https://doi.org/10.3329/dujase.v7i1.62888

Evolution of Random Forest from Decision Tree and Bagging: A Bias-Variance Perspective

Muhammad Ibrahim

Department of Computer Science and Engineering, University of Dhaka, Dhaka-1000, Bangladesh

Email: ibrahim313@du.ac.bd

Received on 24 November 2021, Accepted for publication on 07 June 2022

ABSTRACT

The ensemble methods are one of the most heavily used techniques in machine learning. The random forest arguably spearheads this army of learners. Being sprung from the decision tree in the late 90s, the benefits of a random forest have rightfully attracted practitioners to widely and successfully apply this powerful yet simple-to-understand technique to numerous applications. In this study we explain the evolution of a random forest from a decision tree in the context of bias and variance of learning theory. While doing so, we focus on the interplay between the correlation and generalization error of the random forest. This analysis is expected to enrich the literature of random forests by providing further insight into its working mechanism. These insights will assist the practitioners of the random forest implement this algorithm more wisely and in an informed way.

Keywords: Supervised machine learning, Ensemble learning algorithms, Decision tree, Bagging, Random forests, Bias-variance tradeoff, Correlation.

1. Introduction

The late 90s witnessed a plethora of research works on inventing novel techniques of supervised machine learning. Among them, boosting [11], [12], support vector machines [6], bagging [2] and random forests [18], [4] gained much momentum. As there is no perfect solution to the generic supervised machine learning problem [30], researchers vied each other to develop approximate algorithms to tackle the learning problem. Each of these algorithms has its strengths and weaknesses, thereby offering a rich set of available options for the practitioners.

Ensemble methods [25], soon after their first postures in the early 90s [32], [3], rose to a wider applicability, thanks to their conceptual simplicity coupled with high effectiveness and to the abundance of computer processing power. These algorithms are relatively simpler to understand because the power of grouping together to enhance the individual capabilities has always been well-known to human civilization. These techniques are effective in prediction because they can leverage more than a single thought while predicting future outcomes.

In the sequel of ensemble methods, the term Random Forest (RF) was coined by the venerable statistician Leo Breiman in early 2000s [4]. It is called "random" because it randomly chooses features to split the available data at each node of a decision tree, and it is a "forest" because it deploys a collage of decision trees.

While a good number of works compare and contrast the performance of random forest with its predecessors such as bagging and decision trees from an empirical perspective (i.e., using experimental results), we have not found any work that explains its working principles by demonstrating the relationships among bias, variance, correlation, and error rate. This study aims to fill this gap in the literature.

The rest of this study is structured as follows: Section II introduces the random forest and its benefits. Section III lays the ground of this study. Section IV describes the contribution, i.e., the interplay among bias, variance, error rate and correlation of a random forest from both mathematical and intuitive perspectives. Section V highlights the implications of this study. Section VI briefly discusses some relevant existing works, which is followed by the conclusion in Section VII.

2. Random Forest and Its Benefits

The story of the random forest begins with a decision tree [5]. Being a very popular non-parametric method in the 80s and 90s among the machine learning researchers, a decision tree recursively splits the training data into sub-groups based on the "maximum perceived benefit". For classification task, this maximum benefit is calculated using functions like information gain or gini co-efficient. For regression task, this is calculated by mean squared error. Since a decision tree allows the hypothesis to move almost freely in the hypothesis space and only guided by the available training data, it incurs a very low bias. It, however, for the same reason, incurs a high variance due to its capability of capturing the unnecessary nuances of the training data. Therefore, a small perturbation in the training data may yield a drastically different tree thereby making its prediction highly unstable in many occasions.

To reduce the high variance of a single decision tree, a random forest [4] employs many decision trees learnt from different patches of the (same) training dataset. Moreover, the decision trees are modified to increase the variability among the trees in the following way: instead of considering all the attributes for a split, it selects a random subset of them, thereby offering a chance for the “marginalized” attributes/features to speak out.

A random forest may be considered to be a framework rather than an algorithm because tweaking its various components gives birth to new instances of this generic technique [7]. A wide range of applications [10] have been reaping benefit of this simple yet powerful learning technique. Ibrahim [20] nicely summarizes the benefits of random forests; the major ones include being extremely parallelizable, able to work with both discrete and continuous feature values including missing ones, robust to outliers, in need of almost no parameter tuning, and not being sensitive to feature scaling.

3. Motivation

As mentioned earlier, the random forest’s main working strategy is to reduce the high variance of individual decision trees while, at the same time, retaining their low bias. Towards this end, two important ingredients to investigate are (1) the sample (aka training data) size and (2) the number of attributes to consider for each split of a tree. In bagging [2] which is considered to be a forerunner of random forest, the first of these two ingredients is utilized to reduce variance. A random forest furthers this trend by utilizing both of these ingredients. We elaborate the process below.

In bagging (aka bagged ensemble), a collection of decision trees are employed where each tree is learnt using a bootstrapped copy of the training set. That is, a new training set of the same size is created by sampling-with-replacement instances from the available training set. It is shown that about 63% data are unique in the bootstrapped samples drawn from a single training sample [9]. Applying bootstrapping in a bagged ensemble ensures that the trees become different from each other.

Each tree is trained as follows. (For the sake of simplicity, we consider a regression task.) At the root of a tree, the entire available training data are split into two subsets in the feature space based on the mean squared error. The split-point (considering all the features and their values) that gives the minimum mean squared error is selected. This process is recursively continued for the children nodes until a termination criterion (such as a predefined minimum number of instances at a node or a predefined height of the tree etc.) is met. In a random forest, however, all the features are not considered, rather a small and random subset of the features are chosen to select the split-point. While this strategy does increase variability (or, in more technical term, decrease correlation) among the trees, it does risk increasing the individual tree variances, thereby warranting a deeper understanding the interplay of all the following quantities: ensemble variance, correlation among the trees, individual tree variance, bias, and error rate. In the rest of the article we attempt to shed some light on this interplay. Specifically, we deal with the research question: how can we mathematically expound the trade-off between variance and bias of a random forest and how does it help us understand their relation with error rate and correlation? Knowing the answer to this question would help the practitioners deploy these algorithms in various applications in a more informed way.

4. Interplay among Error Rate, Bias, Variance, and Correlation in A Random Forest

The theory of learning algorithms [15] tells us that for squared loss function (i.e., the regression task), the following equation holds:

$$GE = bias^2 + variance + IE, \qquad (1)$$

where *GE* and *IE* stand for Generalization Error and Irreducible Error respectively. In simple terms, bias is the error incurred due to the limitation of the hypothesis space, variance is the error incurred due to the peculiarities of a particular training set, and irreducible error exists due to the intrinsic noise present in the labels of the data.

Before delving into our analysis we note here that since many research studies such as [13], [14], [27], [31], [29] demonstrate that bootstrapping without replacement works as well as bootstrapping with replacement, in the analysis that follow we use the former setting.

Let the variables *T*, *B*, and *R* denote a single regression tree, a bagged ensemble, and a random forest respectively; and b^2 and σ^2 denote the squared bias and variance of a learner respectively. We use T_s to denote that a tree is learnt using *s* percentage of the available training set; thus T_{63} and T_{100} denote that the tree is learnt from a bootstrap (without replacement) sample and from the full sample respectively. B_s denotes an ensemble of bagged trees where each tree is learnt using *s* percentage of the training set, and similar interpretation is used for R_s. If the meaning is obvious, we omit the subscripts; for example, sometimes we use bare *B* and *R* to denote the default cases for these two learners (bagging and random forest), i.e., B_{63} and R_{63} respectively.

After numerous works such as [8], [22], [23], we ignore the irreducible term of Equation 1 for our analysis. We can thus write for decision tree, bagged ensemble and random forest:

$$Err_T = b^2{}_{\mathrm{T}} + \sigma^2{}_T \qquad (2)$$
$$Err_B = b^2{}_B + \sigma^2{}_B$$
$$Err_R = b^2{}_R + \sigma^2{}_R$$

We begin with the analysis of a bagged ensemble (aka bagging) because it can be considered as a predecessor to a random forest.

A. Bagging for Variance Reduction

It is known that the bias of a bagged ensemble is equal to that of a single tree of the ensemble [17, Ch. 15].[1] Therefore:

$$b^2_B = b^2_{T63}\ , \tag{3}$$

and variance:

$$\sigma^2_B = \rho\sigma^2_{T63} + ((1-\rho)\sigma^2_{T63})\ /\ E\ , \tag{4}$$

where ρ is the correlation between two trees (at a datapoint) and E is size of the ensemble, i.e., the number of trees respectively.[2] If E is large enough, the 2nd term of Equation 4 tends to vanish irrespective of the values of ρ and $\sigma^2\ T$. For an ensemble of identical trees, $\rho = 1$. The idea of bagging is to decrease ρ by using a different training set to learn each tree. However, in reality only one training set is available. To overcome this problem, each tree is learnt from a bootstrap sample thereby making each tree different from one another. But bootstrapping gives rise to two additional concerns: (1) it increases $\sigma^2\ T$ (i.e., $\sigma^2_{T63} > \sigma^2_{T100}$), because individual trees are now learnt using less information than before and hence the overfitting tendency of a tree learnt from a bootstrapped sample is more than that of a tree learnt from the original sample, and (2) it increases bias of individual trees (i.e., b^2_{T63} $(= b^2_B\ ) > b^2_{T100}$), because, again, now the sample space is smaller which, in turn, shrinks the hypothesis space. The good news is, empirically it has been observed the positive effect of decreasing ρ is much greater than the negative effect of increasing σ^2_{T63} (cf. Equation 4), which can be expressed as:

$$\sigma^2_B << \sigma^2_{T100}\ .$$

We can thus summarize the above discussion as follows:

$$b^2_{\mathrm{B}} > b^2_{T100} \tag{5}$$

$$\sigma^2_B << \sigma^2_{T100}\ . \tag{6}$$

Empirically it has been found for the above two equations that:

$$b^2_B + \sigma^2_B < b^2_{T100} + \sigma^2_{T100}$$

$$\Rightarrow Err_B < Err_{T100} \text{ (using Equation 2).}$$

Therefore, from the above mathematical derivations, the following can be said about bagging in a nutshell: the benefit of aggregating predictions of many trees learnt from comparatively smaller training sets (i.e., bootstrapped training sets) outweighs the benefit of using a single prediction from a tree learnt from a comparatively larger training set.

Another way to explain the benefit of the diversity in an ensemble is as follows. By increasing diversity across trees we enable the ensemble to capture increasing amount of non-linear (complex) relationship between labels (by making the ensemble decision boundary smoother when averaging many different axis-parallel decision boundaries), given that the ensemble size is sufficiently large.

B. Adding Further Randomness to Make a Random Forest

The idea of the random forest is to reduce the correlation, ρ even more, yet without substantially increasing single-tree variance, σ^2_{T63} and bias, b^2_{T63} $(= b^2_{R63}\ )$. The way a random forest achieves this is by modifying the tree building procedure as follows. At each node a small number (denoted by K) of features are randomly chosen and the optimal split is then determined over only those K chosen features. An additional benefit of this scheme is, the features which did not get "chance to speak" due to the influence of some other stronger features now get an opportunity to be selected as a split-point, thereby escaping the local minima problem of the pure greedy learning strategy of a decision tree. (We note that there are other possible ways to achieve the same correlation reduction effect, for example, the approach of using even further smaller sub-samples per tree [19].)

The formulation of variance of a random forest is the same as that of bagging (cf. Equation 4) because a random forest simply adds, on top of the bagging's idea of bootstrapping, an extra source of randomness to a tree. Again, empirically it has been found that the (positive) effect of reduction of ρ on σ^2_R is greater than the (negative) effect of rise of σ^2_{T63} (cf. Equation 4), thereby causing:

$$\sigma^2_R < \sigma^2_B << \sigma^2_{T100} \text{ (using Equation 6).} \tag{7}$$

As for the bias of a random forest, it, like bagging, is equal to that of a single tree of the ensemble. As long as K is not too small and the ensemble size is sufficiently large, the additional randomness does not harm the systematic error of the ensemble (i.e., the expected error across multiple ensembles learnt from different samples), thereby causing b^2_R to be roughly of similar level to b^2_B .[3] Thus, we can write:

$$b^2_R >\approx b^2_B > b^2_{T100} \text{ (using Equation 5).} \tag{8}$$

Now we use the same reasoning of the discussion of bagging:

[1] This can be explained as follows. The (squared) bias at a data point x_k is $(l_k - \mathrm{E}[f(x_k)])^2$ where l_k is the label of x_k and $f(x_k)$ is the prediction. Given the model complexity is fixed (a tree with fixed parameters is being used in all cases, i.e., trees are identically distributed [17, Ch. 15]), increasing the number of trees cannot improve the bias component of error. This is because increasing the number of trees to compute their average simply brings the average closer to the true average; it does not systematically increase/decrease the average.

[2] To see a derivation of Equation 4, please see [19].

[3] We note that due to the greedy nature in which each tree is built, it is likely but not necessarily the case that the trees in a random forest have higher bias than the trees in bagged ensemble.

empirically it has been found that considering Equations 7 and 8 we can write:

$b^2_R + \sigma^2_R < b^2_B + \sigma^2_B < b^2_{T100} + \sigma^2_{T100}$

$\Rightarrow Err_R < Err_B < Err_{T100}$ (using Equation 2).

We highlight the following point again: reducing correlation helps only if the single-tree variance, σ^2_{T63} and single-tree bias, b_{T63} are not greatly increased; that is why using a very small number of random features at each node of a random forest does not yield good performance [17, Ch. 15].[4]

From the above mathematical discussion, in a nutshell the following can be said about a random forest: as we reduce correlation ρ by bootstrapping and by selecting a subset of random features, we increase both the variance σ^2_{T63} and squared bias b^2_{T63} of random forest as compared to σ^2_{T100} and b^2_{T100} .[5] However, σ^2_R continues to decrease due to the reduction of ρ (cf. Equation 4). The result of this reduction is, if we continue to increase randomness (by decreasing the number of randomly chosen features used to determine each split), then up to some point the ensemble error, Err_R continues to decrease. Beyond this point, Err_R, however, starts to increase again; and the question is, is higher variance σ^2_{T63} , or higher squared bias b^2_{T63} , or both, responsible for this rise in Err_R? This study is scant in the literature; in fact we found only one such work [17, Ch. 15] who conduct a pilot experiment on a synthetic regression dataset to show that if K is reduced greatly, both the single tree variance and bias increase which in turn increase the error rate. This interesting direction of research needs further attention from researchers.

C. Beyond Standard Random Forest

The random forest, like many other reputed learning algorithms, sprung a good number of variations. We briefly discuss two of them that are relevant to the theme of this article here, namely Extremely Randomized Trees [16] and random forest with aggressive sub-sampling [19].

Geurts et al. [16] almost go wild in adding further randomness in a random forest. They not only select a random subset of features to split the data at a node, but also randomly select the split point among those features. To mitigate the possible negative effect in terms of high variance, they, in contrast to the standard random forest's settings, advocate not to use any subset of the training set to learn a tree, rather they advocate using the entire training set. The authors report better performance in many scenarios that has made the extremely randomized trees quite popular among the practitioners.

Ibrahim [19] explores yet another direction. His study ensues from the study of famous theoreticians Friedman and Hall [14]: whereas the latter investigate what happens when K is tuned in a bagging setting, the former looks into what happens when the sub-sample size per tree is tuned in a random forest setting.

5. Implication of This Study

Being a theoretical explanation of some hidden aspects of a machine learning algorithm, this study does not warrant any experiments. That said, we can link the analysis of this study with a few empirical experiments conducted in the existing literature. If the motivation of many empirical investigations like Geurts et al. [16], Freidman and Hall [14], Ibrahim [19], and the like is not well-understood by the ordinary practitioners, our analysis presented in this paper will be quite helpful for them. Needless to say, to successfully implement of a machine learning algorithm, it is imperative for the practitioners to have a reasonable level of understanding of the theoretical aspects of the algorithm in question.

6. Related Work

There exist some reading materials to understand bias-variance decomposition of randomized tree ensembles such as [4] and [17, Ch. 15]. However, their explanations of the bias and variance of random forests are not particularly thorough, rather they focus on only small parts of the big picture.

A. On Correlation and Strength of Random Forest

Bernard et al. [1] plot strength (i.e., predictive accuracy of individual trees) and correlation between the trees along with error rate of random forests for classification. They also examine the effect of ensemble size. Their findings include: the relationship between strength, correlation and error rate formulated in Theorem 2.3 of Breiman's seminal paper on random forests [4] largely holds in practice.

B. On Parameter Tuning of Random Forest

Segal [28] tunes the minimum node size parameter n_{min} of a tree for the regression setting. Lin and Jeon [24] also work with n_{min} and, in addition, with the number of candidate features at each node K (usually set to $log(\#features) + 1$). Their findings indicate that slight improvement in performance may be found by tuning these parameters, although default values work well in practice. Hastie et al. [17, Ch. 15] comment that while for regression tuning n_{min} helps slightly, for classification it rarely makes any difference.

C. On Theoretical Analysis of Random Forest

The theory behind random forest was not given much attention until recently. Some notables works are: Wager [31], Scornet [26] and the references therein. Most of these works analyze simplified versions of random

[4] For example, Ibrahim [20] shows that an ensemble of completely randomized trees that has very low correlation may not perform well on big data.

[5] Also, σ^2_{T63} is increased as compared to bagger's variance.

forest because the standard random forest where the best split is found among a random subset of the variables is difficult to theoretically analyze.

We see from the above discussion that while some work do analyze the strength, correlation, and generalization error of a random forest, no thorough explanation of the bias-variance interaction of random forests exists in the literature. This article is an attempt to narrow down this gap.

7. Conclusion

Random forest is among the top machine learning algorithms that are simple to understand and deploy and yet highly effective and efficient in performance. Having a clear understanding of a machine learning algorithm is pivotal for successful application of that algorithm. Although many works do study various properties of a random forest from the viewpoint of empirical results, exactly how to explain the interplay among its bias, variance, correlation, and error rate is, to the best of our knowledge, missing in the literature. In this paper we have provided an explanation of random forest's working mechanism from a bias-variance perspective as well as of their interplay with the error rate and correlation of the ensemble. This analysis is expected to assist the practitioners better understand the evolution of a random forest from a single tree, thereby employing this powerful technique in a more informed and wise fashion. Future directions emanated from this study include an empirical investigation (cf. Section IV-B) of the behaviour of single tree variance and bias as the number of randomly chosen features at each split is reduced greatly. Also, empirical investigations like [21] can reap benefit from the analysis presented in this study.

References

1 S. Bernard, L. Heutte, and S. Adam. Towards a better understanding of random forests through the study of strength and correlation. In Emerging Intelligent Computing Technology and Applications. With Aspects of Artificial Intelligence, pages 536–545. Springer, 2009.

2 L. Breiman. Bagging predictors. Machine Learning, 24(2):123–140, 1996.

3 L. Breiman. Stacked regressions. Machine learning, 24(1):49–64, 1996.

4 L. Breiman. Random forests. Machine Learning, 45(1):5–32, 2001.

5 L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen. Classification and regression trees. CRC press, 1984.

6 C. Cortes and V. Vapnik. Support vector machine. Machine learning, 20(3):273–297, 1995.

7 A. Criminisi. Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semisupervised learning. Foundations and Trends® in Computer Graphics and Vision, 7(2-3):81–227, 2011.

8 P. Domingos. A unified bias-variance decomposition. In Proceedings of 17th International Confference on Machine Learning. Stanford CA Morgan Kaufmann, pages 231–238, 2000.

9 B. Efron and R. J Tibshirani. An introduction to the bootstrap. CRC Press, 1994.

10 K. Fawagreh, M. M. Gaber, and E. Elyan. Random forests: from early developments to recent advancements. Systems Science & Control Engineering: An Open Access Journal, 2(1):602–609, 2014.

11 Y. Freund and R. E. Schapire. A desicion-theoretic generalization of on-line learning and an application to boosting. In Computational Learning Theory, pages 23–37. Springer, 1995.

12 J. H. Friedman. Greedy function approximation: a gradient boosting machine.(english summary). Annals of Statistics, 29(5):1189–1232, 2001.

13 J. H. Friedman. Stochastic gradient boosting. Computational Statistics & Data Analysis, 38(4):367–378, 2002.

14 J. H. Friedman and P. Hall. On bagging and nonlinear estimation. Journal of Statistical Planning and Inference, 137(3):669–683, 2007.

15 S. Geman, E. Bienenstock, and R. Doursat. Neural networks and the bias/variance dilemma. Neural Computation, 4(1):1–58, 1992.

16 P. Geurts, D. Ernst, and L. Wehenkel. Extremely randomized trees. Machine Learning, 63(1):3–42, 2006.

17 T. Hastie, R. Tibshirani, and J. Friedman. The elements of statistical learning, 2009.

18 T. K. Ho. The random subspace method for constructing decision forests. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 20(8):832–844, 1998.

19 M. Ibrahim. Reducing correlation of random forest–based learning-to-rank algorithms using subsample size. Computational Intelligence, 35(4):774–798, 2019.

20 M. Ibrahim. An empirical comparison of random forest-based and other learning-to-rank algorithms. Pattern Analysis and Applications, 23(3):1133–1155, 2020.

21 M. Ibrahim. Understanding bias and variance of learning-to-rank algorithms: An empirical framework. Applied Artificial Intelligence, pages 1–34, 2021.

22 R. Kohavi, D. H. Wolpert, et al. Bias plus variance decomposition for zero-one loss functions. In International Confference on Machine Learning (ICML), pages 275–283, 1996.

23 E. B. Kong and T. G. Dietterich. Error-correcting output coding corrects bias and variance. In International Confference on Machine Learning (ICML), pages 313–321, 1995. [24] Y. Lin and Y. Jeon. Random forests and adaptive nearest neighbors. Journal of the American Statistical Association, 101(474):578–590, 2006.

25 D. Opitz and R. Maclin. Popular ensemble methods: An empirical study. Journal of artificial intelligence research, 11:169–198, 1999.

26 E. Scornet. On the asymptotics of random forests. arXiv preprint arXiv:1409.2090, 2014.

27 E. Scornet, G. Biau, and J. Vert. Consistency of random forests. arXiv preprint arXiv:1405.2881, 2014.

28 M. R. Segal. Machine learning benchmarks and random forest regression. 2004.

29 J. Shao et al. Impact of the bootstrap on sample surveys. Statistical Science, 18(2):191–198, 2003.

30 V. Vapnik. The nature of statistical learning theory. Springer, 1999.

31 S. Wager. Asymptotic theory for random forests. arXiv preprint arXiv:1405.0352, 2014.

32 D. H. Wolpert. Stacked generalization. Neural networks, 5(2):241–259, 1992.