Stability of Random Forests and Coverage of Random-Forest Prediction Intervals

Yan Wang, wangyan@wayne.edu; Huaiqing Wu, isuhw@iastate.edu; Dan Nettleton, dnett@iastate.edu

This research was supported in part by the US National Science Foundation under grant HDR:TRIPODS 19-34884.

Take-home message

Random Forests is a provably stable algorithm under mild conditions. One can use the out-of-bag error of Random Forests to construct prediction intervals with guaranteed non-asymptotic coverage. Running Random Forests once, one can not only obtain a point predictor, but also a justified prediction interval for a future test point.

Summary of results

Random Forests (RF) is a popular machine learning algorithm. However, not much is known about its theoretical properties.

- We prove the RF is stable under the mild condition that the squared response Y^2 is not heavy-tailed distributed.
- Our theoretical results hold for the practical version of the RF such as $\mathrm{randomForest}$ in R.
- Primarily based on the stability property, we provide non-asymptotic (and asymptotic) coverage guarantees of prediction intervals constructed from the out-of-bag (OOB) error of the RF.
- RF prediction intervals can be constructed almost without additional computation.

Prediction interval construction methods

Conformal prediction (CP) provides theoretically justified prediction intervals for almost all machine learning algorithms in practice.

- Full CP [1]: any algorithm, distribution-free, but computationally prohibitive
- Split CP [2]: any algorithm, distribution-free, but inefficient data usage
- Jackknife+ [3]: any algorithm, distribution-free, efficient data usage, but computational cost can still be high for modern learning algorithms
- Jackknife+-after-bootstrap [4]: any algorithm, distribution-free, efficient data usage, but the number of bags B is random without stability assumptions and needs to aggregate base learners
- **Ours** [5]: RF only, mild distributional assumptions, efficient data usage, negligible additional computational cost, no need to aggregate tree predictors
- All methods have non-asymptotic coverage guarantees. Our method is based on the stability of bagged algorithms established in Ref. [6].

Stability of Random Forests (in theory)

- Stability of derandomized RF (Number of trees $B=\infty$)

$$P\left(\left|\mathrm{rf}(X)-\mathrm{rf}^{\setminus i}(X)\right|>\varepsilon_2\right)\le\nu_2$$

Step 1: Conditional on training data, all tree predictors output bounded predictive values, and Theorem 8 in Ref. [6] applies.

Step 2: $P(\cdot)=E[P(\cdot\,|\text{training data})]$

- Concentration of (resampling) measure

$$P(|\mathrm{rf}(X)-\mathrm{RF}(X)|>\varepsilon_1)\le\nu_1$$

$$P\left(\left|\mathrm{RF}^{\setminus i}(X)-\mathrm{rf}^{\setminus i}(X)\right|>\varepsilon_3\right)\le\nu_3$$

- Stability of RF (**Theorem 8**): The union bound gives

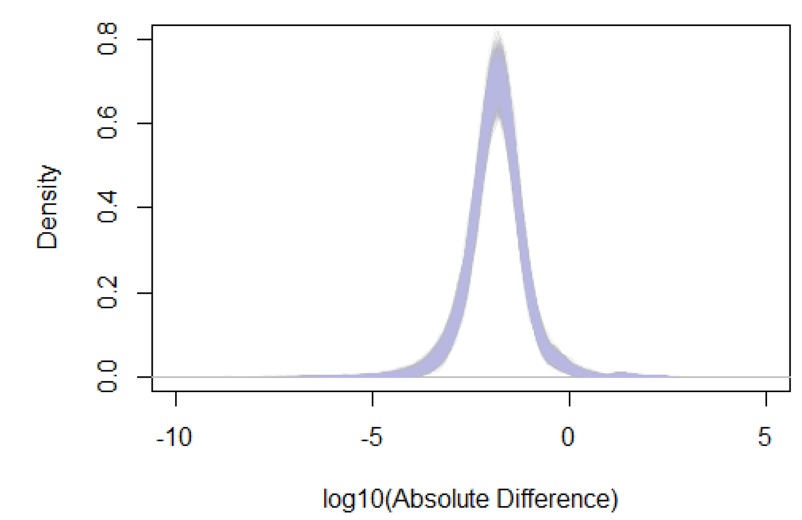
$$P\left(\left|\mathrm{RF}(X)-\mathrm{RF}^{\setminus i}(X)\right|>\varepsilon\right)\le\nu$$

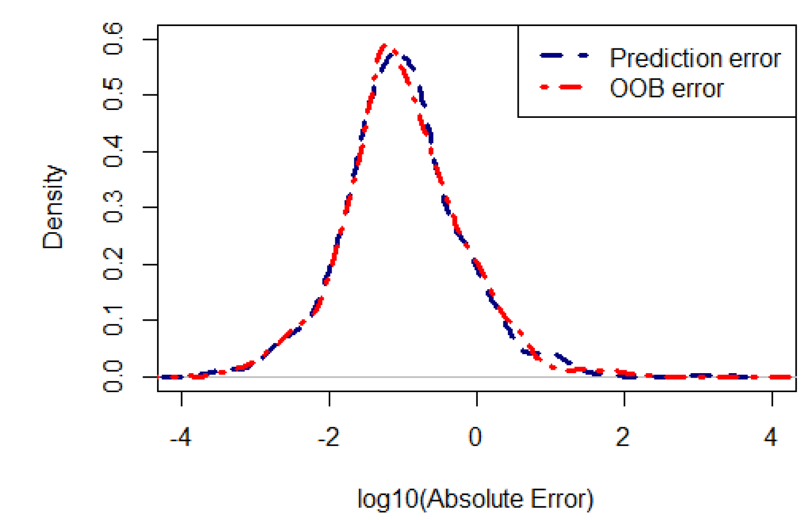
- Both ε and ν approach 0 for non-heavy-tailed Y^2 as $n,B\to\infty$. (Theoretically proved for sub-gamma Y^2.)

Stability of Random Forests (in practice)

Example: Create a virtual dataset with 3000 training data points and 1000 test points, where Y follows the standard Cauchy distribution. Set $B=1000$.

- Left: Density plot of $\log_{10}(|RF(X)-RF^{\setminus i}(X)|)\,,i\in[3000]$
- Right: Density plot of $\log_{10}(|Y-RF(X)|)$ and $\log_{10}(|Y_i-RF^{\setminus i}(X_i)|)$





RF stability persists even if the light-tail assumption is violated, implying the theoretical bound is not necessarily tight.

Comparison of prediction intervals using Random Forests

Method	Output predictors	Prediction interval for future Y	Theoretical coverage	Conditions beyond IID data
Jackknife+ [3]	$\mathrm{RF}^{-i},i\in[n]$	$\left[q_{n,\alpha}^{-}\{\mathrm{RF}^{-i}(X)-R_i^{\mathrm{LOO}}\},q_{n,\alpha}^{+}\{\mathrm{RF}^{-i}(X)+R_i^{\mathrm{LOO}}\}\right]$	$\ge 1-2\alpha$	None
Jackknife+-after-bootstrap [4]	$\mathrm{RF}^{\setminus i},i\in[n]$	$\left[q_{n,\alpha}^{-}\{\mathrm{RF}^{\setminus i}(X)-R_i\},q_{n,\alpha}^{+}\{\mathrm{RF}^{\setminus i}(X)+R_i\}\right]$	$\ge 1-2\alpha$	Binomial number of tress
Jackknife with stability [3]	RF and $\mathrm{RF}^{-i},i\in[n]$	$\mathrm{RF}(X)\pm q_{n,\alpha}\{R_i^{\mathrm{LOO}}+\varepsilon\}$	$\ge 1-\alpha-O(\sqrt{\nu})$	Stability (algorithmic)
Jackknife+-after-bootstrap with stability [4]	$\mathrm{RF}^{\setminus i},i\in[n]$	$\left[q_{n,\alpha}^{-}\{\mathrm{RF}^{\setminus i}(X)-R_i\}-\varepsilon,q_{n,\alpha}^{+}\{\mathrm{RF}^{\setminus i}(X)+R_i\}+\varepsilon\right]$	$\ge 1-\alpha-O(\sqrt{\nu})$	Stability (algorithmic + ensemble)
Jackknife-after-bootstrap [7,8]	RF	$\mathrm{RF}(X)\pm q_{n,\alpha}\{R_i\}$ [7]	No guarantee provided	-
		$\mathrm{RF}(X)\pm q'_{n,\alpha}\{R_i\}$ [8]	$\to 1-\alpha$	Additive model, consistency of RF, etc.
Ours [5] (Jackknife-after-bootstrap with stability)	RF	$\mathrm{RF}(X)\pm q_{n,\alpha}\{R_i+\varepsilon\}$	$\ge 1-\alpha-O(\sqrt{\nu})$	Stability (**Theorem 9**)
		$\mathrm{RF}(X)\pm q_{n,\alpha}\{R_i-\varepsilon\}$	$\le 1-\alpha+\dfrac{1}{n+1}+O(\sqrt{\nu})$	+ Distinct residuals (**Theorem 10**)
		$\mathrm{RF}(X)\pm q_{n,\alpha}\{R_i\}$	$\to 1-\alpha$	+ Uniformly equicontinuous CDF of prediction error and vanishing ε,ν (**Theorem 11**)

- RF^{-i}: Leave-one-out (LOO) RF predictor; $\mathrm{RF}^{\setminus i}$: OOB RF predictor
- $R_i^{\mathrm{LOO}}=\left|Y_i-\mathrm{RF}^{-i}(X_i)\right|$: LOO error; $R_i=\left|Y_i-\mathrm{RF}^{\setminus i}(X_i)\right|$: OOB error
- Note $\mathrm{RF}^{\setminus i}(X_i)$ can be obtained *directly* using packages like $\mathrm{randomForest}$
- $q_{n,\alpha}^{-}$: the $\lfloor(n+1)\alpha\rfloor$-th smallest value
- $q_{n,\alpha}^{+}$: the $\lceil(n+1)(1-\alpha)\rceil$-th smallest value
- $q_{n,\alpha}=q_{n,\alpha}^{+}$; $q'_{n,\alpha}$: the $\lceil n(1-\alpha)\rceil$-th smallest value
- (ε,ν): a pair of stability parameters with $P(|f(X)-f_i(X)|>\varepsilon)\le\nu$, where f_i is either the LOO or OOB predictor

References

[1] V. Vovk, A. Gammerman, and G. Shafer. Algorithmic learning in a random world, Springer, 2005.
[2] J. Lei, *et al.* Distribution-free predictive inference for regression. JASA, 2018.
[3] R. F. Barber, *et al.* Predictive inference with the jackknife+. AoS, 2021.
[4] B. Kim, C. Xu, and R. F. Barber. Predictive inference is free with the jackknife+-after-bootstrap. NeurIPS, 2020.
[5] Y. Wang, H. Wu, and D. Nettleton. Statbility of random forests and coverage of random-forest prediction intervals. NeurIPS, 2023.
[6] J. A. Soloff, R. F. Barber, and R. Willett. Bagging provides assumption-free stability. arXiv:2301.12600.
[7] U. Johansson, *et al.* Regression conformal prediction with random forests. Machine Learning, 2014.
[8] H. Zhang, *et al.* Random forest prediction intervals. The American Statistician, 2019.