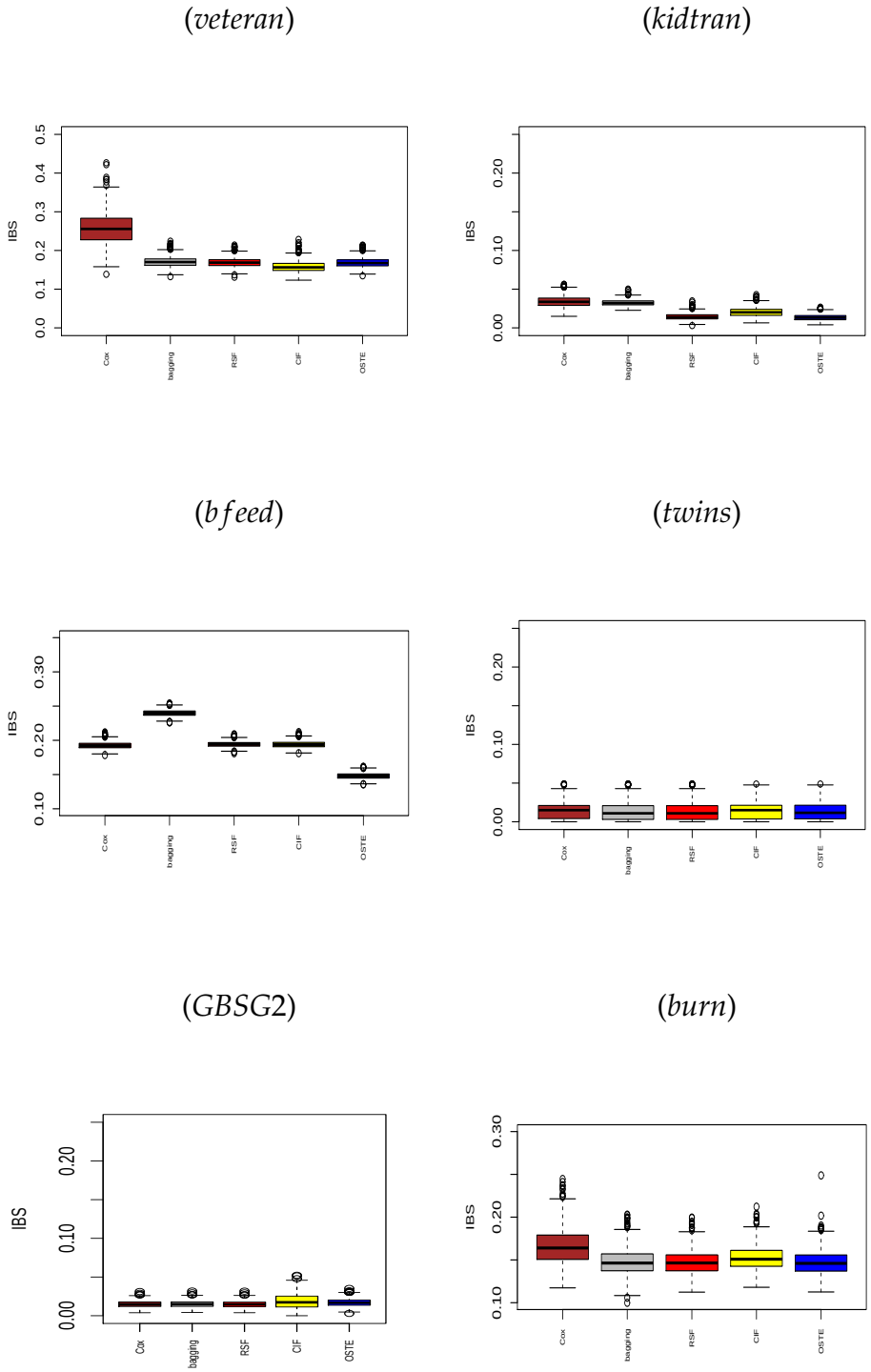

Figure 4.2: *The boxplots showing IBS on the datasets veteran, kidtran, bfeed, twins, GBSG2 and burn. Cox, Bagging, RSF, CIF and OSTE are shown by brown, gray, red, yellow and blue colors, respectively. OSTE shows better performance on kidtran and bfeed datasets while for others datasets the results are similar to the rest of the methods.*

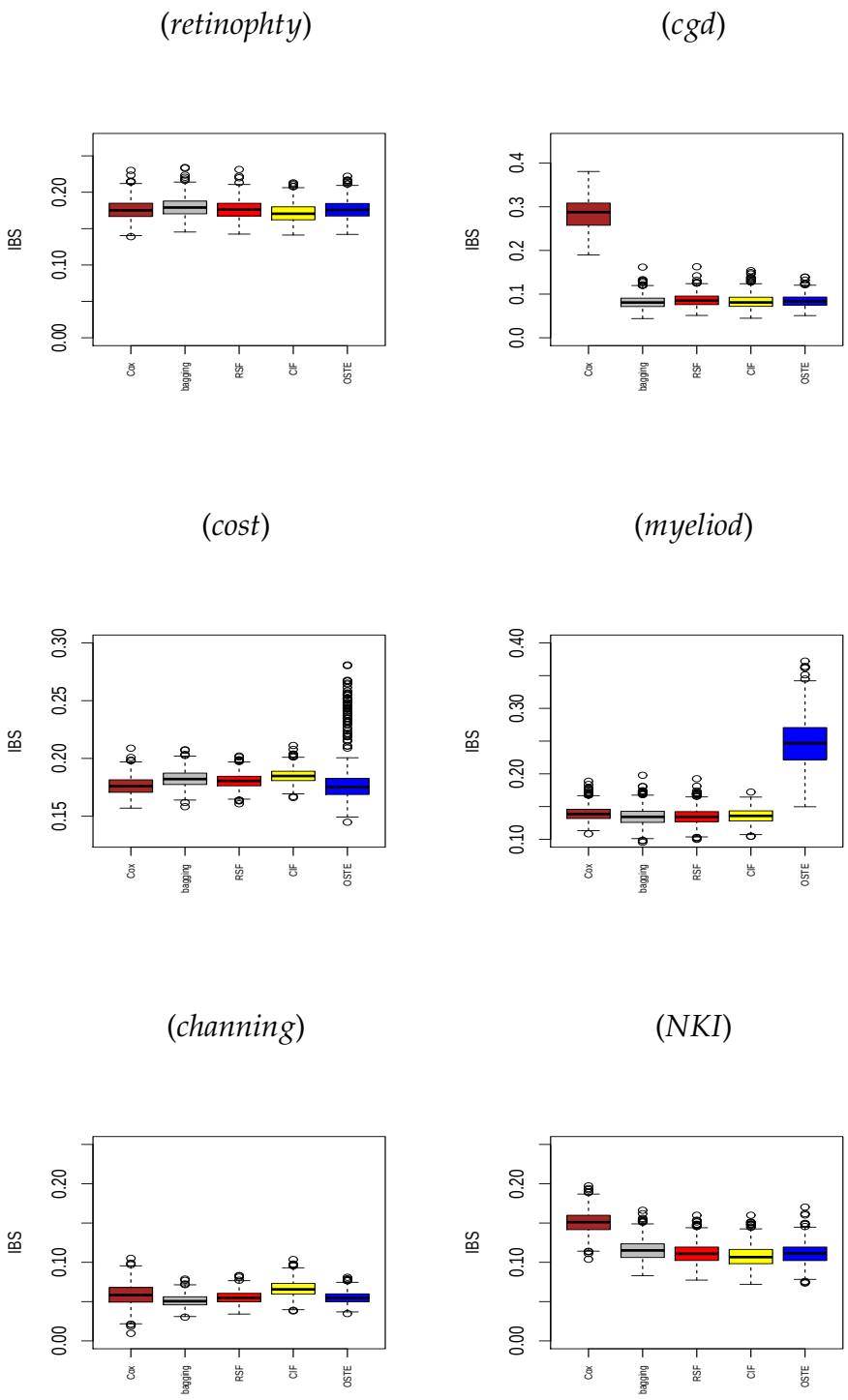

Figure 4.3: *The boxplots showing IBS on the datasets retinophty, cgd, cost, myeliod, channing and NKI. Cox, Bagging, RSF, CIF and OSTE are shown by brown, gray, red, yellow and blue colors, respectively. OSTE shows similar performance on all the datasets except myeliod.*

The boxplots given in Figure 4.3 show the IBS on the datasets retinophty, cgd, cost, myeliod, channing and NKI. Cox, Bagging, RSF, CIF and OSTE are shown by brown, gray, red, yellow and blue colors, respectively. OSTE shows similar performance on all the datasets except myeliod. Cox shows poor performance on cgd and NKI datasets. Results of the methods on other datasets are similar to the rest of the methods.

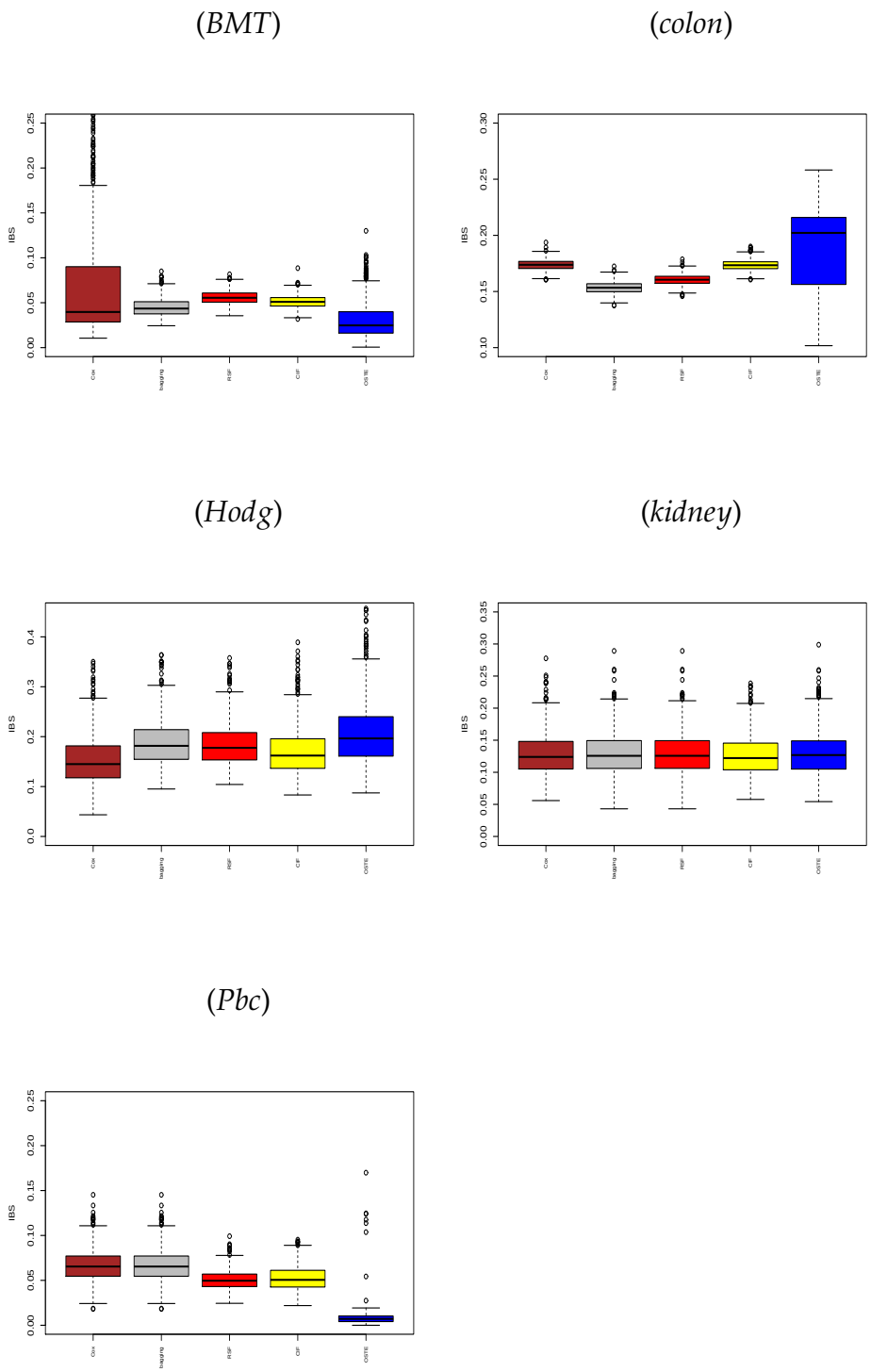

Figure 4.4: *The boxplots showing IBS on the datasets BMT, colon, Hodg, kidney and Pbc. Cox, Bagging, RSF, CIF and OSTE are shown by brown, gray, red, yellow and blue colors, respectively. For kidney dataset OSTE shows similar performance while on Pbc and BMT datasets OSTE shows better results.*

The boxplots given in Figure 4.4 showing IBS on the datasets BMT, colon, Hodg, kidney and Pbc. Cox, Bagging, RSF, CIF and OSTE are shown by brown, gray, red, yellow and blue colours, respectively. For kidney dataset OSTE shows similar performance while the results of OSTE are better on Pbc and BMT datasets.

Due to tree selection with specific patterns, OSTE might give comparatively larger error estimates on some of the random splits of the data into training and testing parts. This may happen when patterns in the selected trees are not in-line with those in the test data. This can be seen in Figure 4.4 for Pbc data, for example.

As OSTE improves RSF by discarding trees from the original forest with adverse effects on its overall efficiency, a further comparison of the two methods is given in terms of feature importance. Feature importance for both the methods is estimated via the permutation method [71]. For both the methods, a variable's permutation importance is estimated by randomly permuting the given variable in the out-of-bag (OOB) data for the tree, and the permuted OOB data is dropped down the tree. The OOB estimate of prediction error is then calculated. The estimate of the variable importance is the difference between this estimate and the OOB error without permutation, averaged over all trees. The larger the permutation importance of a variable, the more predictive the variable.

Variable importance on 4 data sets, burn, bmt, GBSG2 and colon is estimated for both the methods as shown in Figure 4.5. OSTE discards harmful trees from the forest that might have the effects of non-informative features thus giving larger importance values to predictive features compared to random survival forest as shown for the burn and bmt data sets (top panel of Figure 4.5). OSTE fails to achieve this in the cases of colon and GBSB2 data sets (bottom panel of Figure 4.5) which might be a reason of OSTE outperformed by

RSF in the cases of these data sets.

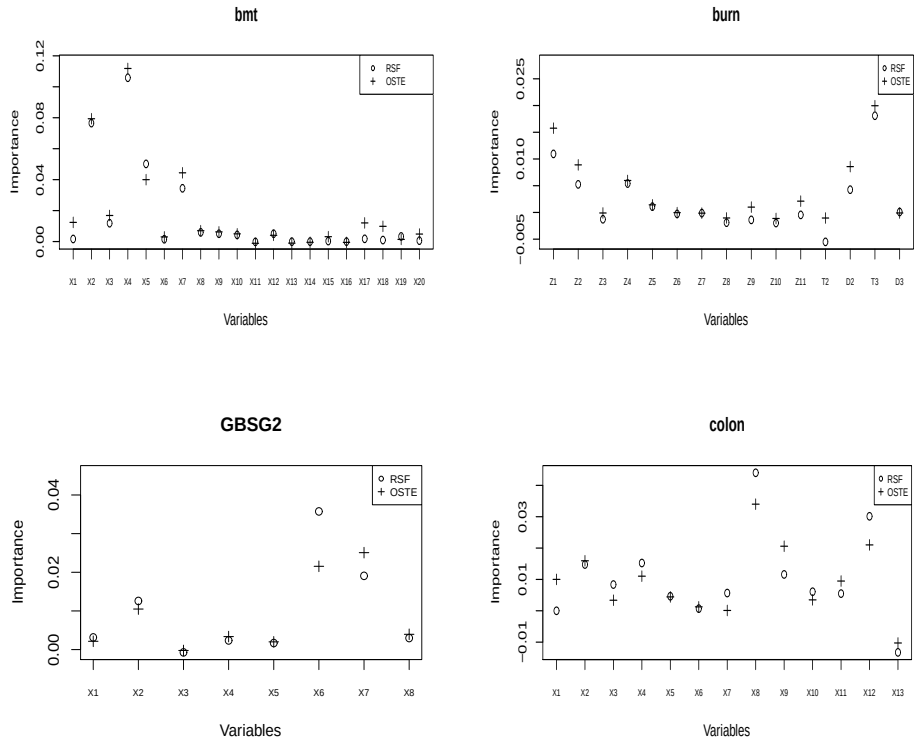

Figure 4.5: *The plots showing feature importance for RSF and OSTE. The dots and + sign shows RSF and OSTE respectively.*

4.4.1 Hyper-parameters assessment

The effect of various number of trees (B) grown in the initial ensemble, proportion of trees (M) selected based on individual accuracy and the number of features p have been assessed on the results of OSTE. For assessing the effect of B, various values are tried in the initial set. The results are given in Figure 4.6. It can be seen that increasing the number of trees from 1000 has no/little effect on the Brier scores on the given datasets. For kidtran dataset, growing more than 1000 trees increase the error.

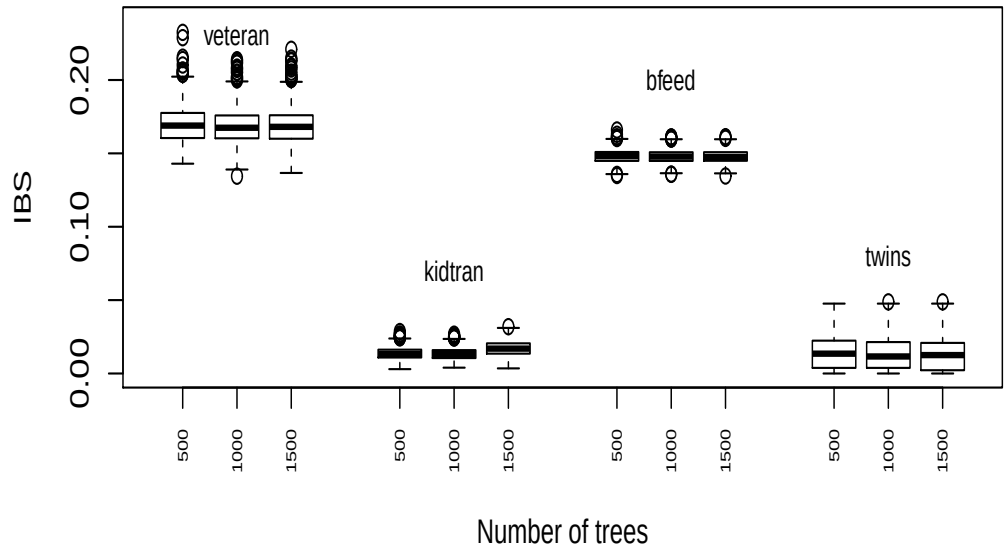

Figure 4.6: *The boxplot showing a comparison of IBS on four datasets for different number of trees B in the inital set.*

OSTE is also checked for various values of M i.e. $5\%, 10\%, \ldots, 60\%$. The results are shown in Figure 4.7. As can be seen in the figure that OSTE gives almost same results by only selecting 5% of trees from the total initial set based on individual accuracy against higher values of *M*. This has led to a final ensemble of sizes 25, 23, 31 and 24 for veteran, kidtran, bfeed and twins datasets respectively. This reveals that a significant reduction in the number of trees used for the final ensemble can be achieved by using OSTE.

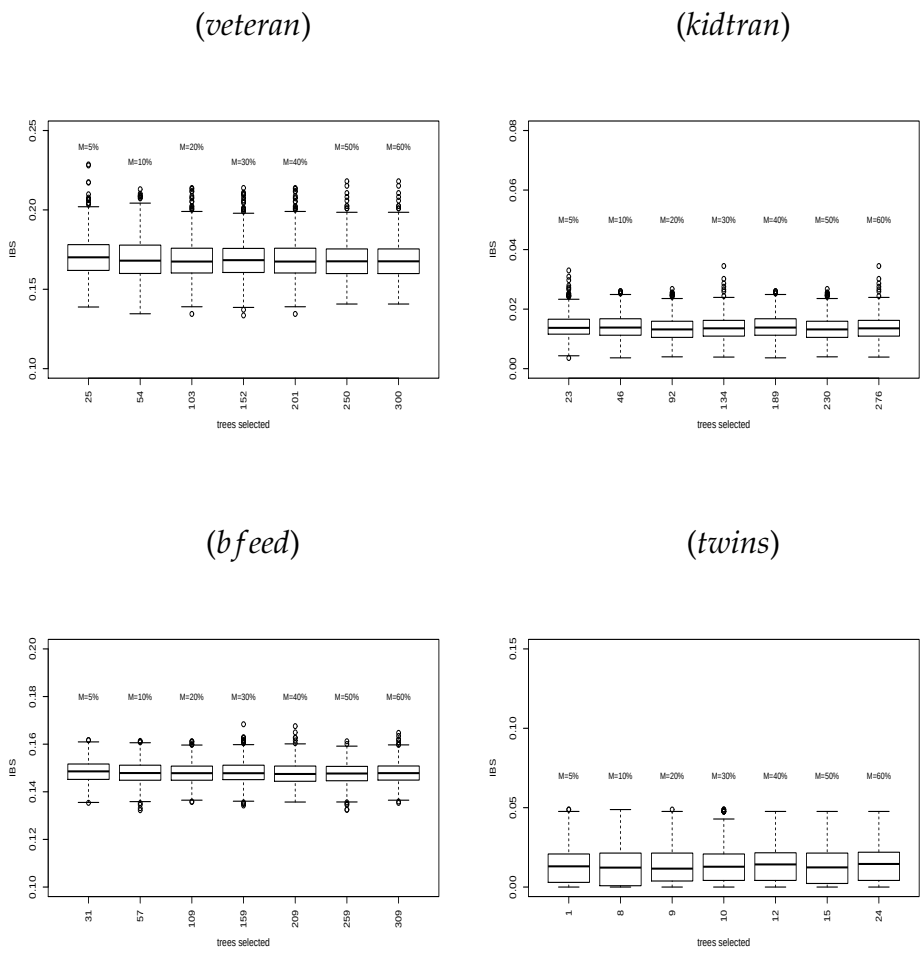

Figure 4.7: *The boxplot showing a comparison of IBS on the datasets for different percentages of total number of trees (M) selected in the first phase. The trees selected by OSTE for the final ensemble are given on the x-axis.*

The effect of the number of features selected at random for splitting the nodes of the trees on IBS are shown in Figure 4.8. The figure shows that there are variations in the results for changing value of p. This suggest that this parameter may be tuned for the corresponding data set.

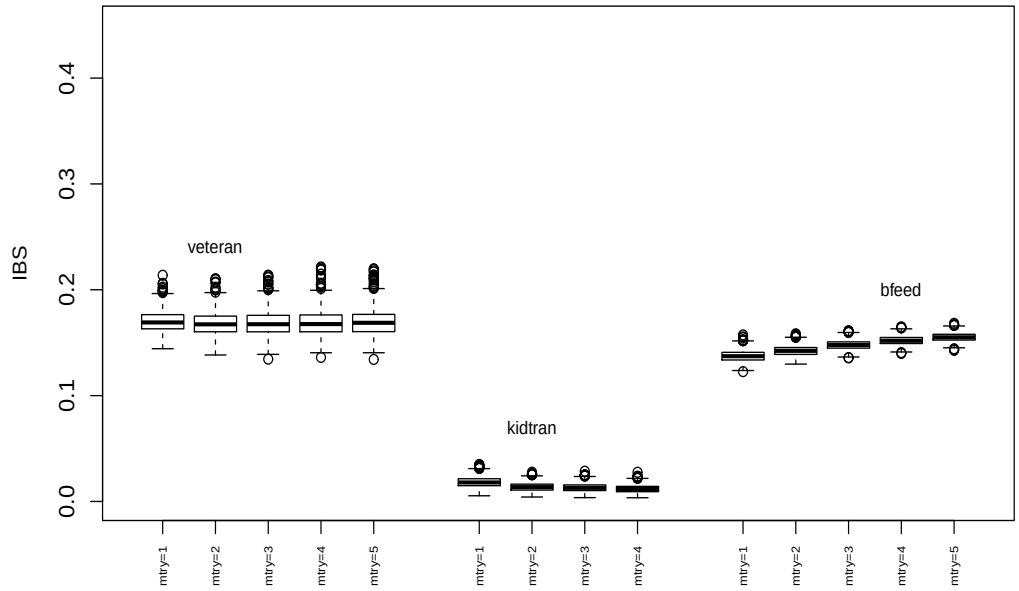

Figure 4.8: *Boxplots showing a comparison of IBS on veteran, kidtran and bfeed datasets for different values of p.*

4.4.2 Size comparison

A comparative analysis of ensemble sizes in terms of the number of survival trees used has also been done. The number of survival trees used in the final ensemble by the methods are given in Table 4.2. The table shows that by choosing $M = 20\%$, a comparable performance could be achieved by a total of 103, 92, 109, 1, 87, 35, 99, 104, 95, 102, 105, 203, 109, 97, 34, 46, 39 and 51 trees for veteran-Pbc datasets, receptively as compared to the other methods using hundreds of survival trees in the corresponding final ensembles. This might be very helpful in reducing computational cost of the ensemble in terms of storage resources.

Table 4.2: *Table showing sizes of ensemble for the datasets. Size of OSTE is shown for M = 20%.*

Dataset	Bagging	RSF	CIF	OSTE
veteran	1000	1500	1000	103
kidtran	1500	1000	1000	92
bfeed	500	1000	500	109
twins	1000	1000	1500	1
VA	1000	1000	1500	87
BMT	1000	1000	1000	35
retinophty	1000	1000	1000	99
cgd	1000	1000	1500	104
channing	1000	1500	1000	95
Burn	1500	1000	1500	102
GBSG2	1500	1500	1500	105
Cost	1500	1000	1000	203
myeliod	1000	1000	1500	109
NKI	1000	1500	1000	97
colon	1500	1500	1500	34
Hodg	1500	1000	1000	46
Kidney	1000	1500	1000	39
Pbc	1000	1000	1500	51

4.5 Chapter summary

This chapter has given the proposed ensemble of optimal survival trees. The metric used in the ensemble formation has also been briefly described. Experiments on 17 benchmark datasets are given using Cox model, bagging, RSF, CIF and OSTE. Performance of the methods is shown by calculating integrated Brier scores on all the datasets. The effect of various hyper-parameters on OSTE has also been checked. Ensemble size comparison of the methods is also done.

Chapter 5

Conclusion

This thesis has aimed at reducing the number of survival trees in the forest in addition to improving its performance. Hence, the idea of "optimal survival trees ensemble" OSTE, is proposed to achieve this goal. Out-of-bag (OOB) observations are used from the bootstrap samples taken from training data as the test subjects to find trees that showed better performance based on C-index. The top ranked survival trees are then assessed on an independent training data for ensemble predictive accuracy. Survival trees that performed well both individually and collectively were selected for the final ensemble. OSTE is then applied on 17 datasets and the results, in terms of integrated Brier score, are compared with some stat-of-the-art method i.e. Cox proportional hazard model, random survival forest, conditional inference forest and bagging survival trees.

Average integrated Brier scores are calculated and Boxplots have been constructed from the integrated Brier scores after applying all the methods on the benchmark data sets. It has been observed that the proposed OSTE is giving better/comparable results to the best of the other methods.

In addition to improved predictive performance, OSTE has also been observed to significantly reduce the number of survival trees in the final ensemble. OSTE consisting of less than 20 survival trees is seen to give comparable results to ensembles of hundreds of survival trees.

Furthermore, the effect of various hyper-parameters on the performance of OSTE has also been checked. In this regard, the effect of changing the number p of features selected at the nodes of the survival trees, proportion M of the top ranked trees and number of trees in the initial ensemble have been checked. p is thus considered to be a tuning parameter of the methods and shall be fine tuned for a data set accordingly. M needs to be no more than 20% as higher values only increase the size of the ensemble with no improvements. As an initial set, a total of 1000 survival trees are suggested to be grown for better results.

The proposed ensemble is implemented in an *R* package "OSTE".

Optimal survival trees ensemble might be used in future to reduce the number of survival trees in the final ensemble to a level that can be interpreted. However, due to the additional filters used in survival trees selection, OSTE is more complex compared to the others and will consequently need more training time. Parallel computing [72] in R could be used in high dimensional settings to reduce training time of the proposed method.

As the proposed OSTE leaves some observations from the training data for internal validation purpose, therefore, some important information could be lost in the learning process, where as the rest of the methods learn from the entire training set. This might negatively affect the performance of OSTE. To solve this problem and further improve OSTE as a future direction, out-of-bag observations could be used for internal validation as well.

Another way to further improve the proposed method is to look for alternative statistics instead of log-rank test while growing survival tree in that this test favours splitting on variables with many possible split points. Maximally selected rank statistic [47] could serve as an off-the-shelve tool for split point selection while growing trees for OSTE.

To prepare the proposed method to work well in high dimensional settings, some state-of-the-art feature selection/dimensionality reduction techniques could be used with OSTE.

Bibliography

[1] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and regression trees*. CRC press, 1984.

[2] H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer, "Random survival forests," *The annals of applied statistics*, pp. 841–860, 2008.

[3] T. Hothorn, B. Lausen, A. Benner, and M. Radespiel-Tröger, "Bagging survival trees," *Statistics in medicine*, vol. 23, no. 1, pp. 77–91, 2004.

[4] L. Breiman, "Bagging predictors," *Machine learning*, vol. 24, no. 2, pp. 123–140, 1996.

[5] I. Bou-Hamad, D. Larocque, H. Ben-Ameur, *et al.*, "A review of survival trees," *Statistics Surveys*, vol. 5, pp. 44–71, 2011.

[6] Z. Khan, *Ensemble of Selected Trees for Classification and Regression*. PhD thesis, University of Essex, 2015.

[7] Z. Khan, A. Gul, O. Mahmoud, M. Miftahuddin, A. Perperoglou, W. Adler, and B. Lausen, "An ensemble of optimal trees for class membership probability estimation," in *Analysis of Large and Complex Data*, pp. 395–409, Springer, 2016.

[8] M. Stevenson and I. EpiCentre, "An introduction to survival analysis," *EpiCentre, IVABS, Massey University*, 2009.

[9] J. Beyersmann, A. Allignol, and M. Schumacher, *Competing risks and multistate models with R*. Springer Science & Business Media, 2011.

[10] J. P. Klein and M. L. Moeschberger, *Survival analysis: techniques for censored and truncated data*. Springer Science & Business Media, 2 005.

[11] J. Beyersmann and T. H. Scheike, "Classical regression models for competing risks," in *Handbook of survival analysis, handbooks of modern statistical methods*, ch. 8, pp. 157–177, Chapman and Hall/CRC, Boca Raton, 2013.

[12] E. T. Lee and O. T. Go, "Survival analysis in public health research," *Annual review of public health*, vol. 18, no. 1, pp. 105–134, 1997.

[13] R. Singh and K. Mukhopadhyay, "Survival analysis in clinical trials: Basics and must know areas," *Perspectives in clinical research*, vol. 2, no. 4, p. 145, 2011.

[14] C. Chai-Adisaksopha, A. Iorio, C. Hillis, W. Lim, and M. Crowther, "A systematic review of using and reporting survival analyses in acute lymphoblastic leukemia literature," *BMC hematology*, vol. 16, no. 1, p. 17, 2016.

[15] D. G. Kleinbaum and M. Klein, "Competing risks survival analysis," *Survival Analysis: A self-learning text*, pp. 391–461, 2005.

[16] T. Clark, M. Bradburn, S. Love, and D. Altman, "Survival analysis part i: basic concepts and first analyses," *The British Journal of Cancer*, vol. 89, no. 2, p. 232, 2003.

[17] E. Marubini and M. G. Valsecchi, *Analysing survival data from clinical trials and observational studies*, vol. 15. John Wiley & Sons, 2004.

[18] F. Steele, "Event history analysis: a national centre for research methods briefing paper," *UK: Centre for Multilevel Modelling, University of Bristol*, 2005.

[19] Ø. Borgan, *Three Contributions to the Encyclopedia of Biostatistics: The Nelson-Aalen, Kaplan-Meier, and Aaalen-Johansen Estimators*. Department of Mathematics, University of Oslo, 1997.

[20] L. Gordon and R. A. Olshen, "Tree-structured survival analysis.," *Cancer treatment reports*, vol. 69, no. 10, pp. 1065–1069, 1985.

[21] M. LeBlanc and J. Crowley, "Survival trees by goodness of split," *Journal of the American Statistical Association*, vol. 88, no. 422, pp. 457–467, 1993.

[22] D. G. Altman, *Practical statistics for medical research*. CRC press, 1990.

[23] B. Damato and A. Taktak, "Survival after treatment of intraocular melanoma," in *Outcome Prediction in Cancer*, pp. 27–41, Elsevier, 2007.

[24] J. Borucka, "Extensions of cox model for non-proportional hazards purpose," *Ekonometria*, no. 3 (45), pp. 85–101, 2014.

[25] P. Meier and E. Kaplan, "Nonparametric estimation from incomplete observations," *Journal of the American Statistical Association*, vol. 53, no. 282, pp. 457–481, 1958.

[26] P. WANG, Y. LI, and C. K. REDDY, "Machine learning for survival analysis: A survey,"

[27] C. David, "Regression models and life tables (with discussion)," *Journal of the Royal Statistical Society*, vol. 34, pp. 187–220, 1972.

[28] T. Hothorn, P. Bühlmann, S. Dudoit, A. Molinaro, and M. J. Van Der Laan, "Survival ensembles," *Biostatistics*, vol. 7, no. 3, pp. 355–373, 2005.

[29] M. N. Wright, T. Dankowski, and A. Ziegler, "Unbiased split variable selection for random survival forests using maximally selected rank statistics," *Statistics in medicine*, vol. 36, no. 8, pp. 1272–1284, 2017.

[30] A. Nardi and M. Schemper, "Comparing cox and parametric models in clinical studies," *Statistics in medicine*, vol. 22, no. 23, pp. 3597–3610, 2003.

[31] R. Zhu, *Tree-based methods for survival analysis and high-dimensional data*. PhD thesis, The University of North Carolina at Chapel Hill, 2013.

[32] Z. Huang, H. Zhang, J. Boss, S. A. Goutman, B. Mukherjee, I. D. Dinov, Y. Guan, *et al.*, "Complete hazard ranking to analyze right-censored data: An als survival study," *PLoS computational biology*, vol. 13, no. 12, p. e1005887, 2017.

[33] F. M. Callaghan, *Classification trees for survival data with competing risks*. PhD thesis, University of Pittsburgh, 2008.

[34] B. Lausen, W. Sauerbrei, and M. Schumacher, "Classification and regression trees (cart) used for the exploration of prognostic factors measured on different scales," in *Computational Statistics:Papers Collected on the Occasion of the 25th Conference on Statistical Computing at Schloss Reisensburg*, pp. 483–496, Physica Verlag, 1994.

[35] A. Ciampi, R. Bush, M. Gospodarowicz, and J. Till, "An approach to classifying prognostic factors related to survival experience for non-hodgkin's lymphoma patients: Based on a series of 982 patients: 1967–1975," *Cancer*, vol. 47, no. 3, pp. 621–627, 1981.

[36] P. Jadwiga Borucka, "Extensions of cox model for non-proportional hazards purpose,"

[37] T. Hothorn, P. Bühlmann, S. Dudoit, A. Molinaro, and M. Van Der Laan, "Survival ensembles.," *Biostatistics (Oxford, England)*, vol. 7, no. 3, p. 355, 2006.

[38] T. G. Dietterich *et al.*, "Ensemble methods in machine learning," *Multiple classifier systems*, vol. 1857, pp. 1–15, 2000.

[39] T. G. Dietterich, "Ensemble methods in machine learning," in *International workshop on multiple classifier systems*, pp. 1–15, Springer, 2000.

[40] F. Dannegger, "Tree stability diagnostics and some remedies for instability," *Statistics in medicine*, vol. 19, no. 4, pp. 475–491, 2000.

[41] C. Mbogning and P. Broët, "Bagging survival tree procedure for variable selection and prediction in the presence of nonsusceptible patients," *BMC bioinformatics*, vol. 17, no. 1, p. 230, 2016.

[42] E. Bingham and H. Mannila, "Random projection in dimensionality reduction: applications to image and text data," in *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 245–250, ACM, 2001.

[43] J. B. Nasejje, H. Mwambi, K. Dheda, and M. Lesosky, "A comparison of the conditional inference survival forest model to random survival forests based on a simulation

study as well as on two applications with time-to-event data," *BMC medical research methodology*, vol. 17, no. 1, p. 115, 2017.

[44] A. Ziegler and I. R. König, "Mining data with random forests: current options for real-world applications," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 4, no. 1, pp. 55–63, 2014.

[45] H. David, *Linear Rank Tests in Survival Analysis: Encyclopedia of Biostatistics*. Wiley Interscience, 2005.

[46] C. Strobl, A.-L. Boulesteix, A. Zeileis, and T. Hothorn, "Bias in random forest variable importance measures: Illustrations, sources and a solution," *BMC bioinformatics*, vol. 8, no. 1, p. 25, 2007.

[47] B. Lausen and M. Schumacher, "Maximally selected rank statistics," *Biometrics*, pp. 73–85, 1992.

[48] T. Hothorn and A. Zeileis, "Generalized maximally selected statistics," *Biometrics*, vol. 64, no. 4, pp. 1263–1269, 2008.

[49] T. Hothorn and B. Lausen, "On the exact distribution of maximally selected rank statistics," *Computational Statistics & Data Analysis*, vol. 43, no. 2, pp. 121–137, 2003.

[50] K. Bache and M. Lichman, "Uci machine learning repository," 2013.

[51] Klein, Moeschberger, and J. Yan, *KMsurv: Data sets from Klein and Moeschberger (1997), Survival Analysis*, 2012. R package version 0.1-5.

[52] T. Therneau, *A Package for Survival Analysis*, 2015. R package 2.38.

[53] U. B. Mogensen, H. Ishwaran, and T. A. Gerds, "Evaluating random forests for survival analysis using prediction error curves," *Journal of Statistical Software*, vol. 50, no. 11, pp. 1–23, 2012.

[54] J. D. Kalbfleisch and R. L. Prentice, *The statistical analysis of failure time data*, vol. 360. John Wiley & Sons, 2011.

[55] T. R. Fleming and D. P. Harrington, *Counting processes and survival analysis*, vol. 169. John Wiley & Sons, 2011.

[56] W. Sauerbrei and P. Royston, "Building multivariable prognostic and diagnostic models: transformation of the predictors by using fractional polynomials," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, vol. 162, no. 1, pp. 71–94, 1999.

[57] U. B. Mogensen, H. Ishwaran, and T. A. Gerds, "Evaluating random forests for survival analysis using prediction error curves," *Journal of statistical software*, vol. 50, no. 11, p. 1, 2012.

[58] R Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016.

[59] M. N. Wright and A. Ziegler, "ranger: A fast implementation of random forests for high dimensional data in C++ and R," *Journal of Statistical Software*, vol. 77, no. 1, pp. 1–17, 2017.

[60] A. Liaw and M. Wiener, "Classification and regression by randomforest," *R News*, vol. 2, no. 3, pp. 18–22, 2002.

[61] A. Peters and T. Hothorn, *ipred: Improved Predictors*, 2015. R package version 0.9-5.

[62] T. Therneau, "Survival: A package for survival analysis. version 2.38," 2015.

[63] B. Efron, "The efficiency of cox's likelihood function for censored data," *Journal of the American statistical Association*, vol. 72, no. 359, pp. 557–565, 1977.

[64] G. W. Brier, "Verification of forecasts expressed in terms of probability," *Monthey Weather Review*, vol. 78, no. 1, pp. 1–3, 1950.

[65] S. Tsouprou, H. Putter, and M. Fiocco, "Measures of discrimination and predictive accuracy for interval censored survival data," 2015.

[66] J. M. Robins and D. M. Finkelstein, "Correcting for noncompliance and dependent censoring in an aids clinical trial with inverse probability of censoring weighted (ipcw) log-rank tests," *Biometrics*, vol. 56, no. 3, pp. 779–788, 2000.

[67] M. S. Rahman, *Validation measures for prognostic models for independent and correlated binary and survival outcomes*. PhD thesis, UCL (University College London), 2012.

[68] R. D'Agostino, J. Griffith, C. Schmid, and N. Terrin, "Measures for evaluating model performance," in *Proceedings of American statistical association biometrics section*, pp. 253–258, 1997.

[69] F. E. Harrell, R. M. Califf, D. B. Pryor, K. L. Lee, and R. A. Rosati, "Evaluating the yield of medical tests," *Jama*, vol. 247, no. 18, pp. 2543–2546, 1982.

[70] A. Mayr and M. Schmid, "Boosting the concordance index for survival data–a unified framework to derive and evaluate biomarker combinations," *PloS one*, vol. 9, no. 1, p. e84483, 2014.

[71] K. K. Nicodemus, J. D. Malley, C. Strobl, and A. Ziegler, "The behaviour of random forest permutation-based variable importance measures under predictor correlation," *BMC bioinformatics*, vol. 11, no. 1, p. 110, 2010.

[72] R Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017.

Appendix A

R-Package

This package consists of function for growing survival trees ensemble, that are grown by the method of random survival forest. The survival trees grown are assessed for both individual and collective performances. The ensemble can give promising results on fewer survival trees selected based on their individual and collective performance in the final ensemble.

Package 'OSTE'

September 30, 2017

Type Package

Title Optimal survival trees ensemble

Version 1.0

Date 2017-09-23

Author Naz Gul, Nosheen Faiz, Zardad Khan and Berthold Lausen

Maintainer Naz Gul <ngul@essex.ac.uk>

Description This package consiste of function for growing survival trees ensemble,
that are grown by the method of random survival forest.
The survival trees grown are assessed for both individual and collective performances.
The ensemble can give promising results on fewer survival trees selected in the final ensemble.

Depends ranger,pec,stats,survival,prodlim

LazyLoad yes

License GPL (>= 2)

R topics documented:

`comb.ranger` *Combining ranger objects for survival analysis*

Description

This function combines two or more than two ranger objects for survival analysis

Usage

```
comb.ranger(...)
```

Arguments

...	ojects of class ranger for survival analysis

Author(s)

Naz Gul, Nosheen Faiz, Zardad Khan and Berthold Lausen.

References

Marvin N. Wright, Andreas Ziegler (2017). ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. Journal of Statistical Software, 77(1), 1-17. doi:10.18637/jss.v077.i01

See Also

OSTE

OSTE *Optimal Survival Trees Ensembles*

Description

This package consists of function for growing survival trees ensemble, that are grown by the method of random survival forest. The survival trees grown are assessed for both individual and collective performances. The ensemble can give promising results on fewer survival trees selected based on their individual and collective performance in the final ensemble.

Details

Package:	OSTE
Type:	Package
Version:	1.0
Date:	2017-09-28
License:	GPL (>= 2)

Author(s)

Naz Gul, Nosheen Faiz, Zardad Khan and Berthold Lausen.

Maintainer: Naz Gul <ngul@essex.ac.uk>

References

Gul, N., Faiz, N., Khan, Z. and Lausen, B.(2017) “Optimal survival trees ensemble”. Journal name to appear

OSTE *Optimal survival trees ensemble*

Description

Optimal survival trees ensemble is the main function of OSTE package that grows a sufficiently large number, t.initial, of survival trees and selects optimal survival trees from the total trees grown by random survival forest. Number of survival trees in the initial set, t.initial, is chosen by the user. If not chosen, then the default t.initial = 500 is used. Based on empirical investigation, t.initial =1000 is recommended.

Usage

```
OSTE(formula = NULL, data, t.initial = NULL, v.size = NULL, mtry = NULL,
      M = NULL,  minimum.node.size = NULL, always.split.features = NULL,
      replace = TRUE, splitting.rule = NULL, info = TRUE)
```

Arguments

formula	Object of class formula describing the required model to be fitted. Interaction terms are not supported in the current version.
data	A nxd matrix or data frame of n observations on d features along with response variables that are described by the formula.
t.initial	Number of survival trees to be grown initially. If equal to NULL then the defalut of t.initial = 500 is taken. A recommended value is t.initial = 1000.
v.size	Portion of data used for validation in the second phase i.e. for assessing survival trees performance in the ensemble. If equal to NULL then the defalut v.size=0.1
mtry	Number of features selected at random at each node of the survival trees for splitting. If equal to NULL then the default sqrt(d) is taken.
M	Percent of the best t.initial survival trees to be selected on the basis of their performance on out-of-bag observations. For selecting 20% of trees, take M=0.2.
minimum.node.size	Minimal node size. If equal to NULL then the default minimum.node.size = 3 is executed.
always.split.features	Vector of variable names if desired to be always selected in addition to the mtry variables tried for splitting.
replace	Whether sampling should be done with or without replacement.
splitting.rule	Splitting rule."logrank", "C" or "maxstat" are suported with default "logrank".
info	If TRUE, displays process status .

Details

Large values are recommended for t.initial for better performance as possible under the available computational resources. The log-rank test statistic is used as defalut, A C-index based splitting rule (Schmid et al. 2015) and maximally selected rank statistics (Wright et al. 2016) are available. The C-index shows better predictive performance in case of high censoring rate, where logrank is best for situations where the data are noisy (Schmid et al. 2015).

Value

`unique.death.times`	Unique death times.
`CHF`	Estimated cumulative hazard function for each observation.
`Survival_Prob`	Estimated survival probability for each observation.
`trees_selected`	Number of trees selected.
`mtry`	Value of mtry used.
`forest`	Saved forest for prediction purposes.

Note

In the case of missing values in any dataset prior action needs to be taken as the fuction can not handle them at the current version. Moreover, the status/delta variable in the data must be code as 0, 1.

Author(s)

Naz Gul, Nosheen Faiz, Zardad Khan and Berthold Lausen.

References

Marvin N. Wright, Andreas Ziegler (2017). ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. Journal of Statistical Software, 77(1), 1-17. doi:10.18637/jss.v077.i01

Terry Therneau, Beth Atkinson and Brian Ripley (2015) rpart: Recursive Partitioning and Regression Trees. R package version 4.1-10. https://CRAN.R-project.org/package=rpart

Ulla B. Mogensen, Hemant Ishwaran, Thomas A. Gerds (2012). Evaluating Random Forests for Survival Analysis Using Prediction Error Curves. Journal of Statistical Software, 50(11), 1-23. URL http://www.jstatsoft.org/v50/i11/.

Schmid, M., Wright, M. N. & Ziegler, A. (2016). On the use of Harrell's C for clinical risk prediction via random survival forests. Expert Syst Appl 63:450-459. http://dx.doi.org/10.1016/j.eswa.2016.07.018.

Wright, M. N., Dankowski, T. & Ziegler, A. (2017). Unbiased split variable selection for random survival forests using maximally selected rank statistics. Stat Med. http://dx.doi.org/10.1002/sim.7212.

Zardad Khan, Asma Gul, Aris Perperoglou, Osama Mahmoud, Werner Adler, Miftahuddin and Berthold Lausen (2015). OTE: Optimal Trees Ensembles for Regression, Classification and Class Membership Probability Estimation. R package version 1.0. https://CRAN.R-project.org/package=OTE

See Also

`predict.OSTE`

Examples

```
#Load the data
data(VETERAN)

#Divide the data into training and test parts

n <- nrow(VETERAN)
trainind <- sample(1:n,n*0.7)
testind <- (1:n)[-trainind]
```

```
# Grow OSTE on the training data

OSTE.fit <- OSTE(Surv(time,status)~.,data=VETERAN[trainind,],t.initial=100)

# Predict on the test data

pred <- predict.OSTE(OSTE.fit,newdata=VETERAN[testind,])

# Index various values

pred$survival_prob

#etc.
```

---

predict.OSTE *Prediction function for* OSTE *object*

Description

This function provides prediction for test data on the trained OSTE object for survival analysis.

Usage

```
predict.OSTE(object, newdata)
```

Arguments

object	An OSTE object.
newdata	New/test data.

Value

CHF	A vector of cumulative hazard function of training data.
survival_prob	A vector of survivalprobability of testing data.
time_points	A vector of unique death times.

Author(s)

Naz Gul, Nosheen Faiz, Zardad Khan and Berthold Lausen.

References

Marvin N. Wright, Andreas Ziegler (2017). ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. Journal of Statistical Software, 77(1), 1-17. doi:10.18637/jss.v077.i01

See Also

OSTE

Examples

```
#Load the data
data(VETERAN)


#Divide the data into training and test parts

n <- nrow(VETERAN)
trainind <- sample(1:n,n*0.7)
testind <- (1:n)[-trainind]

# Grow OSTE on the training data

OSTE.fit <- OSTE(Surv(time,status)~.,data=VETERAN[trainind,])

# Predict on the test data

pred <- predict.OSTE(OSTE.fit,newdata=VETERAN[testind,])

# Index various values

pred$survival_prob
```

`predictSurvProb.ranger` *Survival probabilities for working with* pec *R package*

Description

This function facilitates pec R package to work with the OSTE package.

Author(s)

Naz Gul, Nosheen Faiz, Zardad Khan and Berthold Lausen.

References

Ulla B. Mogensen, Hemant Ishwaran, Thomas A. Gerds (2012). Evaluating Random Forests for Survival Analysis Using Prediction Error Curves. Journal of Statistical Software, 50(11), 1-23. URL http://www.jstatsoft.org/v50/i11/.

See Also

OSTE

VETERAN *Data on randomized trial of two treatment procedures for lung cancer.*

Description

The data set consist of a total 137 observations on 8 variables. The variables consist of the type of lung cancer treatment i.e 1 (standard) and 2 (test drug), cell Type, Status, that denotes the status of the patient as 1 (dead) or 0 (alive), survival time in days since the treatment, Diag, the time since diagnosis in months, age in years, the Karnofsky score, therapy that denotes any prior therapy 0 (none), 1 (yes).

Usage

```
data("VETERAN")
```

Format

A data frame with 137 observations on the following 8 variables.

`trt:` a numeric vector denoting type of lung cancer treatment i.e 1 (standard) and 2 (test drug).

`celltype:` a factor with levels `squamous`, `smallcell`, `adeno` and `large`.

`time:` a numeric vector denoting survival time in days since the treatment.

`status:` a numeric vector that denotes the status of the patient as 1 (dead) or 0 (alive).

`karno:` a numeric vector denoting the Karnofsky score.

`diagtime:` a numeric vector denoting the time since diagnosis in months.

`age:` age in years.

`prior:` a numeric vector denoting prior therapy; 0 (none), 1 (yes).

References

Therneau T (2015). A Package for Survival Analysis in S. version 2.38, <URL: https://CRAN.R-project.org/package=survival>.

Terry M. Therneau and Patricia M. Grambsch (2000). Modeling Survival Data: Extending the Cox Model. Springer, New York. ISBN 0-387-98784-3

Examples

```
#To load the data
data(VETERAN)
# To see the structure
str(VETERAN)
#etc.
```

Index