Journal of Multivariate Analysis 146 (2016) 72–83

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

On the asymptotics of random forests

Erwan Scornet

Sorbonne Universités, UPMC Univ Paris 06, F-75005, Paris, France

ARTICLE INFO

Article history:
Received 7 September 2014
Available online 24 June 2015

AMS 2010 subject classifications:
62G05
62G20

Keywords:
Random forests
Randomization
Consistency
Central limit theorem
Empirical process
Number of trees
q-quantile

ABSTRACT

The last decade has witnessed a growing interest in random forest models which are recognized to exhibit good practical performance, especially in high-dimensional settings. On the theoretical side, however, their predictive power remains largely unexplained, thereby creating a gap between theory and practice. In this paper, we present some asymptotic results on random forests in a regression framework. Firstly, we provide theoretical guarantees to link finite forests used in practice (with a finite number M of trees) to their asymptotic counterparts (with $M = \infty$). Using empirical process theory, we prove a uniform central limit theorem for a large class of random forest estimates, which holds in particular for Breiman's (2001) original forests. Secondly, we show that infinite forest consistency implies finite forest consistency and thus, we state the consistency of several infinite forests. In particular, we prove that q quantile forests – close in spirit to Breiman's (2001) forests but easier to study – are able to combine inconsistent trees to obtain a final consistent prediction, thus highlighting the benefits of random forests compared to single trees.

1. Introduction

Random forests are a class of algorithms used to solve classification and regression problems. As ensemble methods, they grow several trees as base estimates and aggregate them to make a prediction. In order to obtain many different trees based on a single training set, random forests procedures introduce randomness in the tree construction. For instance, trees can be built by randomizing the set of features [14,21], the data set [5,6], or both at the same time [7,10].

Among all random forest algorithms, the most popular one is that of Breiman [7], which relies on CART procedure (Classification and Regression Trees, [8]) to grow the individual trees. As highlighted by several applied studies (see, e.g., [20,13]), Breiman's [7] random forests can outperform state-of-the-art methods. They are recognized for their ability to handle high-dimensional data sets, thus being useful in fields such as genomics [31] and pattern recognition [33], just to name a few. On the computational side, Breiman's [7] forests are easy to run and robust to changes in the parameters they depend on [27, 16]. Besides, extensions have been developed in ranking problems [9], quantile estimation [28], and survival analysis [25]. Interesting new developments in the context of massive data sets have been achieved. For instance, Geurts et al. [17] modified the procedure to reduce calculation time, while other authors extended the procedure to online settings [11,26, and the references therein].

While Breiman's [7] forests are extensively used in practice, some of their mathematical properties remain under active investigation. In fact, most theoretical studies focus on simplified versions of the algorithm, where the forest construction is independent of the training set. Consistency of such simplified models has been proved (e.g., [2,24,11]). However, these results do not extend to Breiman's [7] original forests whose construction critically depends on the whole training set. Recent attempts to bridge the gap between theoretical forest models and Breiman's [7] forests have been made by Wager [37] and Scornet et al. [34] who establish consistency of the original algorithm under suitable assumptions.

E-mail address: erwan.scornet@upmc.fr.

http://dx.doi.org/10.1016/j.jmva.2015.06.009
0047-259X/

Apart from the dependence of the forest construction on the data set, there is another fundamental difference between existing forest models and ones implemented. Indeed, in practice, a forest can only be grown with a finite number M of trees although most theoretical works assume, by convenience, that $M = \infty$. Since the predictor with $M = \infty$ does not depend on the specific tree realizations that form the forest, it is therefore more amenable to analysis. However, surprisingly, no study aims at clarifying the link between finite forests (finite M) and infinite forests ($M = \infty$) even if some authors [29,38] proved results on finite forest predictions at a fixed point $\mathbf{x}$.

In the present paper, our goal is to study the connection between infinite forest models and finite forests used in practice in the context of regression. We start by proving a uniform central limit theorem for various random forests estimates, including Breiman's [7] ones. In Section 3, assuming some regularity on the regression model, we point out that the $\mathbb{L}^2$ risk of any infinite forest is bounded above by the risk of the associated finite forests. Thus infinite forests are better estimate than finite forests in terms of $\mathbb{L}^2$ risk. Under the same assumptions, our analysis also shows that the risks of infinite and finite forests are close, if the number of trees is chosen to be large enough. An interesting corollary of this result is that infinite forest consistency implies finite forest consistency. Finally, in Section 4, we prove the consistency of several infinite random forests. In particular, taking one step towards the understanding of Breiman's [7] forests, we prove that q quantile forests, a variety of forests whose construction depends on the positions $\mathbf{X}_i$'s of the data, are consistent. As for Breiman's [7] forests, each leaf of each tree in q quantile forests contains a small number of points that does not grow to infinity with the sample size. Thus, q quantile forests average inconsistent trees estimate to build a consistent prediction.

We start by giving some notation in Section 2. All proofs are postponed to Section 5.

2. Notation

Throughout the paper, we assume to be given a training sample $\mathcal{D}_n = (\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ of $[0, 1]^d \times \mathbb{R}$-valued independent random variables distributed as the prototype pair $(\mathbf{X}, Y)$, where $\mathbb{E}[Y^2] < \infty$. We aim at predicting the response Y, associated with the random variable $\mathbf{X}$, by estimating the regression function $m(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$. In this context, we use random forests to build an estimate $m_n : [0, 1]^d \to \mathbb{R}$ of m, based on the data set $\mathcal{D}_n$.

A random forest is a collection of M randomized regression trees (for an overview on tree construction, see Chapter 20 in [19]). For the jth tree in the family, the predicted value at point $\mathbf{x}$ is denoted by $m_n(\mathbf{x}, \Theta_j)$, where $\Theta_1, \ldots, \Theta_M$ are independent random variables, distributed as a generic random variable Θ, independent of the sample $\mathcal{D}_n$. This random variable can be used to sample the training set or to select the candidate directions or positions for splitting. The trees are combined to form the finite forest estimate

$$m_{M,n}(\mathbf{x}, \Theta_1, \ldots, \Theta_M) = \frac{1}{M} \sum_{m=1}^{M} m_n(\mathbf{x}, \Theta_m). \tag{1}$$

By the law of large numbers, for any fixed $\mathbf{x}$, conditionally on $\mathcal{D}_n$, the finite forest estimate tends to the infinite forest estimate

$$m_{\infty,n}(\mathbf{x}) = \mathbb{E}_\Theta [m_n(\mathbf{x}, \Theta)].$$

The risk of $m_{\infty,n}$ is defined by

$$R(m_{\infty,n}) = \mathbb{E}[m_{\infty,n}(\mathbf{X}) - m(\mathbf{X})]^2, \tag{2}$$

while the risk of $m_{M,n}$ equals

$$R(m_{M,n}) = \mathbb{E}[m_{M,n}(\mathbf{X}, \Theta_1, \ldots, \Theta_M) - m(\mathbf{X})]^2. \tag{3}$$

It is stressed that both risks $R(m_{\infty,n})$ and $R(m_{M,n})$ are deterministic since the expectation in (2) is over $\mathbf{X}$, $\mathcal{D}_n$, and the expectation in (3) is over $\mathbf{X}$, $\mathcal{D}_n$ and $\Theta_1, \ldots, \Theta_M$. Throughout the paper, we say that $m_{\infty,n}$ (resp. $m_{M,n}$) is $\mathbb{L}^2$ consistent if $R(m_{\infty,n})$ (resp. $R(m_{M,n})$) tends to zero as $n \to \infty$.

As mentioned earlier, there is a large variety of forests, depending on how trees are grown and how the randomness Θ influences the tree construction. For instance, tree construction can be independent of $\mathcal{D}_n$ [1], depend only on the $\mathbf{X}_i$'s [2] or depend on the whole training set [10,17,39]. Throughout the paper, we use Breiman's [7] forests and uniform forests to exemplify our results. In Breiman's [7] original procedure, splits depend on the whole sample and are performed to minimize variance within the two resulting cells. The algorithm stops when each cell contains less than a small pre-specified number of points (typically, 5 in regression). On the other hand, uniform forests are a simpler procedure since, at each node, a coordinate is uniformly selected among $\{1, \ldots, d\}$ and a split position is uniformly chosen in the range of the cell, along the pre-chosen coordinate. The algorithm stops when a full binary tree of level k is built, that is if each cell has been cut exactly k times, where $k \in \mathbb{N}$ is a parameter of the algorithm.

In the rest of the paper, we will repeatedly use the random forest connection function K_n, defined as

$$\begin{aligned} K_n : [0, 1]^d \times [0, 1]^d &\to [0, 1] \\ (\mathbf{x}, \mathbf{z}) &\mapsto \mathbb{P}_\Theta \left[\mathbf{x} \overset{\Theta}{\leftrightarrow} \mathbf{z}\right], \end{aligned}$$