Artificial Intelligence in Medicine 75 (2017) 32–39

Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

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

Piecewise-linear criterion functions in oblique survival tree induction

Malgorzata Kretowska

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

ARTICLE INFO

Article history:
Received 15 July 2016
Received in revised form 7 November 2016
Accepted 28 December 2016

Keywords:
Piecewise-linear criterion function
Survival tree
Oblique splits
Right-censored data

ABSTRACT

Objective: Recursive partitioning is a common, assumption-free method of survival data analysis. It focuses mainly on univariate trees, which use splits based on a single variable in each internal node. In this paper, I provide an extension of an oblique survival tree induction technique, in which axis-parallel splits are replaced by hyperplanes, dividing the feature space into areas with a homogeneous survival experience.

Method and materials: The proposed tree induction algorithm consists of two steps. The first covers the induction of a large tree with internal nodes represented by hyperplanes, whose positions are calculated by the minimization of a piecewise-linear criterion function, the dipolar criterion. The other phase uses a split-complexity algorithm to prune unnecessary tree branches and a 10-fold cross-validation technique to choose the best tree. The terminal nodes of the final tree are characterised by Kaplan–Meier survival functions. A synthetic data set was used to test the performance, while seven real data sets were exploited to validate the proposed method.

Results: The evaluation of the method was focused on two features: predictive ability and tree size. These were compared with two univariate tree models: the conditional inference tree and recursive partitioning for survival trees, respectively. The comparison of the predictive ability, expressed as an integrated Brier score, showed no statistically significant differences ($p = 0.486$) among the three methods. Similar results were obtained for the tree size ($p = 0.11$), which was calculated as a median value over 20 runs of a 10-fold cross-validation.

Conclusions: The predictive ability of trees generated using piecewise-linear criterion functions is comparable to that of univariate tree-based models. Although a similar conclusion may be drawn from the analysis of the tree size, in the majority of the studied cases, the number of nodes of the dipolar tree is one of the smallest among all the methods.

1. Introduction

The prediction of failure time is one of the major tasks in survival analysis. In the medical domain, it often describes the time to death or disease relapse. Cox's proportional hazards model [1] is one of the most common statistical methods used to analyse survival data. This semi-parametric model requires the fulfillment of certain assumptions about an analysed phenomenon that is often difficult to achieve. Some other restrictions concern accelerated failure time models [2], for which the analytical form of the relationship between the survival function and the covariates should be established. The requirements accompanying statistical models result in the development of alternative, assumption-free methods of survival analysis. Among them, tree-based models play an important role.

Survival trees are mainly intended to analyze right-censored data, and their first applications appeared in the eighties. As pointed out by LeBlanc and Crowley [3], tree induction algorithms may be categorized from the point of view of a splitting criterion (i.e., the impurity or the between-node separation measure). The first group covers the algorithms following the CART (Classification and Regression Trees) methodology [4]. Gordon and Olshen [5] used the Wasserstein metric, Davis and Anderson [6] applied exponential log-likelihood loss, LeBlanc and Crowley [7] applied an approximation of the full likelihood for the proportional hazards model, while Therneau et al. [8] used martingale-based residuals from the Cox model. The other group of algorithms is usually based on the Tarone–Ware class of two-sample statistics for censored data, such as the log-rank test [9–11].

Another important aspect of tree induction methods is a stopping criterion. Its appropriate choice causes the final tree to have a good generalization ability; too small or too large trees lead to an under- or overfitting phenomenon. A common way to select the

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

http://dx.doi.org/10.1016/j.artmed.2016.12.004
0933-3657/© 2017 Elsevier B.V. All rights reserved.

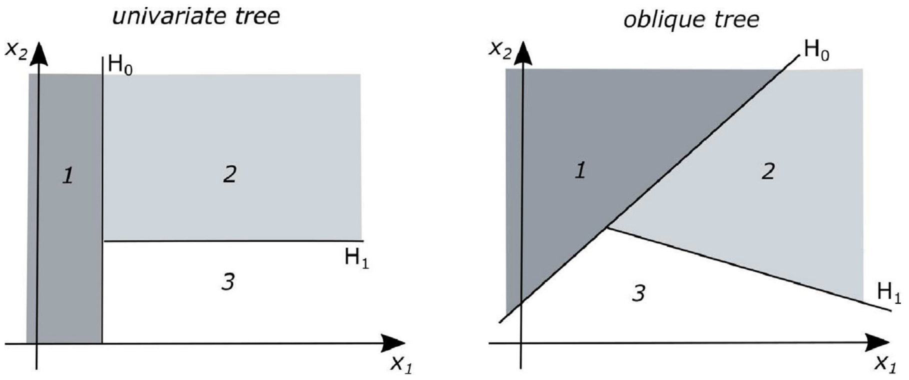

Fig. 1. Division of the feature space into three disjoined areas (1,2,3) by the hyperplanes H_0 and H_1 in univariate and oblique trees.

final tree is to build a large tree and then prune some of its branches. The idea was proposed in [4] as cost-complexity pruning and was then extended to survival trees by LeBlanc and Crowley [11] and referred to as a split-complexity algorithm. Another approach does not separate the pruning phase from the induction process. Rather, a decision on the split importance is made during creation of the node. Hothorn et al. [12] proposed the use of multiple test procedures and to stop the split if the test results are not statistically significant at a given value of α.

Comparisons of different splitting criterion and pruning techniques was presented in [13,14], while a comprehensive overview of tree-based models was provided by Bou–Hamad et al. [15].

Although single trees are now often replaced by more powerful ensembles of trees, they have one undeniable advantage: an insight into data [15], made possible by analysing splits in subsequent internal nodes, which divide the feature space into homogeneous areas. The survival trees are narrowed to univariate trees, in which one split is based on only one variable. In real data, the borders between regions with different survival experiences need not be parallel to the coordinate axes (Fig. 1). If we use a univariate tree to solve this problem, we must create a number of internal nodes instead of one hyperplane.

In this paper, I develop a method of oblique survival tree induction, introduced briefly in [16]. Here, a single split is equivalent to any hyperplane, whose location is determined by the minimization of a convex and piecewise-linear (CPL) criterion function [17] built based on right-censored data. The performance of the final tree, chosen by a split-complexity pruning method [11], was compared with those of two univariate tree-based models: a conditional inference tree [12] and recursive partitioning for survival trees [18] (R package: rpart), which corresponds to a method proposed by LeBlanc and Crowley [7].

The paper consists of 7 sections. Section 2 introduces a definition and basic concepts of survival data. In Section 3, the piecewise-linear criterion function, the dipolar criterion, is presented. An oblique survival tree induction algorithm is described in Section 4. Possible validation measures are presented in Section 5, while Section 6 shows the results of the experiments on synthetic and real data. Section 7 contains the conclusions.

2. Survival data

Observe random variable $P=(\mathbf{X}, T, \Delta)$, where $\mathbf{X}$ is the N-dimensional feature vector, $T=\min(T_0, C)$, T_0 is the survival time with the distribution function f_t, C is the censoring time with the distribution function f_c, and Δ is the censoring indicator $\Delta=I(T_0<C)$. A learning sample, L, consists of M observations $(\mathbf{x}_i, t_i, \delta_i)$, $i=1, 2, \ldots, M$, where $\mathbf{x}_i$ is the N-dimensional feature vector describing the ith patient, t_i is the survival time, and δ_i is the failure indicator, which takes one of two values: 0 for censored observations or 1 for uncensored ones.

The distribution of the survival time may be described by several functions. One of them is a survival function, which represents the probability of surviving beyond the time t: $S(t)=P(T>t)$. One of the most common nonparametric estimators of the survival function is the Kaplan–Meier product-limit estimator [19]. If we assume that the events of interest occur at D distinct times $t_{(1)}<t_{(2)}<\ldots<t_{(D)}$, the estimator is calculated as follows:

$$\hat{S}(t)=\prod_{j|t_{(j)}\leq t}\left(\frac{m_j-d_j}{m_j}\right) \tag{1}$$

where 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 $t_{(j)}$).

3. Dipolar criterion function

CPL criterion functions are common methods used in data analysis. In this paper, a CPL function, the dipolar criterion $\Psi_d(\cdot)$ [17], was used to determine the splits in the internal nodes of survival trees.

Let us introduce the augmented feature and weight vectors:

$$\begin{aligned}\boldsymbol{z} &= [1, x_1, x_2, \ldots, x_N]^T\\ \boldsymbol{v} &= [-\theta, w_1, w_2, \ldots, w_N]^T\end{aligned} \tag{2}$$

For any feature vector $\boldsymbol{z}_j$, $j=1, 2, \ldots, M$ from the learning set L, we can define two piecewise-linear penalty functions:

$$\varphi_j^+(\boldsymbol{v})=\begin{cases}\delta_j-\boldsymbol{v}^T\boldsymbol{z}_j & \text{if} \quad \boldsymbol{v}^T\boldsymbol{z}_j\leq\delta_j\\ 0 & \text{if} \quad \boldsymbol{v}^T\boldsymbol{z}_j>\delta_j\end{cases} \tag{3}$$

and

$$\varphi_j^-(\boldsymbol{v})=\begin{cases}\delta_j+\boldsymbol{v}^T\boldsymbol{z}_j & \text{if} \quad \boldsymbol{v}^T\boldsymbol{z}_j\geq-\delta_j\\ 0 & \text{if} \quad \boldsymbol{v}^T\boldsymbol{z}_j<-\delta_j\end{cases} \tag{4}$$

where $\delta_j\geq 0$ is a margin usually equal to 1. In Fig. 2, we can see graphical representations of $\varphi_j^+(\boldsymbol{v})$ and $\varphi_j^-(\boldsymbol{v})$ compared to the scalar product $\boldsymbol{v}^T\boldsymbol{z}_j$.

If we take into account a hyperplane $H(\boldsymbol{v})=\{\boldsymbol{z}:\boldsymbol{v}^T\boldsymbol{z}=0\}$ (or, equivalently, $H(\boldsymbol{w},\theta)=\{\boldsymbol{x}:\boldsymbol{w}^T\boldsymbol{x}=\theta\}$), the functions $\varphi_j^+(\boldsymbol{v})$ and $\varphi_j^-(\boldsymbol{v})$, associated with a given feature vector $\boldsymbol{z}_j$, penalize for the inappropriate position of $H(\boldsymbol{v})$ toward $\boldsymbol{z}_j$. The minimization of the penalty enforces a correct localisation of $H(\boldsymbol{v})$; in addition, with a margin greater than zero, the hyperplane is unable to pass through $\boldsymbol{z}_j$, which improves the generalization ability.