Statistica Sinica **32** (2022), 1245-1267
doi:https://doi.org/10.5705/ss.202020.0263

CONSISTENCY OF SURVIVAL TREE AND FOREST MODELS: SPLITTING BIAS AND CORRECTION

Yifan Cui, Ruoqing Zhu, Mai Zhou and Michael Kosorok

National University of Singapore, University of Illinois at Urbana-Champaign
University of Kentucky and University of North Carolina at Chapel Hill

Abstract: Random survival forests and survival trees are popular models in statistics and machine learning. However, there is a general lack of understanding regarding the consistency, splitting rules, and influence of the censoring mechanism. In this study, we investigate the statistical properties of existing methods from several interesting perspectives. First, we show that traditional splitting rules with censored outcomes rely on a biased estimation of the within-node failure distribution. To exactly quantify this bias, we develop a concentration bound of the within-node estimation based on samples that are not independent and identically distributed, and apply it to the entire forest. Second, we analyze the entanglement between the failure and censoring distributions caused by univariate splits, and show that without correcting the bias at an internal node, survival tree and forest models can still enjoy consistency under suitable conditions. In particular, we demonstrate this property under two cases: a finite-dimensional case, where the splitting variables and cutting points are chosen randomly, and a high-dimensional case, where the covariates are weakly correlated. Our results also apply to an independent covariate setting, which is commonly used in the random forest literature for high-dimensional sparse models. However, it may be unavoidable that the convergence rate depends on the total number of variables in the failure and censoring distributions. Third, we propose a new splitting rule that compares bias-corrected cumulative hazard functions at each internal node. We show that the rate of consistency of this new model depends only on the number of failure variables. We perform simulation studies to confirm that the proposed bias-correction can substantially benefit the prediction error.

Key words and phrases: Adaptive concentration, bias correction, consistency, random forests, survival analysis.

1. Introduction

Random forests (Breiman (2001)) are among the most popular and powerful machine learning tools. The main advantage of tree-based models (Breiman et al. (1984)) is their nonparametric nature. Although there has been a surge of

---

Corresponding author: Ruoqing Zhu, Department of Statistics, University of Illinois at Urbana-Champaign, Champaign, IL 61820, USA. E-mail: rqzhu@illinois.edu.