h_da
HOCHSCHULE DARMSTADT
UNIVERSITY OF APPLIED SCIENCES

Prediction Uncertainty in Weighted Machine Learning for Survival Data:

Concepts and Application to Organ Transplantation Registry Data

Author: Rehan Butt | Referent: Prof. Dr. Antje Jahn | Coreferent: Prof. Dr. Gunter Grieser

Darmstadt University of Applied Sciences, Faculties of Mathematics and Natural Sciences & Computer Science

Motivation

Event time data analysis is crucial in statistical research, especially in medical and survival studies. Dealing with censored data, where the event of interest is not observed for all subjects, poses significant challenges for predictive modeling. The Inverse Probability of Censoring Weights (IPCW) method offers a solution by generating unbiased predictions from censored event time data [2]. It does so by transforming the problem into a classification framework and producing bootstrap samples in conjunction with weighted resampling. However, accurately estimating the variance of these predictions remains an open problem. Currently, aside from the bootstrap method, there are no methods available in the literature to estimate the variance of predictions calculated from the weighted bootstrap samples produced by the IPCW approach. The bootstrap method, while unbiased, is computationally intensive and may not be feasible in all practical situations.

Therefore, this work is dedicated to developing a nonparametric method for variance estimation tailored to IPC-weighted classification models. By introducing a new estimator, we aim to fill this gap and provide a more efficient and accurate tool for assessing the uncertainty of predictions in the context of censored event time data.

Methodology

In this work, we focus on classification models that are bagged learners based on decision trees. Bagging, or bootstrap aggregating, involves training multiple base learner on different subsets of the data and aggregating their predictions to improve overall performance. To address the challenge of estimating the variance of predictions from IPC-weighted classification models with censored event time data, we developed the Infinitesimal-Jackknife-after-weighted-Bootstrap-unbiased (IJK-AWB-U) estimator. This new method extends existing variance estimation techniques to accommodate IPC-weighted resampling and includes a bias correction for finite bootstrap samples.

Development of the IJK-AWB-U Estimator

- Extension to IPC-Weighted Resampling: Traditional variance estimators like the Infinitesimal Jackknife for Bagged Learners (IJK) are designed for unweighted resampling [3] and are inadequate for IPC-weighted bootstrap samples used in handling censored data. We modified the IJK method to incorporate IPC weights, allowing it to provide unbiased variance estimates in this context.
- Bias Correction: When the number of bootstrap samples (B) is finite, the standard Infinitesimal-Jackknife-after-weighted-Bootstrap (IJK-AWB) estimator can be biased. We derived a bias correction term specific to IPC-weighted resampling, resulting in the IJK-AWB-U estimator that remains unbiased even with finite B.

Simulation Study Following the ADEMP Framework

We conducted an extensive simulation study structured according to the ADEMP (Aim, Data-generating mechanism, Estimand, Methods, Performance measures) framework to evaluate the performance of the IJK-AWB-U estimator.

- Aim: Assess the performance of the IJK-AWB-U estimator in estimating the standard deviation ($\hat{\boldsymbol{\sigma}}$) of predictions ($\hat{\boldsymbol{S}}$) from IPC-weighted bagged decision trees (DTBC).
- Data-Generating Mechanism:
  - We simulated 1000 datasets with varying sample sizes, censoring and event proportions.
  - Event times were generated using Weibull distributions and censoring was introduced to simulate right-censored data. The datasets contains 5 covariates.
- Estimand: The true standard deviation of the prediction, derived from the bagged decision tree model, for a specific individual with predetermined covariate values.
- Methods compared:
  - Infinitesimal-Jackknife-after-weighted-Bootstrap-unbiased (IJK-AWB-U), proposed method
  - Infinitesimal-Jackknife-after-weighted-Bootstrap (IJK-AWB), without bias correction
  - Nonparametric Bootstrap (Boot): Serves as the benchmark, described in the literature as the gold standard for variance estimation, but is computationally intensive [1].
  - Jackknife-after-Bootstrap (JK-AB-U): Known to be unbiased for variance estimates of predictions from a bagged learner under unweighted bootstrap samples [3].
- Performance Measures:
  - Mean Relative Bias (MRB) quantifies the average relative difference between the mean estimated ($\hat{\boldsymbol{\sigma}}$) and the empirical standard deviation of predictions ($\boldsymbol{s}(\hat{\boldsymbol{S}})$), expressed as a percentage. It assesses the bias of the variance estimation methods in comparison to the true variability in the predictions.
  - The Coefficient of Variation (CV) measures the relative variability of the estimated standard deviations ($\hat{\boldsymbol{\sigma}}$) across all simulations.

Application on TxReg Dataset

To assess the practical applicability of our proposed variance estimation method IJK-AWB-U, we applied the variance estimators to the TxReg dataset, which contains retrospective and anonymized data from the German Transplant Registry. The cleaned dataset comprises 15,786 observations, including 19 predictive variables. Our objective was to estimate the standard deviation ($\hat{\boldsymbol{\sigma}}$) of the predictions ($\hat{\boldsymbol{S}}$) generated from IPC-weighted bagged decision trees (DTBC).

Simulation Study Results

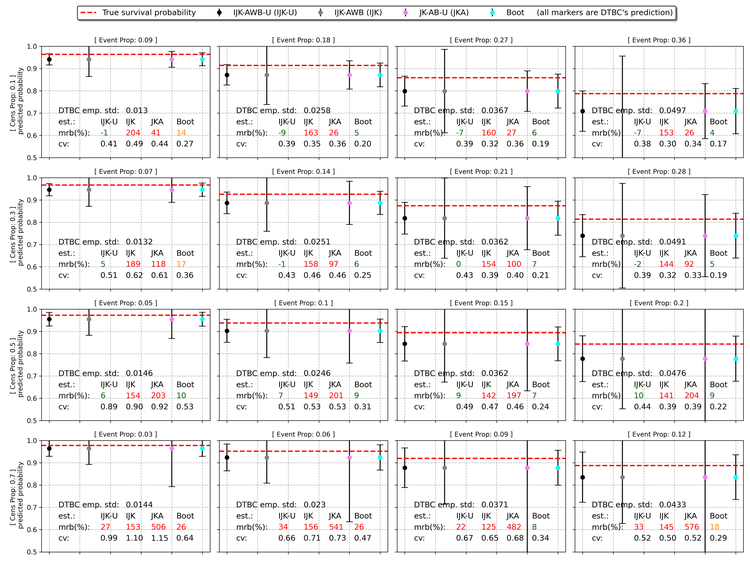

Abbildung 1:Simulation study results for variance estimator's performance with $\boldsymbol{n_{\text{train}} = 1999}$, $\boldsymbol{B = 1000}$. DTBC's prediction represents the mean estimated survival probability $\hat{\boldsymbol{S}}(\boldsymbol{\tau}|\boldsymbol{X_{pred}})$ and the DTBC emp. std stands for $\boldsymbol{s}\left(\hat{\boldsymbol{S}}\right)$ over the **1000** simulation runs. The **4** error bars correspond to $\hat{\boldsymbol{S}}(\boldsymbol{\tau}|\boldsymbol{X_{pred}}) \pm \mathbf{1.96} \cdot \hat{\boldsymbol{\sigma}}$, which provides an approximate **95%** confidence interval of the mean prediction. The mean estimated standard deviations $(\hat{\boldsymbol{\sigma}})$ for each error bar is based on an estimator from the legend. If $|\text{mrb}(\%)| \leq \mathbf{10}$, it is colored green, if $\mathbf{10} < |\text{mrb}(\%)| \leq \mathbf{20}$, it is colored yellow, otherwise its colored red.

The above figure illustrates the performance of various variance estimators for the IPC-weighted bagged decision trees (DTBC) model under different simulation settings, specifically focusing on the event and censoring proportions. Each subplot corresponds to a unique combination of event and censoring proportions, indicated by the Ëvent Propänd Cens Proplabels at the top of each panel.

Our simulation study demonstrated that the IJK-AWB-U estimator provides reliable and accurate variance estimates for the DTB) model's predictions, especially under low to moderate censoring proportions and larger training sample sizes. The estimator effectively corrects the bias present in the IJK-AWB estimator and achieves a favorable balance between accuracy and computational efficiency compared to the nonparametric Bootstrap, which, despite its stability, is computationally intensive. However, under high censoring proportions, the IJK-AWB-U tends to overestimate the variance and exhibit greater variability, with the event proportion having a more pronounced effect on the Coefficient of Variation (CV) than on the Mean Relative Bias (MRB). Additionally, the JK-AB-U estimator was generally less reliable due to its consistent overestimation of variance. These findings highlight the importance of considering both censoring and event proportions when applying the IJK-AWB-U estimator, establishing it as the preferred method for variance estimation in our context.

Application on TxReg Dataset Results

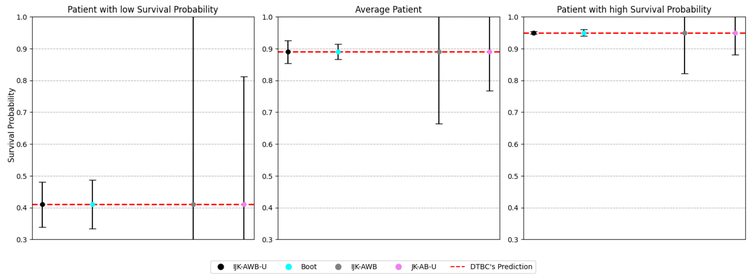

Abbildung 2:Variance Estimates on predicted Survival Probabilities at $\boldsymbol{\tau = 3}$ years of 3 Patients from the TxReg Dataset. The 4 error bars correspond to $\hat{\boldsymbol{S}}\left(\mathbf{1095} \mid \boldsymbol{X_{patient(i)}}\right) \pm \mathbf{1.96} \cdot \hat{\boldsymbol{\sigma}}$, which provides an approximate 95% confidence interval of the prediction. The estimated standard deviations $(\hat{\boldsymbol{\sigma}})$ for each error bar is based on an estimator from the legend.

To illustrate the capabilities of the variance estimators applied to the DTBC model's predictions, we selected three representative patients from the dataset whose three-year survival probabilities were predicted. The patient with the lowest survival probability experienced an event after approximately 0.4 years. The average patient was censored after approximately 5 years, and the patient with the highest survival probability was censored after approximately 3.6 years.

The above figure reveals a clear trend: all four estimators exhibit the widest confidence intervals for the patient with the lowest survival probability (left panel), indicating the highest uncertainty. As survival probability increases (middle and right panels), the confidence intervals become progressively narrower for each method, suggesting increased confidence in the predictions. This pattern is consistent across all estimators, reflecting the DTBC model's greater stability in estimating higher survival probabilities.

The IJK-AWB-U estimator provides narrower confidence intervals that closely align with the Bootstrap estimator, which was identified as the most stable approach in the simulated settings. This alignment suggests that the IJK-AWB-U estimator may offer a reliable approximation of variance in practical applications. In contrast, the IJK-AWB and JK-AB-U estimators consistently produce wider confidence intervals, particularly for lower survival probabilities, indicating a tendency to overestimate variance. This overestimation underscores the limitations of these methods for practical use in variance estimation within the DTBC model with survival data.

Conclusion and Future Work

This study introduced the IJK-AWB-U estimator, demonstrating its ability to provide reliable and accurate variance estimates for IPC-weighted bagged decision trees, particularly under low to moderate censoring and larger sample sizes. Through extensive simulations and application to the TxReg dataset, the IJK-AWB-U outperformed traditional methods like the nonparametric Bootstrap and Jackknife-after-Bootstrap by balancing accuracy with computational efficiency, establishing it as the preferred variance estimation method in this context.

Future research should extend the application of the IJK-AWB-U estimator to a wider range of machine learning models, such as gradient-boosting and neural networks, and across diverse datasets in fields like finance, engineering, and epidemiology. Additionally, exploring hybrid methods that integrate the strengths of IJK-AWB-U with other variance estimation techniques could further enhance predictive accuracy and reliability, broadening its applicability and validating its versatility in handling various censored data scenarios.

Bibliography

[1] Bradley Efron und Robert J. Tibshirani. *An Introduction to the Bootstrap*. Monographs on Statistics and Applied Probability 57. Boca Raton, Florida, USA: Chapman & Hall/CRC, 1993.

[2] Pablo Gonzalez Ginestet u. a. "Stacked Inverse Probability of Censoring Weighted Bagging: A Case Study In the InfCareHIV Register". In: *Journal of the Royal Statistical Society Series C: Applied Statistics* 70.1 (Jan. 2021), S. 51–65. ISSN: 0035-9254. DOI: 10.1111/rssc.12448. eprint: https://academic.oup.com/jrsssc/article-pdf/70/1/51/49158571/rssc_70_1_51.pdf. URL: https://doi.org/10.1111/rssc.12448.

[3] Stefan Wager, Trevor Hastie und Bradley Efron. *Confidence Intervals for Random Forests: The Jackknife and the Infinitesimal Jackknife*. 2014. arXiv: 1311.4555 [stat.ML].