Comparison of Tree-Based Ensembles in Application to Censored Data

Malgorzata Kretowska

Faculty of Computer Science
Bialystok University of Technology
Wiejska 45a, 15-351 Bialystok, Poland
e-mail: m.kretowska@pb.edu.pl

Abstract. In the paper the comparison of ensemble based methods applied to censored survival data was conducted. Bagging survival trees, dipolar survival tree ensemble and random forest were taken into consideration. The prediction ability was evaluated by the integrated Brier score, the prediction measure developed for survival data. Two real datasets with different percentage of censored observations were examined.

Keywords: survival tree, ensemble, random forest, censored observation, survival analysis, Brier score.

1 Introduction

Methods for analysis of classification and regression problems are developing to provide faster, more stable and more accurate prediction. The same goal inspires also the researchers working on survival data. Very often new approaches for classification or regression tasks are then adapted to data with incomplete information. Such incomplete information is an integral part of censored data, which contains observations with unknown failure times. For such data we only know how long the observation has not experienced any failure, but the exact failure time remains unknown.

Except statistical methods, which often require many strict assumptions, survival trees and survival ensembles belong to the most common non-parametric methods for survival data analysis. The fast development of survival trees started in the mid-1980s and lasted for the next ten years [3]. The survival ensemble is quite a new branch of analysis of survival data. First methods were proposed in 2004 - bagging survival trees [8] and relative risk forests [10]. The consecutive approaches were proposed by Kretowska [15], Hothorn [9], and Ishwaran [11].

In this paper the comparison of predictive ability of three ensemble methods was conducted. Bagging survival trees [8] and random survival forest [11] are implemented and available in R packages, while dipolar survival tree ensemble [15] was implemented by the author in C++. In order to compare the predictive ability of the models, the integrated Brier score [6] was applied. Experiments were performed on two data sets with different percentage of censored observations. The first data, Veteran's Administration (*VA*) lung cancer study [4], contains

L. Rutkowski et al. (Eds.): ICAISC 2014, Part I, LNAI 8467, pp. 551–560, 2014.

6.5 percent of censored observations, while the other one - malignant melanoma [1] - 72 percent.

The paper consists of six sections. In Section 2 the definition of survival data as well as the survival time distribution functions are presented. Section 3 contains introduction to survival ensemble and more detailed description of three distinguishes ensemble methods. The definition of the integrated Brier score is given in Section 4. Experimental results are presented in Section 5, while Section 6 summarizes the results .

2 Censored Data

Let T^0 denotes the true survival time and C denotes the true censoring time with distribution functions F and G respectively. We observe a random variable $O = (T, \Delta, \mathbf{X})$, where $T = \min(T^0, C)$ is the time to event, $\Delta = I(T \leq C)$ is a censoring indicator and $\mathbf{X} = (X_1, ..., X_N)$ denotes the set of N covariates from a sample space χ. We have a learning sample $L = (\mathbf{x}_i, t_i, \delta_i)$, $i = 1, 2, ..., n$, where $\mathbf{x}_i$ is N-dimensional covariates vector, t_i - survival time and δ_i - failure indicator, which is equal to 0 for censored cases and 1 for uncensored cases.

The distribution of survival time may be described by several functions:

- survival function
$$S(t) = P(T > t) \tag{1}$$
where $P(\bullet)$ means probability, $S(0) = 1$ and $\lim_{t\to\infty} S(t) = 0$
- density function
$$f(t) = \lim_{\triangle t \to 0} \frac{P(t \leq T < t + \triangle t)}{\triangle t} \tag{2}$$
where $f(t)dt$ is the unconditional probability of failure in the infinitesimal interval $(t, t + dt)$.
- hazard function
$$\lambda(t) = \lim_{\triangle t \to 0} \frac{P(t \leq T < t + \triangle t | T \geq t)}{\triangle t} \tag{3}$$
where $\lambda(t)dt$ is the probability of failure in the in infinitesimal interval $(t, t + dt)$, given survival at time t.
- cumulative hazard function
$$\Lambda(t) = \int_0^t \lambda(u)du = -\log S(t) \tag{4}$$

The estimation of survival function $S(t)$ may be done by using the Kaplan-Meier product limit estimator [13], which is calculated on the base of the learning sample L and is denoted by $\hat{S}(t)$:

$$\hat{S}(t) = \prod_{j|t_{(j)} \leq t} \left(\frac{m_j - d_j}{m_j} \right) \tag{5}$$

where $t_{(1)} < t_{(2)} < \ldots < t_{(D)}$ are distinct, ordered survival times from the learning sample L, in which the event of interest occurred, d_j is the number of events at time $t_{(j)}$ and m_j is the number of patients at risk at $t_{(j)}$ (i.e., the number of patients who are alive at $t_{(j)}$ or experience the event of interest at $t_{(j)}$).

The Nelson-Aalen estimator of cumulative hazard function is defined as:

$$H(t) = \sum_{j|t_{(j)} \leq t} \frac{d_j}{m_j} \tag{6}$$

The 'patients specific' survival probability function is given by $S(t|\mathbf{x}) = P(T > t|\mathbf{X} = \mathbf{x})$. The conditional survival probability function for the new patient with covariates vector $\mathbf{x}_{new}$ is denoted by $\hat{S}(t|\mathbf{x}_{new})$. Similarly $H(t|\mathbf{x}_{new})$ means a conditional cumulative hazard function.

3 Ensembles of Survival Trees

An ensemble is a set of k single predictors, often trees. Depending on the data, the ensemble may solve classification, regression or survival problems. In case of censored survival data single predictors are usually survival trees, which have the ability to cope with censored observations. Unlike the ensemble for classification and regression problems, the ensemble of survival trees does not return the exact predicted value. The outcome for a given observation is a distribution function of survival time. Thus, analyzing such a function, the time intervals with higher and smaller probability of failure occurrence may be distinguished for the observation.

Each single tree is built on the base of bootstrap sample drawing with replacement from the learning data. A general algorithm of building and using the ensemble is given as follows:

1. Draw k bootstrap samples $(L_1, L_2, \ldots, L_k)$ of size n with replacement from L
2. Induct k single trees T_i based on each bootstrap sample L_i, $i = 1, 2, \ldots, k$
3. Having a new observation $\mathbf{x}_{new}$, drop it down each of k single trees
4. On the base of the results of k single trees, calculate a function $f(t|\mathbf{x}_{new})$, being an outcome of the whole ensemble

Comparing various approaches to building the ensembles, the differences are visible in steps 2 and 4 of the above algorithm.

3.1 Bagging Survival Trees

The approach was proposed by Hothorn *et al.* [8]. The authors did not focus on special splitting criterion for single tree induction. They used a method previously proposed by LeBlanc and Crowley [16] which employed a measure based on Poisson deviance residuals. They presented an original method of calculating the function $f(t|\mathbf{x}_{new})$, which takes a form of aggregated Kaplan-Meier survival function: $\hat{S}_A(t|\mathbf{x}_{new})$. Step 4 is here divided into two parts:

4a Build aggregated sample $L_A(\mathbf{x}_{new}) = \{L_1(\mathbf{x}_{new}); L_2(\mathbf{x}_{new}), \ldots, L_k(\mathbf{x}_{new})\}$, where $L_i(\mathbf{x}_{new})$ is a set of observations from the bootstrap sample L_i that reached the same leaf node of the tree T_i as the observation $\mathbf{x}_{new}$.

4b On the base of aggregated sample $L_A(\mathbf{x}_{new})$, compute the Kaplan-Meier aggregated survival function for a new observation $\mathbf{x}_{new}$: $\hat{S}_A(t|\mathbf{x}_{new})$

3.2 Dipolar Survival Trees Ensemble

Unlike the bagging survival tree, which is an example of univariate tree, the single dipolar survival tree [14] belongs to multivariate approaches. It means that each internal node contains the split which is based not only on one variables (e.g $x_i > c$), but a linear combination of input variables is examined. The test takes the form of a hyperplane: $H(\mathbf{w}, \theta) = \{\mathbf{x} : \mathbf{w}^T\mathbf{x} = \theta\}$. If a given feature vector $\mathbf{x}$ is situated on the positive site of the hyperplane the test returns the value greater or equal to 0, in the other case the test returns the negative value. The values of $\mathbf{w}$ and θ are calculated by the minimization of dipolar criterion function [2].

Dipolar survival trees ensemble [15] is build according to the general rules presented above. Similarly to bagging survival trees, the result of the whole ensemble for a new features vector $\mathbf{x}_{new}$ is calculated as an aggregated survival function $\hat{S}_A(t|\mathbf{x}_{new})$.

3.3 Random Survival Forest

Randon survival forest was proposed by Ishwaran *et al.* [11]. The method differs from the previous ones, both in the induction process and in the way the results are calculated. During the induction process the randomization is injected into each node generation. It means that the best split is not chosen by the analysis of the whole set of available variables but a subset of variables is selected. Then, basing on this subset, the split that maximizes survival difference between two child nodes is chosen.

The results of the whole ensemble is calculated as the average of cumulative hazards functions received for each single tree. Step no. 4 is here divided into three parts:

4a For each survival tree T_i, $i = 1, 2, \ldots, k$, determine a set $L_i(\mathbf{x}_{new})$ containing the covariates vectors from the bootstrap sample L_i which belong to the same leaf node as $\mathbf{x}_{new}$

4b For each set $L_i(\mathbf{x}_{new})$ calculate the Nelson-Aalen estimator of CHF: $H_i^*(t|\mathbf{x}_{new})$, $i = 1, 2, \ldots, k$

4c Calculate the average of CHF to obtain the ensemble CHF:

$$H(t|\mathbf{x}_{new}) = \frac{1}{k}\sum_{i=1}^{k} H_i^*(t|\mathbf{x}_{new}) \tag{7}$$

4 Model Validation

In case of censored survival data where the exact failure time for a given subject may be unknown, the classical validation measures used in regression problems are not applicable. Indexes which are used in survival analysis do not calculate the differences between the given and predicted failure times, they rather use the differences between survival functions [6,5] or the order of predicted and given survival times [7]. The integrated Brier score [6] belongs to the first types of indexes. For a fixed time point t the contribution to the Brier score is divided into three groups:

1. $t_i \leq t$ and $\delta_i = 1$
2. $t_i > t$ and ($\delta_i = 1$ or $\delta_i = 0$)
3. $t_i \leq t$ and $\delta_i = 0$

For the observations belonging to group 1 the failure occurred before t and the event status at t is equal to 0, so in the Brier score we present this as $(0 - \hat{S}(t|\mathbf{x}_i))^2 = \hat{S}(t|\mathbf{x}_i)^2$. The observations of group 2 do not experienced any event at time t, hence the event status at t is equal to 1 and the contribution to the Brier score is: $(1 - \hat{S}(t|\mathbf{x}_i))^2$. The contribution to the Brier score for observation of group 3 can not be calculated, because the event status at t is unknown for them. Since the observations of group 3 do not have any contribution to the Brier score, the loss of information should be compensate by additional weighting of the existing contributions. The observations in group 1 have the weight $\hat{G}(t_i)^{-1}$ and those in group 2 the weight $\hat{G}(t)^{-1}$, where $\hat{G}(t)$ denotes the Kaplan-Meier estimator of the censoring distribution. It is calculated on the base of observations $(t_i, 1 - \delta_i)$. The definition of the Brier score is given as:

$$BS(t) = \frac{1}{n} \sum_{i=1}^{N} (\hat{S}(t|\mathbf{x}_i)^2 I(t_i \leq t \wedge \delta_i = 1) \hat{G}(t_i)^{-1} + (1 - \hat{S}(t|\mathbf{x}_i))^2 I(t_i > t) \hat{G}(t)^{-1}) \tag{8}$$

where $I(condition)$ is equal to 1 if the condition is fulfilled, 0 otherwise. The BS equal to 0 means the best prediction.

The integrated Brier score is calculated as:

$$IBS = \frac{1}{max(t_i)} \int_0^{max(t_i)} BS(t)dt \tag{9}$$

5 Experimental Results

The comparison of three ensemble methods in application to censored survival data was conducted. Experimental results were performed on the base of two real data sets with different percentage of censored observations. The value of the integrated Brier score, given in the paper, is the average value of the index calculated for 20 runs of 10-fold cross-validation. The random survival forest (RSF) is implemented in R package 'randomForestSRC' [12]. Since the package

uses Harrell's concordance index [7] as a prediction measure, package 'pred' [17] was used to calculate the integrated Brier score. The second aggregation technique is the bagging survival trees method (*BST*) proposed by Hothorn *et al.* [8], which is implemented in 'ipred' package [18].

The first analyzed dataset contains the information from the Veteran's Administration (*VA*) lung cancer study [4]. In this trial, male patients with advanced inoperable tumors were randomized to either standard (69 subjects) or test chemotherapy (68 subjects). Only 9 subjects from 137 were censored. Detailed description of the variables is given in table 1.

Table 1. Description of *VA lung cancer* data

Variable name	Description
Variables assessed at the time of randomization	
Treat	Chemotherapy (0-standard, 1-test)
Cell	Cell type (0-squamous, 1-small, 2-adeno, 3-large)
Prior	Prior therapy (0-no, 1-yes)
KPS	Karnofsky rating
DiagTime	Disease duration
Age	Age
Outcome variables	
Time	Survival time
Status	Failure indicator (0- censored observation, 1- death)

In table 2 the integrated Brier scores (IBS) for *VA lung cancer* data are presented. The experiments were conducted for the ensembles with different number of single trees: 50, 100, 200, 500, 1000. The results for RSF do not depend on the number of single trees, for 100 trees as well as for 1000 trees the IBS equals 0.104. The best results are for bagging survival trees method, for 1000 trees IBS equals 0.098. The most visible influence of the number of trees is for DST ensemble technique. For 50 trees the IBS equals 0.119, then decreasing with increased number of trees, riches the value 0.104 for 1000 trees, what is comparable with the IBS received for RSF.

Table 2. The integrated Brier scores received for *VA lung cancer* data

Number of trees	RSF	BST	DST Ensemble
50	0.105	0.102	0.119
100	0.104	0.101	0.111
200	0.109	0.101	0.108
500	0.103	0.099	0.105
1000	0.104	0.098	0.104

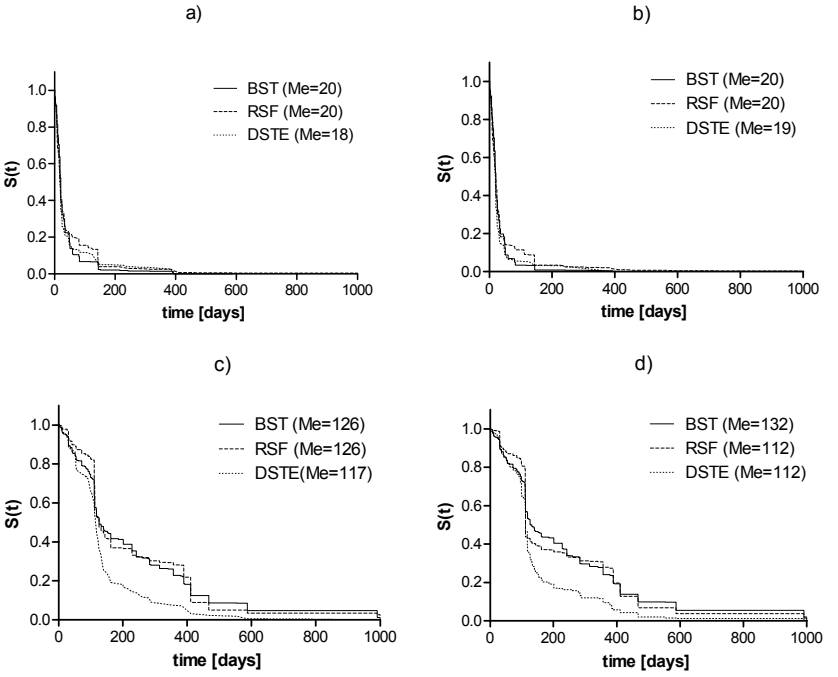

Fig. 1. Survival functions for *VA lung cancer* data a) Treat=0, KPS=20; b)Treat=1, KPS=20; c)Treat=0, KPS=80; d)Treat=1, KPS=80

In figure 1 the survival functions for *VA lung cancer* data are presented. The functions were calculated for patients with standard or test chemotherapy with Karnofsky rating equals 20 or 80. Disease duration and age were fixed as their median values (5 and 62, respectively), Cell and Prior were fixed as 0. For each observation the survival functions received as the results of BST, RSF and DSTE are presented. In figure 1a) and 1b) the functions are quite similar for all the examined methods. The differences exist for functions in figures 1c)

Table 3. Description of *malignant melanoma* data

Variable name	Description
Variables assessed at the time of operation	
Sex	The patients sex (1-male, 0-female)
Age	Age (years)
Thickness	Tumour thickness (cm)
Ulcer	Indicator of ulceration (0-absent, 1-present)
Outcome variables	
Time	Survival time (days)
Status	Failure indicator (0- censored observation, 1- death from melanoma)

Table 4. The integrated Brier scores received for *malignant melanoma* data

Number of trees	RSF	BST	DST Ensemble
50	0.151	0.149	0.149
100	0.152	0.147	0.150
200	0.152	0.148	0.148
500	0.155	0.148	0.147
1000	0.153	0.150	0.146

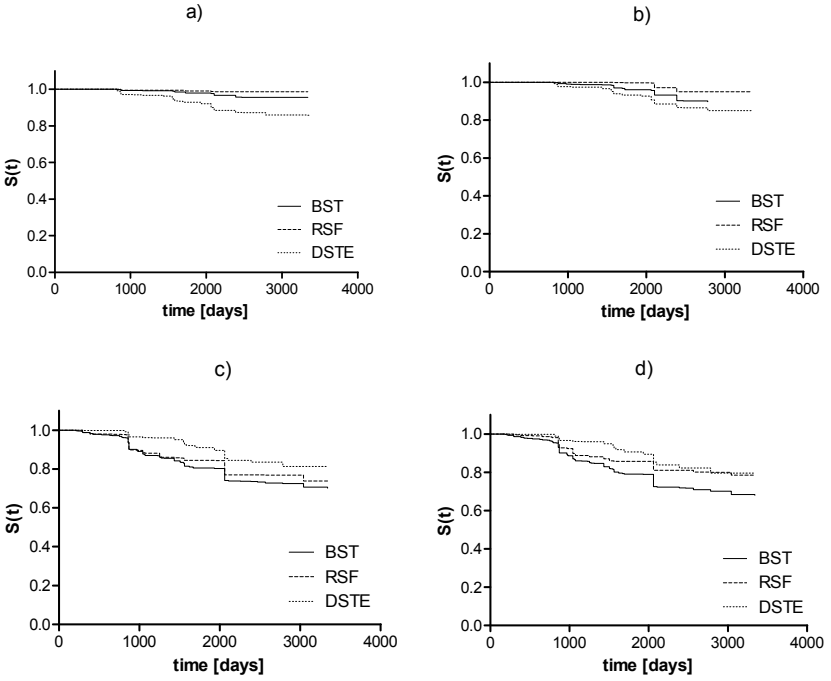

Fig. 2. Survival functions for *malignant melanoma* data a) Sex=0, Thickness=0.97; b) Sex=1, Thickness=0.97; c) Sex=0, Thickness=3.56; d) Sex=1, Thickness=3.56

and 1d). The survival function received for DSTE gives the most pessimistic prediction, especially for time greater than 150 days. Comparing, for example, the probability of survival for 200 days, BST and RSF give the value about 0.4, while for DSTE the probability equals 0.2. Median survival times, also presented in figure 1, are similar for three methods. Analyzing the graphs one could say that the type of treatment does not influence the survival, while Karnofsky rating has a great impact on patients survival.

The other data set contains the information on 205 patients (148 censored cases) with malignant melanoma following radical operation. The data was

collected at Odense University Hospital in Denmark by K.T. Drzewiecki [1]. Each patient is described by four variables presented in table 3.

Table 4 presents the integrated Brier scores received for *malignant melanoma* data. The results for RSF do not depend on the number of trees and the integrated Bries scores take the values from the range [0.151; 0.155]. For bagging survival trees the best result is for 100 trees - IBS=0.147, while for 1000 trees IBS equals 0.15. The best results are for DST ensemble and the minimal value of IBS is equal to 0.146 for 1000 trees.

Figure 2 presents survival functions received for *malignant melanoma* data. The influence of sex and tumor thickness was verified. Variable "Thickness" was fixed as its lower and upper quartiles: 0.97 and 3.56, respectively. The experiments were conducted for 54 years old people without ulceration. As we could see, sex do not influence the survival. The differences are visible between figures with different values of Thickness: the prediction is worse for patients with greater tumor thickness. The results received for BST, RSF and DSTE show the main tendency of survival changes in a similar manner, but the exact prediction is slightly different for them.

6 Conclusions

In the paper the prediction ability of tree-based ensemble methods was verified. The analysis covered the results of three techniques: bagging survival trees, random survival forest and dipolar survival tree ensemble. The prediction ability was tested by calculating the integrated Brier score. The analysis was conducted on the base of two medical data sets. The analysis did not show that one method outperformed the results of two others. The best value of the integrated Brier score in case of *VA lung cancer* data was for bagging survival forest, in case of the other data set - *malignant melanoma* - the best result was achieved by dipolar survival tree ensemble.

Acknowledgements. This work was supported by the grant S/WI/2/1013 from Bialystok University of Technology.

References

1. Andersen, P.K., Borgan, O., Gill, R.D.: Statistical Models based on Counting Processes. Springer, New York (1993)
2. Bobrowski, L., Kretowska, M., Kretowski, M.: Design of neural classifying networks by using dipolar criterions. In: Proc. of the Third Conference on Neural Networks and Their Applications, Kule, Poland, pp. 689–694 (1997)
3. Bou-Hamad, I., Larocque, D., Ben-Ameur, H.: A rewiew of survival trees. Statistics Surveys 5, 44–71 (2011)
4. Kalbfleisch, J.D., Prentice, R.L.: The statistical analysis of failure time data. John Wiley & Sons, New York (1980)

5. Gerds, T.A., Cai, T., Schumacher, M.: The performance of risk prediction models. Biometrical Journal 50(4), 457–478 (2008)
6. Graf, E., Schmoor, C., Sauerbrei, W., Schumacher, M.: Assessment and comparison of prognostic classification schemes for survival data. Statistics in Medicine 18, 2529–2545 (1999)
7. Harrell, F.E., Califf, R.M., Pryor, D.B., Lee, K.L., Rosati, R.A.: Evaluating the yield of medical tests. Journal of American Medical Association 247, 2543–2546 (1982)
8. Hothorn, T., Lausen, B., Benner, A., Radespiel-Troger, M.: Bagging survival trees. Statistics in Medicine 23, 77–91 (2004)
9. Hothorn, T., Buhlmann, P., Dudoit, S., Molinaro, A.M., van der Laan, M.J.: Survival ensembles. Biostatistics 7, 355–373 (2006)
10. Ishwaran, H., Blackstone, E.H., Pothier, C.E., Lauer, M.S.: Relative risk forest for exercise heart rate recovery as a predictor of mortality. Journal of the American Statistical Association 99, 591–600 (2004)
11. Ishwaran, H., Kogalur, U.B., Blackstone, E.H., Lauer, M.S.: Random survival forests. Annals of Applied Statistics 2, 841–860 (2008)
12. Ishwaran, H., Kogalur, U.B.: Random Forests for Survival, Regression and Classification (RF-SRC), R package version 1.3 (2013)
13. Kaplan, E.L., Meier, P.: Nonparametric estimation from incomplete observations. Journal of the American Statistical Association 5, 457–481 (1958)
14. Kretowska, M.: Dipolar regression trees in survival analysis. Biocybernetics and Biomedical Engineering 24(3), 25–33 (2004)
15. Krętowska, M.: Random forest of dipolar trees for survival prediction. In: Rutkowski, L., Tadeusiewicz, R., Zadeh, L.A., Żurada, J.M. (eds.) ICAISC 2006. LNCS (LNAI), vol. 4029, pp. 909–918. Springer, Heidelberg (2006)
16. LeBlanc, M., Crowley, J.: Relative risk trees for censored survival data. Biometrics 48, 411–425 (1993)
17. Mogensen, U.B., Ishwaran, H., Gerds, T.A.: Evaluating Random Forests for Survival Analysis Using Prediction Error Curves. Journal of Statistical Software 50(11), 1–23 (2012), http://www.jstatsoft.org/v50/i11/
18. Peters, A., Hothorn, T.: ipred: Improved Predictors, R package version 0.9-2 (2013), http://CRAN.R-project.org/package=ipred