The Journal of Open Source Software

aorsf: An R package for supervised learning using the oblique random survival forest

Byron C. Jaeger [1], **Sawyer Welden**[1], **Kristin Lenoir**[1], **and Nicholas M Pajewski**[1]

1 Wake Forest University School of Medicine

DOI: 10.21105/joss.04705

Software

- Review
- Repository
- Archive

Editor: Daniel S. Katz

Reviewers:

- @danielskatz

Submitted: 06 August 2022
Published: 27 September 2022

License
Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License (CC BY 4.0).

Summary

Risk prediction is a type of supervised learning where the goal is to predict the probability that a person will experience an event within a specific amount of time. This kind of prediction may be useful in clinical settings, where identifying patients who are at high risk for experiencing an adverse health outcome can help guide strategies for prevention and treatment. The oblique random survival forest (RSF) is a supervised learning technique that has obtained high prediction accuracy in general benchmarks for risk prediction (Jaeger et al., 2019). However, computational overhead and a lack of tools for interpretation make it difficult to use the oblique RSF in applied settings.

`aorsf` is an R package with fast algorithms to fit and interpret oblique RSFs, allowing users to obtain an accurate risk prediction model without losing efficiency or interpretability. `aorsf` supports exploration and customization of oblique RSFs, allowing users to select the fitting procedure for an oblique RSF or supply their own procedure (see examples provided in documentation for `orsf_control_custom()`). Whereas existing software for oblique RSFs does not support estimation of variable importance (VI), `aorsf` provides multiple techniques to estimate VI with the added flexibilty of allowing users to supply their own functions where applicable (see documentation for `orsf_vi()`).

Statement of need

The purpose of `aorsf` is to allow oblique RSFs to be fit and interpreted efficiently. The target audience includes both **analysts** aiming to develop an accurate risk prediction model (e.g., see Segar et al. (2021)) and **researchers** who want to conduct experiments comparing different techniques for fitting oblique RSFs (e.g., see Katuwal, Suganthan, & Zhang (2020)).

Related software

The `obliqueRF` and `RLT` R packages support oblique random forests (RFs) for classification and regression, but not survival. The `ranger`, `randomForestSRC`, and `party` packages support axis-based RSFs (see Background section) but not oblique RSFs. The `obliqueRSF` R package fits oblique RSFs, but has high computational overhead, provides a limited set of tools to interpret oblique RSFs, and does not enable customization of routines used to fit oblique RSFs.

Background

Random forests (RFs) are large sets of de-correlated decision trees (Breiman, 2001). Trees in the RF may be axis-based or oblique. Axis-based trees split data using a single predictor, creating decision boundaries that are perpendicular or parallel to the axes of the predictor space.

Oblique trees split data using a linear combination of predictors, creating decision boundaries that are neither parallel nor perpendicular to the axes of their contributing predictors. The increased flexibility of their decision boundaries allows oblique trees to produce more controlled partitions of a predictor space compared to their axis-based counterparts (**Figure 1**). However, the added flexibility has a computational cost, as the number of potential oblique decision boundaries to search through may be much larger than the number of axis-based boundaries.

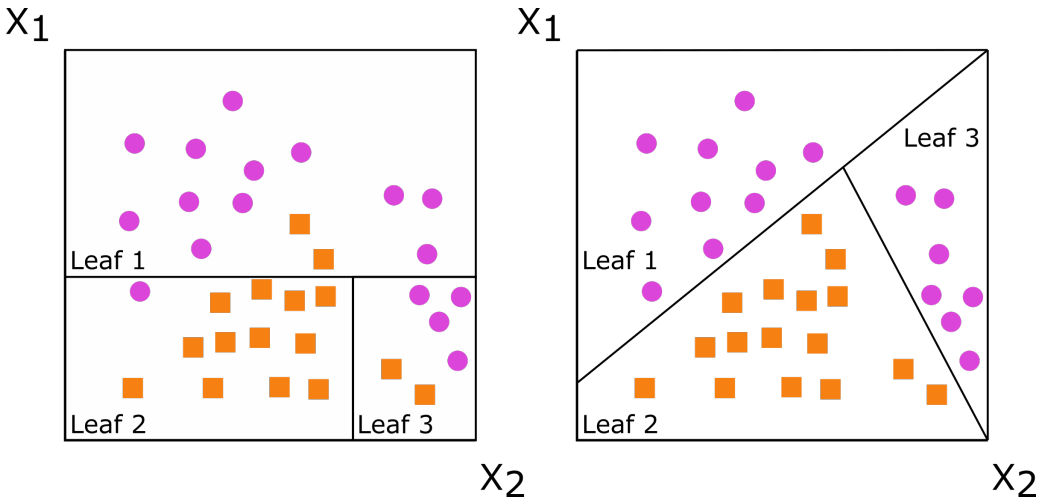

Figure 1: Decision trees with axis-based splitting (left) and oblique splitting (right). Cases are orange squares; controls are purple circles. Both trees partition the predictor space defined by variables X1 and X2, but the oblique splits do a better job of separating the two classes.

Censoring

A common challenge in risk prediction is censoring. The term 'censor' indicates partial observation of something. Censoring occurs often in medical studies that track the elapsed time from a baseline visit to the occurrence of an event. For example, censoring can occur if study coordinators lose contact with a participant or the study concludes before the participant experiences the event.

Decision trees can engage with censored outcomes by recursively partitioning data using standard descriptive tests (e.g., the log-rank test) and applying estimation techniques for censored outcomes to generate predicted values in leaf nodes, such as the survival curve or the cumulative hazard function (Ishwaran, Kogalur, Blackstone, & Lauer, 2008).

Newton Raphson scoring

The efficiency of `aorsf` is primarily attributed to its use of Newton Raphson scoring to identify linear combinations of predictor variables. `aorsf` uses the same approach as the `survival` package to complete this estimation procedure efficiently. Full details on the steps involved have been made available by Therneau (2022). Briefly, a vector of estimated regression coefficients, $\hat{\beta}$, is updated in each step of the procedure based on its first derivative, $U(\hat{\beta})$, and second derivative, $H(\hat{\beta})$:

$$\hat{\beta}^{k+1} = \hat{\beta}^{k} + U(\hat{\beta} = \hat{\beta}^{k})\, H^{-1}(\hat{\beta} = \hat{\beta}^{k})$$

While it is standard practice in statistical modeling to iteratively update $\hat{\beta}$ until a convergence threshold is met, the default approach in `aorsf` only completes one iteration (see `orsf_control_fast()`). The rationale for this approach is based on two points. First, while

completing more iterations reduces bias in the regression coefficients, general benchmarking experiments have found it does not improve the discrimination or Brier score of the oblique RSF (Jaeger et al., 2022). Second, computing U and H requires computation and exponentiation of the vector $X\hat{\beta}$, where X is the matrix of predictor values, but these steps can be skipped on the first iteration if an initial value of $\hat{\beta} = 0$ is chosen, allowing for a reduction in required computation.

Variable importance

Estimation of VI is a common technique used for interpretation of RFs. Permutation, a standard approach to estimate VI, measures how much a RF's prediction accuracy decreases after randomly permuting values of a given variable. If the variable is important, de-stabilizing decisions based on the variable should reduce the RF's prediction accuracy.

`aorsf` includes multiple functions to estimate VI, including permutation VI and a faster approach to estimate VI using analysis of variance (Menze, Kelm, Splitthoff, Koethe, & Hamprecht, 2011). In addition, `aorsf` includes a novel technique for VI that is designed for compatibility with oblique decision trees: 'negation importance.' Similar to permutation, negation VI measures the reduction in an oblique RFs prediction accuracy after de-stabilizing decisions based on a variable. However, instead of permuting values of a variable, negation VI flips the sign of all coefficients linked to a variable (i.e., it negates them). As the magnitude of a coefficient increases, so does the probability that negating it will change the oblique RF's predictions.

Benchmarking

General benchmarks of prediction accuracy, computational efficiency, and estimation of VI using `aorsf` and several other software packages are available (Jaeger et al., 2022). In these benchmarks, `aorsf` has matched or exceeded the prediction accuracy of `obliqueRSF` while running hundreds of times faster. Additionally, simulation studies have found that negation VI improves the oblique RSF's ability to discriminate between signal and noise predictors compared to existing VI techniques.

Acknowledgements

The development of this software was supported by the Center for Biomedical Informatics at Wake Forest University School of Medicine and the National Center for Advancing Translational Sciences (NCATS), National Institutes of Health (NIH), through Grant Award Number UL1TR001420. Dr. Pajewski was additionally supported by grant award P30 AG021332 from the NIH. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. doi:10.1023/A:1010933404324

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., & Lauer, M. S. (2008). Random survival forests. *The Annals of Applied Statistics*, *2*(3), 841–860. doi:10.1214/08-AOAS169

Jaeger, B., Long, D. L., Long, D. M., Sims, M., Szychowski, J. M., Min, Y.-I., Mcclure, L. A., et al. (2019). Oblique random survival forests. *The Annals of Applied Statistics*, *13*(3), 1847–1883. doi:10.1214/19-AOAS1261

Jaeger, B., Welden, S., Lenoir, K., Speiser, J. L., Segar, M. W., Pandey, A., & Pajewski, N. M. (2022). Accelerated and interpretable oblique random survival forests. arXiv. doi:10.48550/ARXIV.2208.01129

Katuwal, R., Suganthan, P. N., & Zhang, L. (2020). Heterogeneous oblique random forest. *Pattern Recognition*, *99*, 107078. doi:10.1016/j.patcog.2019.107078

Menze, B. H., Kelm, B. M., Splitthoff, D. N., Koethe, U., & Hamprecht, F. A. (2011). On oblique random forests. *Joint european conference on machine learning and knowledge discovery in databases* (pp. 453–469). Springer.

Segar, M. W., Jaeger, B. C., Patel, K. V., Nambi, V., Ndumele, C. E., Correa, A., Butler, J., et al. (2021). Development and validation of machine learning–based race-specific models to predict 10-year risk of heart failure: A multicohort analysis. *Circulation*, *143*(24), 2370–2383. doi:10.1161/circulationaha.120.053134

Therneau, T. (2022, April). Survival package source code documentation. Retrieved from https://github.com/therneau/survival/blob/5440691d44abea537b08aeb60153a31654d66a9b/noweb