The Journal of Open Source Software

iRF: extracting interactions from random forests

Sumanta Basu[1, a, b], **Karl Kumbier**[1, c], **James B. Brown**[c, d, e], **and Bin Yu**[c, f]

1 Denotes equal contribution **a** Department of Biological Statistics and Computational Biology, Cornell University **b** Department of Statistical Science, Cornell University **c** Statistics Department, University of California, Berkeley **d** Centre for Computational Biology, School of Biosciences, University of Birmingham **e** Molecular Ecosystems Biology Department, Biosciences Area, Lawrence Berkeley National Laboratory **f** Department of Electrical Engineering and Computer Sciences, University of California, Berkeley

DOI: 10.21105/joss.01077

Software

- Review
- Repository
- Archive

Submitted: 04 October 2018
Published: 05 December 2018

Summary

Random forests (Breiman, 2001) are a popular class of supervised learning models that have demonstrated impressive empirical success across a wide variety of problems. The predictive accuracy of random forests stems from their ability to learn high-order, non-linear interactions in large datasets. Although approaches exist for evaluating the importance of individual features in a fitted random forest, identifying interactions that drive predictive accuracy remains a challenge. This challenge is in large part due to the enourmous number of interactions that must be considered (i.e. there are $O(p^s)$ possible interactions of size s among p features) and the instability of random forest decision paths.

The iterative Random Forest algorithm (iRF), and corresponding `iRF` R package, take a step towards addressing these issues with a computationally tractable approach to search for important interactions in a fitted random forest (Basu, Kumbier, Brown, & Yu, 2018,Kumbier, Basu, Brown, Celniker, & Yu (2018)). Our algorithm grows a series of feature weighted random forests (Amaratunga, Cabrera, & Lee, 2008) to perform soft regularization on the model based on predictive features. We then search for prevalent interactions in the fitted random forest using a generalization of random intersection trees (Shah & Meinshausen, 2014). Finally, we assess the stability of recovered interactions by repeating this search across random forests trained on bootstrap samples of the data. The `iRF` R package combines these steps into a single workflow. It is based on the source codes from the R packages `randomForest` (Liaw & Wiener, 2002) and `FSInteract` (Shah & Meinshausen, 2014). A detailed vignette is available here.

Acknowledgements

This research was supported in part by grants NHGRI U01HG007031, ARO W911NF1710005, ONR N00014-16-1-2664, DOE DE-AC02-05CH11231, NHGRI R00 HG006698, DOE (SBIR/STTR) Award DE-SC0017069, DOE DE-AC02-05CH11231, and NSF DMS-1613002. We thank the Center for Science of Information (CSoI), a US NSF Science and Technology Center, under grant agreement CCF-0939370. Research reported in this publication was supported by the National Library Of Medicine of the NIH under Award Number T32LM012417. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. BY acknowledges support from the Miller Institute for her Miller Professorship in 2016-2017. SB acknowledges the support of UC Berkeley and LBNL, where he conducted most of his work on this paper as a postdoc. We thank P. Bickel and S. Shrotriya for helpful discussions and comments.

References

Amaratunga, D., Cabrera, J., & Lee, Y.-S. (2008). Enriched random forests. *Bioinformatics*, *24*(18), 2010–2014. doi:10.1093/bioinformatics/btn356

Basu, S., Kumbier, K., Brown, J. B., & Yu, B. (2018). Iterative random forests to discover predictive and stable high-order interactions. *Proceedings of the National Academy of Sciences*, 201711236. doi:10.1073/pnas.1711236115

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

Kumbier, K., Basu, S., Brown, J. B., Celniker, S., & Yu, B. (2018). Refining interaction search trhough signed iterative random forests. *bioRxiv*. doi:10.1101/467498

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R news*, *2*(3), 18–22.

Shah, R. D., & Meinshausen, N. (2014). Random intersection trees. *The Journal of Machine Learning Research*, *15*(1), 629–654.