Computer Methods and Programs in Biomedicine 159 (2018) 185–198

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.elsevier.com/locate/cmpb

Tree-based models for survival data with competing risks

Malgorzata Kretowska

Faculty of Computer Science, Bialystok University of Technology, Wiejska 45a, Bialystok 15-351, Poland

ARTICLE INFO

Article history:
Received 22 November 2017
Revised 13 March 2018
Accepted 20 March 2018

Keywords:
Oblique tree
Ensemble of trees
Piecewise-linear criterion function
Survival analysis
Competing risks

ABSTRACT

Objective. Tree-based models belong to common, assumption-free methods of data analysis. Their application in survival data is narrowed to univariate models, which partition the feature space with axis-parallel hyperplanes, meaning that each internal node involves a single feature. In this paper, I extend the idea of oblique survival tree induction for competing risks by modifying a piecewise-linear criterion function. Additionally, the use of tree-based ensembles to analyze the competing events is proposed.

Method and materials. Two types of competing risks trees are proposed: a single event tree designed for analysis of the event of interest and a composite event tree, in which all the competing events are taken into account. The induction process is similar, except that the calculation of the criterion function is minimized for the individual tree nodes generation. These two tree types were also used for building the ensembles with aggregated cumulative incidence functions as an outcome. Nine real data sets, as well as a simulated data set, were taken to assess performance of the models, while detailed analysis was conducted on the basis of follicular cell lymphoma data.

Results. The evaluation was focused on two measures: the prediction error expressed by an integrated Brier score (IBS), and the ranked measure of predictive ability calculated as a time-truncated concordance index (C–index). The proposed techniques were compared with the existing approaches of the Fine–Gray subdistribution hazard model, Fine–Gray regression model with backward elimination, and random survival forest for competing risks. The results for both the IBS and the C–index indicated statistically significant differences between these methods ($p < .0001$).

Conclusions. The prediction error of the individual trees was similar to the other methods, but the results of the C–index differ in comparison to the Fine–Gray subdistribution hazard model and the Fine–Gray regression with backward elimination. The ensembles prediction ability was comparable to existing algorithms, but their IBS values were better than either random survival forest or Fine–Gray regression with backward elimination.

1. Introduction

In survival analysis, we usually measure the time to one predefined event of interest, specifically, death or disease relapse. The straightforward extension of survival data is a competing risks data, which enables analyzing the risk of several different events–an event of interest and one or more competing events. Methods developed to cope with such types of data should take into account not only the variety of events but also the observations with no event assigned, the so-called right-censored cases. Such observations end before any event occurs, and the only information they provide is the event-free survival time.

Among tree-based methods developed to cope with competing risks, we can distinguish single trees and ensembles of predictors. The general tree induction algorithm consists of two phases. The first, the growing phase, builds the whole tree with its internal and terminal nodes, while the other one, the pruning phase, reduces the size of the tree (the number of nodes) to prevent overfitting the data [1].

Single tree models applied to competing risks are divided into two separate groups: *single event* (or univariate) and *composite events* (or multivariate) techniques [2,3]. The first group covers the methods aiming at analysis of the event of interest, pooling all the competing events. Callaghan [2] proposed two separate tree models that used Gray's two-sample test of cumulative incidence function (CIF) [4] to measure between-node heterogeneity and, in the other case, the event-specific martingale residuals to measure within-node homogeneity. Ibrahim and Kudus [3] suggested using the deviance of the between-node difference that is derived from the likelihood ratio test statistic of the proportional hazards model of subdistribution [5]. The approach based on the likelihood ratio test for the splitting rule is also proposed by Xu et al. [6]. In

E-mail address: m.kretowska@pb.edu.pl

https://doi.org/10.1016/j.cmpb.2018.03.017
0169-2607/

multivariate trees, the competing risks are considered as separate events, and the splitting criterion is based on multivariate tests for comparing CIFs [2,3]. The most common pruning procedure utilizes the methodology proposed by LeBlanc and Crowley [7], who developed the pruning algorithm described in CART [1] for analysis of survival data. In composite event trees, Ibrahim and Kudus [3] adopted Segal's pruning algorithm [8].

A tree-based ensemble is a set of k decision, regression or survival trees built on the base of bootstrap samples, drawing from the learning set. The induction of a single tree belonging to the ensemble does not require the pruning phase. Ishwaran et al. [9] proposed a random survival forest in which a generalized log-rank test based on the weighted difference of cause-specific Nelson–Aalen estimates in the two child nodes and Gray's test are used as the splitting rule for single event trees. Proposed by them, a composite splitting rule uses a combination of the cause-specific splitting rules across the event types. For each tree, the estimators of CIF, the cause j mortality, cumulative event-specific hazard function and event-free survival are obtained. The ensemble estimators are calculated as the average over k trees. The approach is available in R package randomForestSRC [10]. Mogensen and Gerds [11] replaced censored event status with a jackknife pseudo value, that caused the random forest to be built for uncensored data. They proposed a pseudo split criterion, which was then compared with the Gini index.

The tree induction procedures developed to cope with survival data with competing risks are limited to univariate trees, in which a single split refers to only one variable and usually has a form of $x_i < c$, where x_i is a variable and $c \in R$. In this paper, I extend the methodology proposed in [12], where a convex piecewise-linear criterion function, the dipolar criterion, was used in the oblique survival tree induction. As a result, a tree with internal nodes containing the splits of the form of any hyperplane, which need not be parallel to the coordinate axes, is created. The basic procedure was dedicated for right–censored data. In this paper, I adopt the method to survival data with competing risks by changing the way the dipolar criterion function is calculated, both in univariate and multivariate approaches. It leads to trees that divide the feature space into disjoined areas with homogeneous failure experiences characterized by cumulative incidence functions. Based on this methodology, I propose also a tree-based ensemble, introduced in [13], to be used in competing risks analysis. The methods are then compared with already existing approaches, such as the Fine–Gray subdistribution hazard model [5], the Fine–Gray regression model with backward elimination [14] and the random survival forest for competing risks with log–rank splitting [15]. A time-truncated concordance index (C–index) for competing risks [16] and an integrated Brier score (IBS) [17,18] are applied to validate the model performance. A follicular cell lymphoma dataset [19] is used to illustrate the capabilities of the proposed approaches, while results on a synthetic dataset show the performance of the models when the true survival time distribution is known.

This paper consists of 8 sections. Section 2 introduces a definition and basic concepts of survival data with competing risks. In Section 3, the idea of dipolar criterion function is recalled and its modifications aiming at competing risks analysis, together with their application in univariate and multivariate trees induction, are presented. A tree-based ensemble is described in Section 4. Possible validation measures are presented in Section 5, while Section 6 shows the results of the experiments on simulated and real data. Section 7 analyzes the results, and Section 8 contains the conclusions.

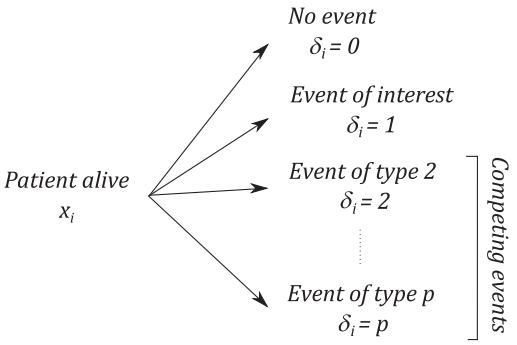

Fig. 1. Competing risks data.

2. Competing risks data

In a competing risks study a patient is observed from a starting event until the first of p ($p > 1$) final events occurs. The follow-up of right-censored observations is interrupted by other causes. The learning sample, L, for competing risk data is defined as $L = (\mathbf{x}_i, t_i, \delta_i)$, $i = 1, 2, \ldots, M$, where $\mathbf{x}_i$ is N-dimensional covariates vector, t_i is a time to the first event observed, while $\delta_i = \{0, 1, \ldots, p\}$ indicates the cause of failure; $\delta_i = 0$ represents right-censored observations (i.e., observations without any failure occurring); $\delta_i = 1$ denotes an event of interest; and $\delta_i > 1$ denotes competing events (Fig. 1).

The distribution of the random variable T (time), for an event of type i ($i = 1, 2, \ldots, p$) may be represented by several functions. Some of the most popular functions are a CIF defined as the probability that the event of type i occurs at or before time t [20],

$$F_i(t) = P(T \le t, \delta = i) \tag{1}$$

and a survival function,

$$S_i(t) = P(T > t, \delta = i). \tag{2}$$

The estimator of the CIF is calculated as

$$\hat{F}_i(t) = \sum_{j|t_j \le t} \frac{d_{ij}}{m_j} \hat{S}(t_{j-1}) \tag{3}$$

where $t_{(1)} < t_{(2)} < \ldots < t_{(D)}$ are distinct, ordered, uncensored time points from the learning sample L, d_{ij} is the number of events of type i at time $t_{(j)}$, 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 an event at $t_{(j)}$), and $\hat{S}(t)$ is the Kaplan–Meier estimator of the probability of being free of any event by time t. It is calculated as

$$\hat{S}(t) = \prod_{j|t_{(j)} \le t} \left(\frac{m_j - d_j}{m_j} \right) \tag{4}$$

where d_j is the number of events at time $t_{(j)}$.

The patient-specific CIF for the event of type i is given by $F_i(t|\mathbf{x}) = P(T \le t, \delta = i | \mathbf{X} = \mathbf{x})$. The conditional CIF for the new patient with the covariate vector $\mathbf{x}_{new}$ is denoted by $\hat{F}_i(t|\mathbf{x}_{new})$.

3. Dipolar survival trees for competing risks

A binary tree is a structure built from internal and terminal nodes that divide the feature space into disjoined areas containing similar elements – feature vectors. The definition of similarity depends on the analyzed problem. In survival analysis, two feature vectors are similar if their failure times are alike. In the classical top-down algorithm, the induction process of a tree starts from a root node. Based on the learning dataset, the algorithm decides how to split the feature space to obtain the best outcome in the