Hypertension statistics for US adults: an open-source web application for analysis and visualization of US NHANES data

Reviewer: 1

Comments to the Author

This manuscript adds two contributions to the literature of random forests. 1) The development of a computational efficient oblique random forest algorithm for right censored/survival data and 2) An alternative approach to estimate variable important using what is called “negation variable importance”. The authors apply their approach on both simulated and real data sets. Overall, the manuscript is well written, and the manuscript is in line with JCGS’s scopes and aims. The authors do an excellent job outlining their approach and planned future directions.

Below are a few minor comments that I would like to have addressed.

1. The validity of the Cox partial likelihood function requires the non-informative censoring assumption (true survival time and censoring time are independent given the features X). I did not see this explicitly stated as an assumption in Section 2.3.

**Response**: We agree with the reviewer and have modified the manuscript to include this information. Specifically, because non-informative censoring presents a possible limitation for our approach, we have added the following text to Section 5.2:

“The validity of the Cox partial likelihood function depends on an assumption of non-informative censoring (*i.e.*, true survival time and censoring time are independent given the predictors). Thus, the oblique RSF may be improved by incorporating methods to account for informative censoring prior to identifying a linear combination of variables.”

1. For the simulations proposed in Section 3.2.3, the authors use the simsurv package. The underlying data generating model was not included so it is unclear how exactly the data were simulated. More clarity on the data generation procedure would be appreciated. I am guessing that the authors generated data under a PH model. However, it would be advantageous to see the performance of aosrf under data generated under a non-PH model.

**Response**:

1. It would be good to see how the various methods perform under high censoring, especially since several of the datasets that are included have censoring > 80%.

**Response**:

Reviewer: 2

Comments to the Author

1. Lines 42-49 on page 6: I would like to request further clarification on the process for obtaining linear combinations of predictor variables through the use of Newton-Raphson scoring.

**Response**:

1. Equation in line 13 on page 7, I have some doubts as to whether the sign on the right-hand side should be negative (positive in Fisher scoring with the information matrix). Furthermore, it may be better to bold the beta symbol in the equation to be consistent with the previous formula.

**Response**:

Associate Editor’s Comments to the Author:

This paper improves the computational efficiency and interpretation of oblique random survival forests (RSFs). The paper is clearly written. The numerical experiments are exceptionally well-planned, executed, and presented. Two reviewers’ and below comments (mostly minor) need to be addressed.

1. The package associated with this manuscript, aorsf, is one of the supported engines by tidymodels (<https://urldefense.com/v3/__https://parsnip.tidymodels.org/reference/details_rand_forest_aorsf.html__>;!!GA8Xfdg!yI2a4uGp0KH2vYEOZyxWCuA5MK9PUwHaWFlkvdEAr73zIbT1eicot7N-Mj6CaKHRP3OUkh030vi3ORph5zQfVgCSRYXfHA$ ) and is widely used. This paper has potentially high impact.

**Response**:

1. The numerical experiments are extensive and well-planned. They set a high standard for JCGS articles. It boasts 21 data sets with a total of 35 learning tasks. Performances of learners are compared using the Bayesian linear mixed model.

**Response**:

1. Conceptually, how does oblique RSF compare to first creating linear (PCA, PLS) and nonlinear features (UMP) and then applying standard axis RSF? Authors don’t need to make extra comparisons. A brief explanation suffices.

**Response**:

1. page 7, line 13: What are the and functions? Score and Hessian? Please define.

**Response**:

1. page 7, line 23 and footnote: should be .

**Response**:

1. Page 12, line 2: number and density of layers means number and width of layers or number and dropout rates of layers?

**Response**:

1. For boosting, is the learning rate tuned? How are the neural networks tuned (depth and width of layers, drop out rate, etc)?

**Response**:

1. In general, the paper lacks details on how several competing learning methods are tuned. For reproducibility, (1) how the data sets are retrieved and pre-processed and (2) the code and script for numerical experiments need to be made public.

**Response**:

1. Please briefly explain what nested cross-validation is.

**Response**: