Comments on: A random forest guided tour

Stefan Wager¹

Published online: 19 April 2016
© Sociedad de Estadística e Investigación Operativa 2016

1 Introduction

First of all, the authors deserve thanks for their numerous foundational contributions to the theory of random forests, and for this excellent review paper. This article will be a valuable resource to researchers who want to know what statisticians have to say about random forests, and who want to understand when and why forest-based methods can be expected to work.

Since their introduction by [Breiman \(2001\)](#), random forests have posed something of a conundrum to statisticians: on the one hand, they consistently achieve outstanding “out-of-the-box” empirical performance across numerous application areas; but, on the other hand, understanding how they achieve this performance from a theoretical perspective has been challenging. Biau, Scornet, and many others have taken considerable strides towards improving our theoretical understanding of forests, and the present review paper provides a helpful overview of this literature. However, we are still far from having a complete theory about why random forests work so well in practice. Thus, in the spirit guiding interested readers through conceptual thickets surrounding the random forests literature, the goal of this note is to highlight some open questions that may motivate further research.

This comment refers to the invited paper available at: doi:[10.1007/s11749-016-0481-7](https://doi.org/10.1007/s11749-016-0481-7).

✉ Stefan Wager
swager@stanford.edu

¹ Department of Statistics, Stanford University, Stanford, USA

2 What makes for a good splitting rule?

Most random forests rely on trees built using the CART splitting rules of [Breiman et al. \(1984\)](#), i.e., splits chosen by greedy minimization of squared-error loss for regression or Gini impurity for classification. These splitting rules have been found to work well again and again in real-world applications, and so a practitioner should feel confident in using them without hesitation. However, from a theoretical perspective, these splitting rules are essentially heuristics; and this lack of conceptual grounding makes it more difficult to extend the use of random forests to new application areas.

There has been considerable interest in using random forests for a wide variety of tasks, ranging from quantile estimation ([Meinshausen 2006](#)) and causal inference ([Wager and Athey 2015](#)) to survival analysis ([Ishwaran et al. 2008](#)). In each new application area, we need to decide which splitting rule to use for our trees, and this can be difficult: for now, there is not a single well-established recipe for doing so. At one extreme, in the context of causal inference, [Athey and Imbens \(2015\)](#) consider several candidate splitting rules that are carefully tailored to their application area, while at the other extreme, for quantile estimation, [Meinshausen \(2006\)](#) just uses the same splitting rule as for regression. Having a better understanding of why CART works well in the classification and regression settings could be helpful in developing appropriate splitting rules for new problems.

From a different perspective, to prove consistency results for high-dimensional regression forests, both [Biau \(2012\)](#) and [Wager and Walther \(2015\)](#) use splitting rules with a distinctly different flavor from CART: in both cases, the authors construct splitting rules that privilege splitting on “important” features beyond what a greedy rule would do. At the very least, these results show that CART may not be the best option in all regimes. It would be desirable to have a deeper understanding of the kinds of problems for which we should use CART.

3 Why is feature sampling helpful?

As described in the paper, almost all implementations of random forests rely on randomizing the choice of splitting variables: at each step of the algorithm, we first sample m features at random from the p available features, and then split on one of those features. In the empirical literature, this randomized splitting procedure is a well-established idea going back at least to [Amit and Geman \(1997\)](#). In practice, feature sampling often acts as a form of variance-reducing regularization ([Hastie et al. 2009](#), section 15.4), although counter examples are also available (e.g., [Wager et al. 2014](#), figure 6). It would be interesting to have theoretical results that could explain when and why this procedure is a good idea, and perhaps even guide the choice of m . Much of the recent work on the consistency of random forests allows for feature sampling—but this still falls short of answering why the technique helps.

At least superficially, the feature sampling in random forests seems to have an obvious connection to the dropout algorithm of [Srivastava et al. \(2014\)](#) for training neural networks. While running stochastic gradient descent, dropout involves randomly sampling the set of features that are considered in any step of the algorithm. However, in a

contrast to randomized splitting in a forest, the effect of dropout training is fairly well understood: for example, in the case of single-layer models, dropout can be understood as a form of data-adaptive ridge-like regularization (Baldi and Sadowski 2014; Wager et al. 2013). Fleshing out the connections between dropout and random forests could provide new insights about the role of feature sampling in growing trees.

4 What about Bayes?

Finally, as a third challenge to the theoretical literature on random forests, it would be interesting to study how random forests relate to Bayesian estimation. For example, it would be nice to know if feature sampling can be motivated using Bayesian principles. Bayesian forest algorithms have been found to work well in practice (Chipman et al 2010), and there has also been interest in using the posterior distributions of such forests for statistical inference (Green and Kern 2012; Hill 2011). However, our theoretical understanding of such procedures is still fairly limited, e.g., in terms of how to interpret the prior information encoded by a Bayesian forest, or in establishing their frequentist consistency properties. Gaining a better understanding of what random forests are doing from a Bayesian perspective could lead to valuable new insights.

References

- Amit Y, Geman D (1997) Shape quantization and recognition with randomized trees. *Neural Comput* 9(7):1545–1588
- Athey S, Imbens G (2015) Machine learning methods for estimating heterogeneous causal effects. *arXiv preprint arXiv:1504.01132*
- Baldi P, Sadowski P (2014) The dropout learning algorithm. *Artif Intell* 210:78–122
- Biau G (2012) Analysis of a random forests model. *J Mach Learn Res* 13:1063–1095
- Breiman L (2001) Random forests. *Mach Learn* 45(1):5–32
- Breiman L, Friedman J, Stone CJ, Olshen RA (1984) *Classification and regression trees*. CRC Press, Boca Raton
- Chipman HA, George EI, McCulloch RE (2010) BART: Bayesian additive regression trees. *Ann Appl Stat* 4(1):266–298
- Green DP, Kern HL (2012) Modeling heterogeneous treatment effects in survey experiments with Bayesian additive regression trees. *Public Opin Q* 76(3):491–511
- Hastie T, Tibshirani R, Friedman J (2009) *The elements of statistical learning*. Springer, New York
- Hill JL (2011) Bayesian nonparametric modeling for causal inference. *J Comput Graph Stat* 20(1):217–240
- Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS (2008) Random survival forests. *Ann Appl Stat* 2: 841–860
- Meinshausen N (2006) Quantile regression forests. *J Mach Learn Res* 7:983–999
- Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R (2014) Dropout: a simple way to prevent neural networks from overfitting. *J Mach Learn Res* 15(1):1929–1958
- Wager S, Athey S (2015) Estimation and inference of heterogeneous treatment effects using random forests. *arXiv preprint arXiv:1510.04342*
- Wager S, Walther G (2015) Uniform convergence of random forests via adaptive concentration. *arXiv preprint arXiv:1503.06388*
- Wager S, Wang S, Liang PS (2013) Dropout training as adaptive regularization. In: Burges CJC, Bottou L, Welling M, Ghahramani Z, Weinberger KQ (eds) *Advances in neural information processing systems*, pp 351–359
- Wager S, Hastie T, Efron B (2014) Confidence intervals for random forests: the jackknife and the infinitesimal jackknife. *J Mach Learn Res* 15(1):1625–1651