Expert Systems With Applications 63 (2016) 450–459

Contents lists available at ScienceDirect

Expert Systems With Applications

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

On the use of Harrell's C for clinical risk prediction via random survival forests

Matthias Schmid[a,*], Marvin N. Wright[b], Andreas Ziegler[b,c,d,e]

[a] Institut für Medizinische Biometrie, Informatik und Epidemiologie, Rheinische Friedrich-Wilhelms-Universität Bonn, Sigmund-Freud-Str. 25, 53127 Bonn, Germany
[b] Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Ratzeburger Allee 160, Geb. 24, 23562 Lübeck, Germany
[c] Zentrum für Klinische Studien, Universität zu Lübeck, Ratzeburger Allee 160, Geb. 24, 23562 Lübeck, Germany
[d] School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, King Edward Avenue, Scottsville 3209, Pietermaritzburg, South Africa
[e] Deutsches Zentrum für Herz-Kreislauf-Forschung, Standort Hamburg/Kiel/Lübeck, Oudenarder Str. 16, 13347 Berlin, Germany

ARTICLE INFO

Article history:
Received 13 March 2016
Revised 11 July 2016
Accepted 12 July 2016
Available online 16 July 2016

Keywords:
Concordance index
Event history analysis
Log-rank statistic
Random survival forests
Risk prediction
Split rules

ABSTRACT

Random survival forests (RSF) are a powerful method for risk prediction of right-censored outcomes in biomedical research. RSF use the log-rank split criterion to form an ensemble of survival trees. The most common approach to evaluate the prediction accuracy of a RSF model is Harrell's concordance index for survival data ('*C* index'). Conceptually, this strategy implies that the split criterion in RSF is different from the evaluation criterion of interest. This discrepancy can be overcome by using Harrell's *C* for both node splitting and evaluation. We compare the difference between the two split criteria analytically and in simulation studies with respect to the preference of more unbalanced splits, termed end-cut preference (ECP). Specifically, we show that the log-rank statistic has a stronger ECP compared to the *C* index. In simulation studies and with the help of two medical data sets we demonstrate that the accuracy of RSF predictions, as measured by Harrell's *C*, can be improved if the log-rank statistic is replaced by the *C* index for node splitting. This is especially true in situations where the censoring rate or the fraction of informative continuous predictor variables is high. Conversely, log-rank splitting is preferable in noisy scenarios. Both *C*-based and log-rank splitting are implemented in the R package `ranger`. We recommend Harrell's *C* as split criterion for use in smaller scale clinical studies and the log-rank split criterion for use in large-scale 'omics' studies.

1. Introduction

Random forests are among the most powerful methods for risk prediction in the biomedical sciences. The basic idea of random forests is to fit an ensemble of classification and regression trees (CART) to bootstrap samples that are generated from a set of learning data (Breiman, 2001). Ensemble predictions are obtained by averaging predictions from the individual trees (Kruppa et al., 2014). An important element of random forests is that only a small number of the predictor variables is made available for splitting, which is done at random in each node of a tree. With this randomization element, trees are decorrelated, and the variance of the ensemble prediction is reduced. The random selection of predictors also constitutes the main difference between random forests and earlier tree-based ensemble methods, such as bootstrap aggregating (bagging; Breiman, 1996).

Random forests were originally proposed for classifying dichotomous outcomes (Breiman, 2001) and have been extended over the past 15 years in a number of ways. For example, various methods have been developed for judging the importance of predictor variables, which may serve as a basis for variable selection (Díaz-Uriarte & Alvarez de Andrés, 2006; Ishwaran, Kogalur, Chen & Minn, 2011). It is also possible to estimate individual probabilities for both dichotomous and categorical outcomes (Kruppa, Schwarz, Arminger & Ziegler, 2013) and to analyze continuous outcomes as well as right-censored event times (Ishwaran, Kogalur, Blackstone & Lauer, 2008). Finally, considerable progress has been made in understanding the statistical properties of random forests, including results on consistency and asymptotic normality (Arlot & Genuer, 2014; Biau, 2012; Mentch & Hooker, 2016; Scornet, Biau & Vert, 2015; Wager & Athey, 2015; Wager & Walther, 2015).

* Corresponding author.
E-mail addresses: matthias.schmid@imbie.uni-bonn.de, schmid@imbie.meb.uni-bonn.de (M. Schmid), wright@imbs.uni-luebeck.de (M.N. Wright), ziegler@imbs.uni-luebeck.de (A. Ziegler).

http://dx.doi.org/10.1016/j.eswa.2016.07.018
0957-4174/

Reviews can be found elsewhere; see, e.g., Boulesteix, Janitza, Kruppa & König (2012); Kruppa et al. (2014); Touw et al. (2013); Ziegler & König (2014).

The standard approach to analyze survival outcomes with random forests is termed *random survival forests* (RSF; Ishwaran et al., 2008). In RSF, an ensemble of survival trees is built, and tree splitting is performed by maximizing the log-rank statistic in each node. Ensemble predictions are given by averages over the cumulative hazard estimates in the terminal nodes of the trees, as estimated by the Nelson–Aalen estimator. The most common approach to evaluate the predictive performance of the ensemble is the calculation of the *C* statistic for survival data, also termed 'Harrell's *C*' (Harrell, Califf, Pryor, Lee & Rosati, 1982; Ishwaran et al., 2008). A value of $C = 0.5$ corresponds to a non-informative prediction rule whereas $C = 1$ corresponds to perfect association, implying that Harrell's *C* is an easy-to-interpret coefficient that accounts for the whole range of the observed survival times (Schmid & Potapov, 2012). In biomedical applications, in particular in the analysis of gene expression data, *C* often ranges between the values 0.6 and 0.75. For example, estimates in this range were reported, among others, by Van Belle, Pelckmans, Van Huffel and Suykens (2011a), Schröder, Culhane, Quackenbush and Haibe-Kains (2011) and Zhang, Xia, Lu, Sun and Wang (2013). A remaining disadvantage of the RSF approach with *C*-based evaluation, however, is that the split criterion used for tree building is different from the performance criterion used to measure prediction accuracy. As a result, the performance measure of interest, i.e., Harrell's *C*, may not be fully optimized by the log-rank splits and may even have characteristics that are not reflected by the log-rank statistic.

In this work, we therefore investigate whether the performance of RSF can be improved if Harrell's *C* is used for *both* node splitting and the evaluation of prediction accuracy. In other words, the idea is to replace the log-rank split criterion by Harrell's *C* and to determine split points that are optimal with respect to Harrell's *C* in each node. In Section 2 we first provide a description of the random forest algorithm for survival data, which is followed by theoretical considerations on the two split criteria. In two simulation studies and with the re-analysis of two cancer data sets (Section 3) we finally demonstrate that the use of Harrell's *C* can lead to systematic improvements in the predictive performance of RSF.

2. Methods

2.1. Random survival forests

Algorithm 1 provides a description of the RSF algorithm for n independent observations and p predictor variables. Before the algorithm starts, the number of trees, termed `ntree`, of the RSF and the number of predictor variables `mtry` available for splitting at each node need to be defined. Recently, Lopes (2015) derived the limiting distribution of the prediction error for dichotomous endpoints and showed how this finding may be used for determining optimized values of `ntree`. Kruppa et al. (2013) demonstrated how the hyper-parameter `mtry` can be optimized.

An important feature of the RSF algorithm is the use of the log-rank statistic to split observations at each node and in every tree (Step 2 in Fig. 1). The log-rank statistic will be formally introduced below. At a specific node, the variable and the split point that maximize the log-rank statistic over all possible split points and all `mtry` variables are used for splitting. With this approach, the dissimilarity of the survival curves in the two children nodes is maximized. An alternative criterion for node splitting in Step 2 is Harrell's *C*, which will also be considered below.

The performance of the random survival forest is evaluated using independent test data in Steps 3 and 4 of the algorithm. If no independent data are available, the out-of-bag data generated in Step 1 are used to evaluate the predictive performance. It is important to note that several summary measures are available in Step 3 of the algorithm. For example, Kaplan–Meier or Nelson–Aalen estimates can be derived in each terminal node, and results may be averaged over all trees. In addition, confidence intervals can be obtained for these estimators (Mentch & Hooker, 2016; Wager, Hastie & Efron, 2014; Wager & Walther, 2015).

2.2. The C statistic and the log-rank statistic as split criteria

In this section we introduce the use of Harrell's *C* as split criterion, and we start with a theoretical analysis of both Harrell's *C* and the log-rank statistic. Specifically, we show that both split criteria are special cases of the Gehan statistic (Gehan, 1965) and that they can be obtained from the Gehan statistic by applying different standardization and weighting schemes. Since these schemes are different, differences can be expected between the two criteria regarding their splitting behavior in RSF. First, we introduce basic notation and provide formal definitions of the log-rank, *C* and Gehan statistics. Second, we analyze the connection between the measures. Third, we provide a description of how Harrell's *C* should be used as a split criterion in random forests for survival data.

Notation

Throughout this paper we assume that RSF are fitted to a set of independent and identically distributed data of size n. The data are represented by vectors $(\tilde{T}_i, \Delta_i, X_{i1}, \ldots, X_{ip})$, $i = 1, \ldots, n$, where $\tilde{T}_i$ is a possibly right-censored continuous survival time and $(X_{i1}, \ldots, X_{ip})^\top$ is a vector of predictor variables. It is assumed that $\tilde{T}_i$ is the minimum of the true survival time T_i and an independent continuous censoring time C_i. The variable $\Delta_i := \mathrm{I}(T_i \leq C_i)$ indicates whether T_i has been fully observed ($\Delta_i = 1$) or not ($\Delta_i = 0$). To simplify notation, we assume that there are no tied observed survival times in the data. A predictor variable X_j, $j \in \{1, \ldots, p\}$, is called non-informative if the distribution of $\tilde{T}$ conditional on X_j does not depend on X_j. Otherwise, X_j is called informative.

The events are observed at K ordered time points $t_{(1)} < \ldots < t_{(K)}$ with $K \leq n$. The numbers of events and observations at risk at $t_{(k)}$, $k = 1, \ldots, K$, are denoted by d_k and Y_k, respectively.

As described in Step 3 of the RSF algorithm (Fig. 1), the outcome of an RSF is calculated from the cumulative hazard estimates in the terminal nodes. A one-dimensional score $\eta_i \in \mathbb{R}$ is estimated for each observation $i = 1, \ldots, n$, by averaging the cumulative hazard estimates over all trees and all time points $t_{(k)}$.

Definition of Harrell's C

Harrell's *C* (Harrell et al., 1982) is given by

$$C = \frac{\sum_{i,j} \mathrm{I}(\tilde{T}_i > \tilde{T}_j) \cdot \mathrm{I}(\eta_j > \eta_i) \cdot \Delta_j}{\sum_{i,j} \mathrm{I}(\tilde{T}_i > \tilde{T}_j) \cdot \Delta_j}, \quad (1)$$

where the indices i and j refer to pairs of observations in the sample. The *C* statistic is the number of concordant pairs of observations divided by the number of comparable pairs. Multiplication by the factor Δ_j in Eq. (1) discards pairs of observations that are not comparable because the smaller survival time is censored, i.e., $\Delta_j = 0$.

Harrell's *C* is designed to estimate the concordance probability $\mathrm{P}(\eta_j > \eta_i \,|\, T_i > T_j)$, which compares the rankings of two independent pairs of survival times T_i, T_j and predictions η_i, η_j. The concordance probability evaluates whether large values of η_i are associated with small values of T_i and vice versa. Harrell' *C* can also be interpreted as a summary measure of the area(s) under the time-dependent ROC curves (Heagerty & Zheng, 2005; Schmid & Potapov, 2012). A value of $C = 0.5$ corresponds to a non-informative prediction rule, whereas