PA

Comparing Random Forest with Logistic Regression for Predicting Class-Imbalanced Civil War Onset Data: A Comment

Yu Wang

Department of Political Science, Unversity of Rochester, Rochester, NY , USA.
Email:ywang @ur.rochester.edu

Keywords:data analysis algorithms, forecasting, learning, random forests, boosting

Introduction

In an interesting and provocative paper, Muchlinski et al. ( ) make an important contribution by emphasizing the significance of predictive accuracy and empirically training a highly accurate random forest model. With an area under the curve (AUC) of . , their random forest model outperforms by a large margin three leading logistic regression models: Fearon and Laitin ( ) with an AUC of . , Collier and Hoe ler ( ) with an AUC of . , and Hegre and Sambanis ( ) with an AUC of . . The improvement is dramatic, and the paper has quickly established itself in the machine learning/prediction-inclined community in our discipline (Cederman and Weidmann ; Cranmer and Desmarais).

Muchlinski et al. ( ) have emphasized in their paper the importance of cross validation in evaluating their model•s predictive accuracy and applied tenfold cross validation throughout to tune the parameters. When evaluating the performance of their model, however, the authors have veered away from this approach and used models trained with the whole dataset instead. This leads to several incorrect presentations and interpretations of their results. In this comment, I point out and correct this error with respect to cross validation. I also report better prediction results using AdaBoosted trees and gradient boosted trees.

Spot the Error

One way to quickly spot the error is to notice that while the reported AUC of random forest is . based on cross validation, the area under the dot…dash curve is substantially larger than . (Figure). For the purpose of comparison, I have added a dashed rectangle with a height of , a width of . (from x = 0.1 to x = 1), and an area of . . The real AUC as presented in Figure in the original article is . rather than . , and the model is trained with the entirety of the dataset.

To be sure, Muchlinski et al. ( ) have used cross validation to tune the parameters such as the number of variables to randomly sample as candidates for each split when constructing each tree. Once the parameters are selected, however, the authors trained the random forest model using the whole dataset. As the model is then used to predict samples that it has seen during the training process, it is no surprise that an AUC of . obtained this way is higher than . based on cross validation. The same error has a ected the receiver operating characteristic (ROC) curves and the separation plots for all the classifiers.

Author•s note I would like to thank Randall Stone, Curtis Signorino, Kevin Clarke, Jiebo Luo, Henry Kautz and Sally Thurston at the University of Rochester, the editor, the anonymous reviewer, and David Muchlinski. All remaining errors are my own. The replication materials (Wang) for all the figures and tables in this paper and in the online appendix are available at the Political Analysis dataverse site.

The replication materials (Wang) are available at the Political Analysis dataverse site.

Political Analysis ( )
vol. : …
DOI: . /pan. .

Corresponding author
Yu Wang

Edited by
Je Gill