COBRA in Cause Specific Survival Prediction

A Project Report Submitted
for the Course

Advanced Statistical Algorithms
MA691

by

Shrey Jani 180102095
Sukrit Bagaria 180108043
Amrit Ayushman 180103009
Pankaj Kumar 180123031
Mridul Garg 180123029

to the

Asst. Prof. Arabin Kumar Dey

DEPARTMENT OF MATHEMATICS
INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI
GUWAHATI - 781039, INDIA

November 2021

1 Random Survival Forest for Competing Risk

Random survival forest is a very popular tree-based method in predicting survival function given a set of covariates. Random Forests which were introduced for specific analysis of right censored data and are used in predicting different kinds of survival metrics, including mortality, survival function and hazard function.Here we have implemented to competing risks using random forest.The method is fully non-parametric and can be used for selecting event-specific variables and for estimating the cumulative incidence function.

Individuals subject to competing risks are observed from study entry to the occurrence of the event of interest, a competing event, or often, before the individual can experience one of the events, that person is right censored. Formally, let T_i^o be the event time for the i th subject, $i = 1, \ldots..n$, and let δ_i° be the event type, $\delta_i^0 \in \{1, \ldots, J\}$, where $J \geqslant 1$. Let C_i^0 denote the censoring time for individual i such that the actual time of event T_i^o is unobserved and one only observes $T_i = \min\left(T_i^\theta, C_i^\sigma\right)$ and the event indicator $\delta_i = \delta_i^\theta I\left(T_i^\theta \leqslant C_i^\circ\right)$. When $\delta_i = 0$, the individual is said to be censored at T_i; otherwise if $\delta_i = j > 0$, the individual is said to have an event of type j at time T_i. The observed data are $(T_i, \delta_i, \mathbf{x}i)\, 1 \leqslant i \leqslant n$ where $\mathbf{x}_i$ is a p dimensional vector of covariates.

The cause-specific hazard function for event j given covariates x is

$$\alpha_j(t \mid \mathbf{x}) = \lim \Delta t \to 0 \frac{\mathbb{P}\left\{t \leqslant T^e \leqslant t + \Delta t, \delta^0 = j \,|T^e \geqslant t, \mathbf{x}|\right.}{\Delta t} = \frac{fj(t \mid \mathbf{x})}{S(t \mid \mathbf{x})}$$

Where, $S(t \mid \mathbf{x}) = \mathbb{P}\left\{T^0 \gg t|\mathbf{x}|\right.$ is the event-free survival probability function given $\mathbf{x}$. The cause-specific hazard function describes the instantancous risk of event j for subjects that currently are event-free. The probability of an event is determined using the cumulative incidence function (CIF), defined as the probability of experiencing an event of type j by time t. The CIF and cause-specific hazard function are related according to

$$F_j(t \mid \mathbf{x}) = \int_0^t S(s- \mid \mathbf{x})\alpha_j(s \mid \mathbf{x})\mathrm{d}s = \int_0^1 \exp\left(-\int_0^s \sum_{t=1}^J \alpha_l(u \mid \mathbf{x})\mathrm{d}u\right)\alpha_j(s \mid \mathbf{x})\mathrm{d}s$$

A covariate that reduces the cause-specific hazard of a competing risk increases the event-free survival probability and thereby indirectly increases the cumulative incidence of event j. Thus, covariates found to change the t-year risk of event j are those that change the cause-specific hazard function of event j and those that change the cause-specific hazard functions of the competing risks.

In addition to estimating the CIF, a 1D summary of the cumulative incidence re-

ferred to as the expected number of life years lost due to cause j is proposed in the paper. In right-censored data, it is not feasible to get a reliable estimate of the expected lifetime. Therefore, for a fixed time point τ, we consider the restricted mean lifetime conditional on x : $\int_0^\tau S(t \mid \mathrm{x})\mathrm{d}t$. The truncation time point τ is chosen such that the probability of being uncensored at τ is bounded away from zero: $P\left(C_i^o > \tau\right) \geqslant \epsilon > 0$. This relation is extended to the case with covariates and note the relation $S(t \mid \mathbf{x}) + \sum_{i=1}^{\prime} F_i(t \mid \mathbf{x}) = 1$. which holds for all values $t \leqslant \tau$ and all $\mathbf{x}$. The expected number of years lost before time τ is

$$L(\tau \mid \mathbf{x}) = \tau - \int_0^\tau S(t \mid \mathbf{x})\mathrm{d}t = \int_0^\tau \sum_{l=1}^{J} F_l(t \mid \mathbf{x})\mathrm{d}t$$

The summary value is $M_j(r \mid \mathbf{x}) = \int_0^t F_j(t \mid \mathbf{x})\mathrm{d}t$, which the above shows equals the expected number of life years lost due to cause j before time τ. We shall also call $M_j(\tau \mid x)$ the cause- j mortality.

2 COMPETING RISK FORESTS

In RSF ,each tree is grown using an independent bootstrap sample of the learning data using random feature selection at each node. RSF trees are generally grown very deeply with many terminal nodes (the ends of the tree). Trees in competing risk forests differ in the splitting rule and estimated values. To grow a competing risk forest, a single competing risk tree is grown in each bootstrap sample. The splitting rules are either event specific, or combine event-specific splitting rules across the J events.

2.1 Terminal Node estimators:

The estimators are calculated within the terminal node of each RSF tree and then aggregated to form the ensemble.

1.Eventspecificensemble :

Let $(T_i, \delta_i, \mathbf{x}i)\, 1 \leqslant i \leqslant n$ denote the learning data.A RSF tree is grown using an independent bootstrap sample of the learning data.Let $c_{i,b}$ be the number of times case i occurs in bootstrap sample b. To define the CIF for the bth trees, take a case's covariate x and drop it down the tree. Let $h_b(x)$ denote the indices for cases from the learning data whose covariates share the terminal node with x. Denoting node specific event counts by $N_{j,t}(t \mid \mathbf{x}) = \sum_{i \in h_{b,(\mathbf{x})}} c_{i,b} I\left\{T_i \leq t, \delta_i = j\right\}$ and the number

at risk by $Y_b(t \mid \mathbf{x}) = \sum_{i \in h_b,(\mathbf{x})} c_{i,b} I \mid T_i \geqslant t \Big\}$, we define x′s CIF as

$$\hat{F}j, b(t \mid \mathbf{x}) = \int 0^t \hat{S}b(u- \mid \mathbf{x}) Yb(u \mid \mathbf{x})^{-1} N_{j,b}(\, \mathrm{d}u \mid \mathbf{x})$$

where $\hat{S}_b(t \mid \mathbf{x}) = \prod u \leq t \left(1 - \sum_j N_{j,b}(\mathrm{du} \mid \mathbf{x}) / Y_b\Big(u \mid \mathbf{x}\Big)\right)$ is x′s Kaplan-Meier estimate of event-free survival The ensemble estimates of the CIF and the cause- f mortality, respectively, equal

$$\bar{F}j(t \mid \mathbf{x}) = \frac{1}{B} \sum_{b=1}^{B} \hat{F}_{j,b}(t \mid \mathbf{x}), \quad \bar{M}_j(\tau \mid \mathbf{x}) = \int_0^\tau \bar{F}_j(t \mid \mathbf{x}) \mathrm{dt} := \frac{1}{B} \sum_{b=1}^{B} \hat{M}_{j,b}(\tau \mid \mathbf{x})$$

For reporting an internal error rate, out-of-bag (OOB) ensembles are used.The OOB data are used to construct the OOB ensemble. Let $\mathcal{O}_i, \subset \{1, \ldots, B\}$ be the index set of trees where $c_{i,b} = 0$; i.e. O_i, records trees where case i is OOB. The OOB ensemble estimates of the CIF and the cause j mortality are, respectively, given by

$$\bar{F}_j^{\mathrm{oob}}\left(t \mid \mathbf{x}_i\right) = \frac{1}{|\mathcal{O}_i|} \sum_{h \in \mathcal{O}_i} \hat{F}_{j,b}\left(t \mid \mathbf{x}_i\right)$$

$$\bar{M}_j^{\mathrm{oob}}\ \left(\tau \mid \mathbf{x}_i\right) = \int_0^\tau \bar{F}_j^{\mathrm{oob}}\ \left(t \mid \mathbf{x}_i\right) \mathrm{d}t := \frac{1}{|\mathcal{O}_i|} \sum_{h \in \mathcal{O}_i} \hat{M}_{j,b}\left(\tau \mid \mathbf{x}_i\right)$$

The OOB predicted value for a case does not use event time outcome information for that case, and, therefore, because it is a cross-validation based estimator, it can be used for estimation of the prediction error.
2.Event free ensemble: The forest event-free survival is estimated using the ensemble.

$$\bar{S}\left(\tau \mid \mathbf{x}_i\right) \sum_{b=1}^{B} \hat{S}_b(t \mid \mathbf{x}) / B$$

2.2 Splitting Rule:(Generalised log rank test)

In the setting with competing risk, this is a test of the null hypothesis $H_0 : \alpha_{\mu l}(t) = \alpha_j(t)$ for all $t \leqslant \tau$. The test is based on the weighted difference of the cause-specific Nelson-Aalen estimates in the two daughter nodes. Specifically, for a split at the value c for variable x, the splitting rule is

$$L_j^{LR}(x, c) = \frac{1}{\hat{\sigma} j^{LR}(x, c)} \sum k = 1^m W_j\left(t_k\right) \left(d_{j.l}\left(t_k\right) - \frac{d_j\left(t_k\right) Y_1\left(t_k\right)}{Y\left(t_k\right)}\right)$$

where the variance estimate is given by,

$$\left(\hat{\sigma} j^{\mathrm{LR}}(x, c)\right)^2 = \sum k = 1^m W_j\left(t_k\right)^2 d_j\left(t_k\right) \frac{Y_l\left(t_k\right)}{Y\left(t_k\right)} \left(1 - \frac{Y_l\left(t_k\right)}{Y\left(t_k\right)}\right) \left(\frac{Y\left(t_k\right) - d_j\left(t_k\right)}{Y\left(t_k\right) - 1}\right)$$

3 Combined Regression Strategy (COBRA)

COBRA which stands for Combined Regression it is a boosting method used to combine multiple weak learners. We can combine multiple initial estimators of the regression function instead of building a linear or convex optimized function over a selection of basic estimators. We can use them as a collective indicator of the proximity between the training data and test observation. This local distance approach will be fast and efficient. Which performs asymptotically in the L^2 sense as the best combination of the basic estimators in the collective. Given a collection of basic estimators $r_1, .., r_M$, the idea behind this combining method is an "unanimity" concept. It creates a prediction mapping for each weak learner on the training data. These are then used while predicting on the test data to find existing data points that are close to the considered point. At each time t we have a new observation x_t, and we compute the K experts predictions $p_1(x_t), p_2(x_t)), ..., p_k(x_t))$. Then, the idea is to average realizations of y, not used to generate the experts, that have predictions in the same neighbourhood (in the Euclidean sense) of $p_1(x_t), p_2(x_t)), ..., p_k(x_t))$. The step of searching for realizations in these neighbourhoods is called the consensus step.

4 Prediction Performance:

o assess prediction performance, the concordance index and the prediction error defined by the integrated Brier score (BS) is used.The concordance index (C-index) is related to the area under the receiver operating characteristic curve and estimates the probability that, in a randomly selected pair of cases, the case that fails first had a worse predicted outcome. The BS is the squared difference between actual and predicted outcome. Individuals are ranked by ensemble cause-j mortality. The time-truncated concordance index for competing risks, which in our setting is

$$C_j(\tau) = \mathbb{P}\left\{\bar{M} j\left(\tau \mid \mathbf{x}_i\right) > \bar{M}_j\left(\tau \mid \mathbf{x}_i\right) \mid T_i^o \leqslant \tau, \delta_i^o = j \text{ and } \left(T_i^o < T_{i'}^o \text{ or } \delta_{i'}^o \neq j\right)\right\}$$

Thus, the ensemble prediction of the cumulative incidence is concordant with the outcome if either the case with the higher cause-j mortality has event j before the

other case has an event of cause j or if the other case has a competing event. The time-dependent BS and its integral (IBS) to assess the performance of the ensemble CIF is also considered:

$$\mathrm{IBS}_j(\tau) = \int_0^\tau \mathrm{BS}_j(t)\mathrm{d}t = \int_0^\tau \mathrm{E}\left\{ I\left\{T_i^o \leqslant t, \delta_i = j\right\} - \hat{F}_j(t \mid \mathbf{X})\right\}^2 \mathrm{d}t.$$

5 Implementation & Results

The implementation majorly contains two parts: Random Survival Forests with competing risks and COBRA. Both the parts are implemented from scratch. The RFS implementation is based on the research paper "Random Survival forests for Competing Risks" by Ishwaran et al. and the COBRA implementation is based on the research paper "COBRA: A Combined Regression Strategy" by Biau et al. RFS and cobra are implemented as python classes named *RandomSurvivalForest* and *COBRA* respectively. Below are cumulative survival functions learned at some of the leaf nodes of a decision tree in the RFS trained.

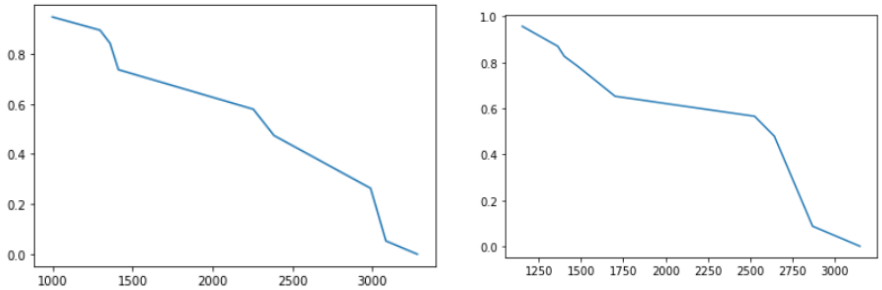

Figure 1: Cumulative survival functions learned at 2 leaf nodes of a decision tree in the RFS trained.

References

[1] Gérard Biau et al. "COBRA: A combined regression strategy." In: *Journal of Multivariate Analysis* 146 (2016). Special Issue on Statistical Models and Methods for High or Infinite Dimensional Spaces, pp. 18–28. ISSN: 0047-259X. DOI: https://doi.org/10.1016/j.jmva.2015.04.007. URL: https://www.sciencedirect.com/science/article/pii/S0047259X15000950.

[2] Arabin Dey and Anshul Juneja. "Some variations on Random Survival Forest with application to Cancer Research." In: (Sept. 2017).

[3] Hemant Ishwaran et al. "Random survival forests for competing risks." In: *Biostatistics* 15.4 (Apr. 2014), pp. 757–773. ISSN: 1465-4644. DOI: 10.1093/biostatistics/kxu010. eprint: https://academic.oup.com/biostatistics/article-pdf/15/4/757/28922269/kxu010.pdf. URL: https://doi.org/10.1093/biostatistics/kxu010.