ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2. Available from http://www.i6doc.com/en/.

Adapting Random Forests to Cope with Heavily Censored Datasets in Survival Analysis

Tossapol Pomsuwan and Alex A. Freitas

School of Computing
University of Kent — Canterbury, UK

Abstract. We address a survival analysis task where the goal is to predict the time passed until a subject is diagnosed with an age-related disease. The main challenge is that subjects' data are very often censored, i.e., their time to diagnosis is only partly known. We propose a new Random Forest variant to cope with censored data, and evaluate it in experiments predicting the time to diagnosis of 8 age-related diseases, for data from the English Longitudinal Study of Ageing (ELSA) database. In these experiments, the proposed Random Forest variant, in general, outperformed a well-known Random Forest variant for censored data.

1 Introduction

We address the survival analysis problem of predicting the time passed from a "baseline date" (when a subject is visited by a nurse) until the date of diagnosis of some age-related disease. The datasets analyzed in this work were derived from the English Longitudinal Study of Ageing (ELSA) [1] — a survey of ageing and quality of life among people aged 50 and over. This paper focuses on the biomedical data in ELSA, such as the results of blood tests and other data collected by nurses, and information about the subjects' age-related diseases.

The main challenge in survival analysis is to cope with censored data [2]. Censoring occurs when observed instances have some information available for estimating the survival time but the information is incomplete. For example, an individual is lost to follow up, drops out of the study, or does not experience the event of interest (a disease's diagnosis in this paper) before the study ends. Classical regression methods cannot effectively handle censorship, since censorship introduces uncertainty into the value of the target variable to be predicted.

This paper proposes a new variant of the Random Forest algorithm for coping with censoring in survival data. We choose Random Forests [3, 4] due to the algorithm's good performance of achieving high predictive accuracy in general, using the power of an ensemble of decision trees to make more robust predictions.

This paper is organised as follows. Section 2 reviews background on survival analysis. Section 3 describes the creation of the datasets used in the experiments. Section 4 describes the proposed Random Forest variant. Section 5 reports experimental results, and Section 6 presents the conclusion.

ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2.
Available from http://www.i6doc.com/en/.

2 Background on Survival Analysis

Survival analysis methods aim at analyzing or predicting the time until the occurrence of an event of interest [5]. Although the time-to-event to be predicted is a numerical variable, survival analysis is very different from classical regression. The main difference is the presence of data censoring in survival analysis, which cannot be effectively handled by traditional linear regression methods.

We focus on right-censoring (as opposed to left-censoring) [5], which is common in medical research and occurs very often in our datasets. Right-censoring occurs when the subject dropped out of the study before its end and no event of interest occurred before the drop out, or when the study ends before the event of interest occurred for a subject. Note that in right-censoring the last observed time for a subject is a lower bound for the unknown event occurrence time.

[6] introduced Inverse Probability of Censoring (IPC) weights, where positive weights are assigned to uncensored instances while the weights of the censored ones are 0. Hence, a subject with a long survival time is assigned a large IPC weight, which is inversely proportional to her/his probability of being censored. The probability of censoring is estimated based on a censoring probability function, i.e., the probability that the censored time is greater than t, denoted $G(t)$, as shown in Equation (1) , where n_j is the number of subjects in the risk set at time j, i.e., the set containing subjects who have survived at least to time j, and c_j is the number of subjects who were censored at time j.

$$G(t) = \prod_{j=0}^{t} \left(1 - \frac{c_j}{n_j}\right) \tag{1}$$

Then, the IPC weight of each instance i (w_i) is estimated by Equation (2):

$$w_i = \begin{cases} \frac{1}{G(t_i)}, & \text{if } i \text{ is uncensored} \\ 0, & \text{Otherwise} \end{cases} \tag{2}$$

Where t_i is the survival time observed for the i-th uncensored subject, and the value of $G(t_i)$ is given by Equation (1).

The IPC weight technique for coping with censorship in Random Forest, called Survival Ensemble, was introduced in [7]. This is used as the baseline method in our experiments, and has also been used in several survival analysis studies [8, 9, 10]. Survival Ensemble learns a Random Forest model where each tree was derived from bootstrap data where each instance is sampled with a probability based on its IPC weight, so that censored instances (with IPC weight = 0) are not used in the tree-building process. This has the clear disadvantage of ignoring potentially very useful information for building the model, namely the censored instances' feature values and their relationships with the partially observed survival time (until the time of censoring). This disadvantage is particularly serious in the datasets used in our experiments, when the large majority of instances are censored.

ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2. Available from http://www.i6doc.com/en/.

3 Dataset Creation

The datasets created in this paper were derived from the English Longitudinal Study of Ageing (ELSA) [1] — www.elsa-project.ac.uk/. The data was collected across 8 waves, with a two-year gap between consecutive waves. In the created datasets, the instances represent subjects in the ELSA database and the predictive features represent mainly biomedical data collected by nurses in wave 2 (the "baseline" wave), including age and gender. We use as baseline wave 2 because this was the first wave when a nurse collected biomedical data from subjects. We created 8 datasets, all with the same set of 44 predictive features from wave 2, but each with a different target variable (to be predicted) measuring the time passed (in months) from the date when a subject received a nurse visit in wave 2 until the date when the subject was first diagnosed with a given disease. The 8 age-related diseases used as target variables are: Angina, Heart Attack, Diabetes, Stroke, Arthritis, Alzheimer's, Cancer and Psychiatric disorder.

The main challenge in our dataset creation was to define a pair of "target" and "uncensorship status" variables for each of the above 8 age-related diseases. Each "uncensorship status" variable takes the value "1" or "0" to indicate whether or not a subject's target variable value is uncensored (fully known) or censored (partly known), respectively. Hence, if a subject has uncensored status = 1, that subject's target variable records the true time passed until her/his first diagnosis of a disease; whilst if a subject has uncensored status = 0, that subject's target variable records only the last time when the subject was observed not to have the diagnosis, which is a lower bound for true value of the target variable.

To determine the value of each subject's target and uncensored status variables, we first tried to obtain the values of ELSA database variables indicating the date the subject was first diagnosed with a disease, and then distinguish between two cases. First, if the values of those two variables were known for a subject, she/is is considered uncensored, and her/his target variable value is directly computed as the number of months passed between the nurse visit to that subject in wave 2 and the date of the subject's first diagnosis for that disease. In the second case, however, the variables indicating year and month of first diagnosis have missing values for a subject (a very common scenario), thus the subject is considered censored. In this case, the computation of the target variable value is much more complex, and it is summarized here into three steps. First, we combine information from several ELSA variables to create a new set of intermediary variables, each indicating whether or not the patient was diagnosed with a given disease at a given wave. Second, by comparing the values of these intermediary variables for a given subject and a given disease across all waves, we determine the last date when the subject was observed and still did not have the diagnosis for that disease. Third, using results of the previous step, the target variable value for each subject and each disease is computed as the number of months passed between the nurse visit to the subject in wave 2 and the last date when the subject was known not to be diagnosed with the disease.

ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2.
Available from http://www.i6doc.com/en/.

In this case, however, that target variable value is a lower-bound to the unknown number of months until first diagnosis (i.e. the subject is censored), and so the "uncensored status" variable is set to 0.

4 The Proposed Random Forest Variant

The proposed variant of Random Forests for survival analysis is based on the idea of imputing the value of the target variable of each censored instance based on its individual lower-bound and upper-bound. The calculation of the target value's lower-bound for each instance was explained in Section 3. The target value's upper-bound for each instance is computed as the number of months passed between the date of the nurse visit and the end of wave 8 (last wave) for censored subjects only. In addition, in order to increase the variance of the trees in the forest, this imputation process is applied independently for every bootstrap training set, as shown in Figure 1. Therefore, the same censored instance may contain different imputed target values in different bootstrap samples. This means that an imputed value of a censored instance is a uniformly random value between its lower-bound and upper-bound. Note that this imputation allows all other procedures of the Random Forest algorithm to be used without modification.

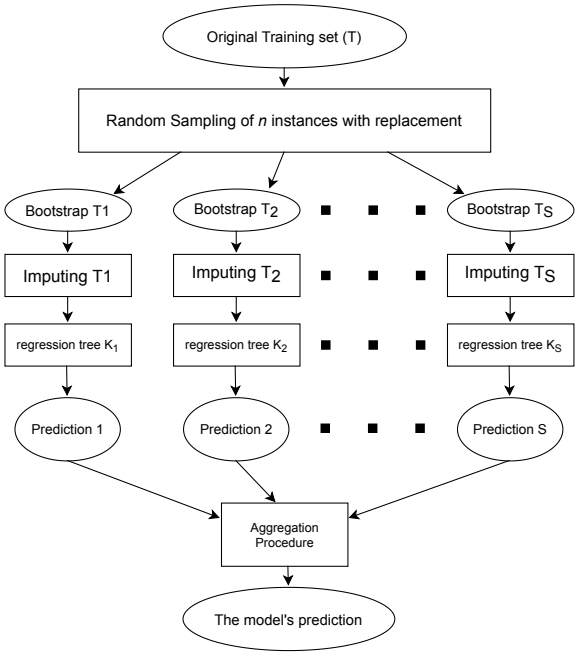

Fig. 1: Overview of the proposed Random Forest variant

ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2. Available from http://www.i6doc.com/en/.

5 Computational Results

We report results for 8 datasets, created as described in Section 3. Recall that each dataset concerns a different age-related disease where the target variable to be predicted is the time passed (in months) until the first diagnosis of that disease for a subject, whilst all datasets have the same predictive features (derived from the baseline wave 2 in ELSA). The predictive performance of the Random Forest (RF) models was estimated by the Concordance index (C-index), which can be interpreted as the probability of correctly ordering the predicted survival values for a randomly chosen pair of subjects whose actual survival times are different. The C-index was adapted for censored data as described in [11], by considering the concordance of actual survival times (diagnosis times in this work) versus predicted survival times among pairs of subjects whose survival outcomes can be ordered with respect to their survival times, i.e., among pairs where both subjects were diagnosed with a certain disease, or one subject was diagnosed before the other subject is censored. The C-index is computed by Equation (3) where *Usable(i,j)* and *Agreed_order(i,j)* are Boolean variables taking the value *true* if the subject pair (i,j) can be ordered and their target variable values agree as defined by Equation (4), where $\widehat{T}_i$ and T_i denote the predicted and actual target values of the i-th subject, respectively.

$$\text{C-index} = \frac{|\{(i,j)|\text{ Usable }(i,j)\text{ AND Agreed\_order }(i,j)\}|}{|\{(i,j)|\text{Usable}(i,j)\}|} \tag{3}$$

$$\text{Agreed\_order}(i,j) = \begin{cases} \text{Yes, if } \widehat{T}_i > \widehat{T}_j \text{ and } T_i > T_j \\ \text{Yes, if } \widehat{T}_j > \widehat{T}_i \text{ and } T_j > T_i \\ \text{No, otherwise} \end{cases} \tag{4}$$

All experiments were performed using nested cross-validation, where 5-fold inner cross-validation performs hyper-parameter tuning and 10-fold outer cross-validation estimates predictive performance. We tuned two hyper-parameters of the Random Forest algorithm: the node size (the minimum number of instances allowed at leaf nodes) and mtry (the number of randomly sampled candidate features at each tree node). The tuning procedure tried 3 node size values (5, 7 and 10) and 4 mtry values (4, 7, 10, 13), leading to 12 combinations.

Table 1 shows the C-index values and the Standard Error of the Mean (SEM) (over 10-fold cross-validation) obtained by two variants of Random Forest (RF), the RF described in [7] (IPC weight approach), and our proposed Random Target-Imputation Forest (RTIF) method. The second column of this table shows the relative frequency (ratio) of uncensored instances in each dataset. Note that in all datasets the large majority of instances are censored. As reported in the C-index columns, the proposed RTIF obtained the best result in 7 out of the 8 datasets. Both RF variants had poor performance (C-index around 0.5) in 3 datasets (Arthritis, Cancer and Psychiatric disorder); whilst the three highest C-index values are 0.7742, 0.7443 and 0.6366, obtained by the proposed RTIF in Alzheimer, Diabetes and Stroke datasets, respectively.

ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2. Available from http://www.i6doc.com/en/.

Table 1: Predictive performance obtained by two variants of Random Forests.

Dataset		IPC_weight [7]		proposed RTIF	
Disease	uncensoring ratio	C-index	SEM	C-index	SEM
Angina	165/6488 (2.5%)	0.5691	0.0297	**0.5723**	0.0288
HeartAtt	186/6607 (2.8%)	0.5894	0.0206	**0.6228**	0.0266
Diabetes	416/6500 (6.4%)	0.6796	0.0196	**0.7443**	0.0171
Stroke	270/6632 (4.1%)	0.5983	0.0215	**0.6366**	0.0335
Arthritis	784/4276 (18.3%)	0.5068	0.0136	**0.5078**	0.0196
Alzheimer	69/6825 (1.0%)	0.6576	0.0389	**0.7742**	0.0338
Cancer	562/6386 (8.8%)	0.5071	0.0171	**0.5135**	0.0199
Psychiatric	219/5972 (3.5%)	**0.4834**	0.0225	0.4692	0.0335

6 Conclusions

We have proposed a new Random Forest variant for coping with heavy censoring in survival data. This variant stochastically imputes the values of the target variable for censored instances by using instance-specific lower and upper bounds. The experimental results have shown that our proposed variant in general outperformed the baseline method [7] in 8 age-related disease datasets.

References

[1] S. Clemens, A. Phelps, et al. English Longitudinal Study of Ageing: Waves 0-8, 2019.

[2] E. L. Kaplan and Paul Meier. Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, 53(282):457–481, 1958.

[3] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[4] Wouter G. Touw, Jumamurat R. Bayjanov, and et al. Data mining in the life science swith random forest. *Briefings in Bioinformatics*, 14(3):315–326, 2013.

[5] David G Kleinbaum and Mitchel Klein. *Introduction to Survival Analysis*, pages 1–54. Springer, 2012.

[6] James M. Robins and Andrea Rotnitzky. Recovery of Information and Adjustment for Dependent Censoring Using Surrogate Markers. pages 297–331. Birkhäuser Boston, 1992.

[7] Torsten Hothorn, Peter Buhlmann, Sandrine Dudoit, Annette Molinaro, and Mark J. Van Der Laan. Survival ensembles. *Biostatistics*, 7(3):355–373, 2006.

[8] Hemant Ishwaran, Udaya B. Kogalur, Eugene H. Blackstone, and Michael S. Lauer. Random survival forests. *Annals of Applied Statistics*, 2(3):841–860, 2008.

[9] David M Vock, Julian Wolfson, et al. Adapting machine learning techniques to censored time-to-event health record data. *Journal of Biomedical Informatics*, 61:119–31, 2016.

[10] Nikolaj Tollenaar and Peter G.M. Van Der Heijden. Optimizing predictive performance of criminal recidivism models using registration data with binary and survival outcomes. *PLoS ONE*, 14(3):e0213245, Mar 2019.

[11] Frank E Harrell, Kerry L Lee, and Daniel B Mark. Tutorial in Biostatisticsmultivariable Prognostic Models: Issues in Developing Models, Evaluating Assumptions and Adequacy,and Measuring and Reducing Errors. *Statistics in Medicine*, 15(4):361–387, 1996.