Paper Review: Consistency of Random Forests

Erwan Scornet, Gerard Biau, Jean-Philippe Vert

March 9, 2016

Introduction

Problem Setup

Main Results

Discussion and Future Work

Introduction

Problem Setup

Main Results

Discussion and Future Work

Background

- Random forests is an ensemble learning method for classification and regression that constructs a number of randomized decision trees during the training phase and predicts by averaging the results.
- It was first introduced by Breiman(2001).

Applications

- Random forests can be used to deal with big data and high-dimensional models (Sparsity).
- It is widely used in bioinformatics, survival analysis, quantile regression, ecology, etc.

Introduction

Problem Setup

Main Results

Discussion and Future Work

Decision Trees

- Decision trees can be applied to both regression and classification problems.
- Regression trees are used to predict a quantitative response and classification trees are used to predict a qualitative response.

Regression Trees

- Given a training sample $D_n = (X_1, Y_1), \dots, (X_n, Y_n)$ in $[0,1]^p \times R$, the objective is to estimate $m_n : [0,1]^p \to R$ of the function $m(x) = E\left[Y | X = x\right]$.
- p: dimension of predictors; n: size of the sample (training sets)
- X: input random vector used to estimate $\hat{Y}$, i.e. $m_n(X)$.
- In general trees, $X \in R^p$ rather than $[0,1]^p$ stated in this paper.

Example of Regression Trees

- "Hitters" data set
- Response: log salary of a baseball play
- Predictors: number of years that he has played in the major leagues; number of hits that he made in the previous year.

Example of Regression Trees

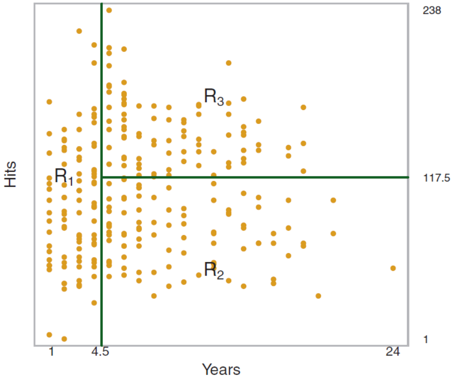


Figure : The three-region partition for "Hitters" data set from regression tree

Example of Regression Trees

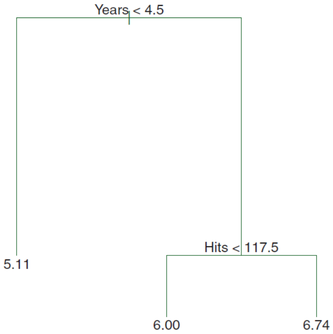


Figure : The regression tree for "Hitters" data set

Notations of Regression Trees

- $m_{try} \in \{1, \ldots, p\}$: number of pre-selected directions for splitting, when $m_{try} < p$, it can used to dealt with high-dimensional data.
- $a_n \in \{1, \ldots, n\}$: number of sampled data points in each tree. If $a_n < n$, it is sub-sampling, and can be used to deal with Big Data.
- $t_n \in \{1, \ldots, a_n\}$: number of leaves(cells) in each tree. If $t_n < a_n$, trees are not fully developed; if $t_n = a_n$, trees are fully developed, i.e. each leave has one number.
- A: a generic cell; $N_n(A)$: number of data points falling in A.
- j: direction of predictor of jth splitting; z: position of cut along the jth coordinate.
- $A_L = \left\{x \in A : x^{(j)} < z\right\}$, $A_R = \left\{x \in A : x^{(j)} \geq z\right\}$

CART-Split Criterion of Regression Trees

- $L_n(j,z) = \frac{1}{N_n(A)} \sum_{i=1}^{n} (Y_i - \overline{Y}_A)^2 1_{X_i \in A} - \frac{1}{N_n(A)} \sum_{i=1}^{n} (Y_i - \overline{Y}_{A_L} 1_{X_i^{(j)} < z} - \overline{Y}_{A_R} 1_{X_i^{(j)} \geq z})^2 1_{X_i \in A}$.
- $(j_n^*, z_n^*) = \underbrace{argmax}_{j \in M_{try}, (j,z) \in C_A} L_n(j,z)$.
- M_{try}: the set of selected predictors to build the tree.
- C_A: the set of all possible cuts in A.
- By CART-Splitting Criterion, we will build a regression tree with m_{try} predictors, a_n data points and t_n leaves.

Classification Trees

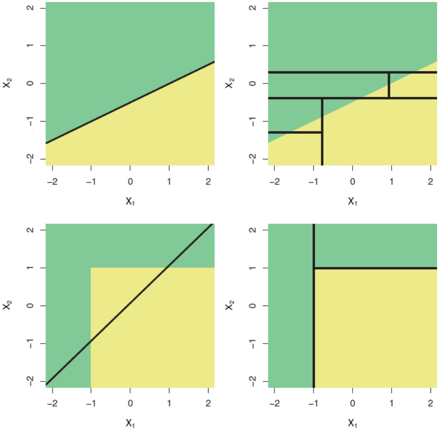


Figure : Decision Trees vs Linear Regression

Advantages and Disadvantages of Trees

- **Advantages:**

  1. Trees are very easy to explain to people.
  2. Decision trees are more close to human decision-making mode.
  3. Trees can be displayed graphically.
  4. Trees can easily handle qualitative predictors without the need to create dummy variables.

- **Disadvantages:**

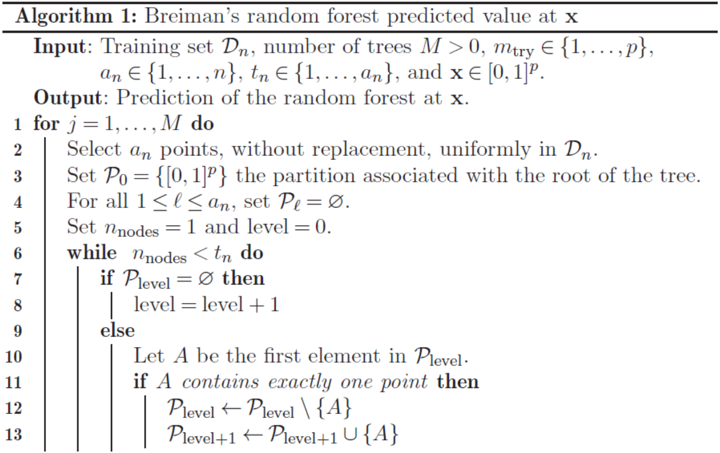
  1. Trees generally don't have the same level of predictive accuracy as other approaches.
  2. Trees can be very non-robust. A small change in the data may cause a large change in the final estimated tree.

- So, we need random forests method to overcome these disadvantages.

Random Forests

- Random forests contains many trees by bootstrap M trees with replacement.
- $m_{M,n}(x; \Theta_1, \ldots, \Theta_M, D_n) = \frac{1}{M} \sum_{j=1}^{M} m_n(x; \Theta_j, D_n) \qquad (1)$
- **Notations:**
- x: query point used to predict value of y.
- D_n: training sample.
- $\Theta_1, \ldots, \Theta_M$: independent random variables distributed as a generic random variable Θ.

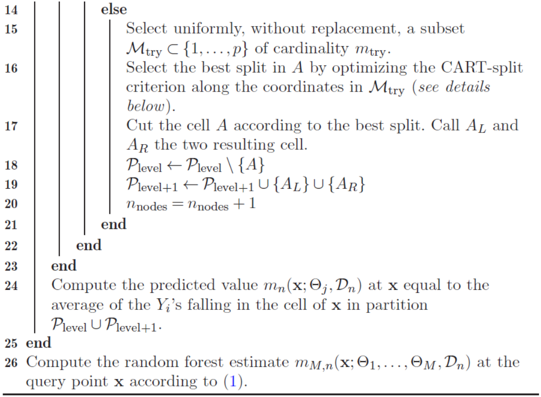
Algorithm of Random Forests



Algorithm 1: Breiman's random forest predicted value at $\mathbf{x}$

```
Input: Training set $\mathcal{D}_n$, number of trees $M > 0$, $m_{\text{try}} \in \{1, \ldots, p\}$,
       $a_n \in \{1, \ldots, n\}$, $t_n \in \{1, \ldots, a_n\}$, and $\mathbf{x} \in [0,1]^p$.
Output: Prediction of the random forest at $\mathbf{x}$.
for $j = 1, \ldots, M$ do
    Select $a_n$ points, without replacement, uniformly in $\mathcal{D}_n$.
    Set $\mathcal{P}_0 = \{[0,1]^p\}$ the partition associated with the root of the tree.
    For all $1 \leq \ell \leq a_n$, set $\mathcal{P}_\ell = \varnothing$.
    Set $n_{\text{nodes}} = 1$ and $\text{level} = 0$.
    while $n_{\text{nodes}} < t_n$ do
        if $\mathcal{P}_{\text{level}} = \varnothing$ then
            $\text{level} = \text{level} + 1$
        else
            Let $A$ be the first element in $\mathcal{P}_{\text{level}}$.
            if $A$ contains exactly one point then
                $\mathcal{P}_{\text{level}} \leftarrow \mathcal{P}_{\text{level}} \setminus \{A\}$
                $\mathcal{P}_{\text{level}+1} \leftarrow \mathcal{P}_{\text{level}+1} \cup \{A\}$
```

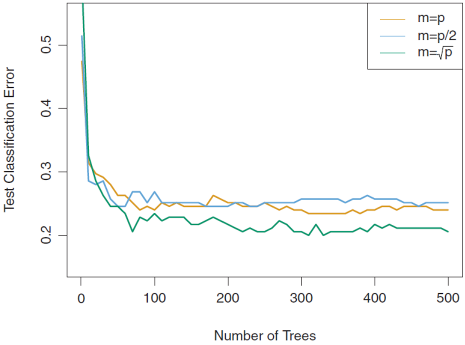
Algorithm of Random Forests



```
            else
                Select uniformly, without replacement, a subset
                $\mathcal{M}_{\text{try}} \subset \{1, \ldots, p\}$ of cardinality $m_{\text{try}}$.
                Select the best split in $A$ by optimizing the CART-split
                criterion along the coordinates in $\mathcal{M}_{\text{try}}$ (see details
                below).
                Cut the cell $A$ according to the best split. Call $A_L$ and
                $A_R$ the two resulting cell.
                $\mathcal{P}_{\text{level}} \leftarrow \mathcal{P}_{\text{level}} \setminus \{A\}$
                $\mathcal{P}_{\text{level}+1} \leftarrow \mathcal{P}_{\text{level}+1} \cup \{A_L\} \cup \{A_R\}$
                $n_{\text{nodes}} = n_{\text{nodes}} + 1$
            end
        end
    end
    Compute the predicted value $m_n(\mathbf{x}; \Theta_j, \mathcal{D}_n)$ at $\mathbf{x}$ equal to the
    average of the $Y_i$'s falling in the cell of $\mathbf{x}$ in partition
    $\mathcal{P}_{\text{level}} \cup \mathcal{P}_{\text{level}+1}$.
end
Compute the random forest estimate $m_{M,n}(\mathbf{x}; \Theta_1, \ldots, \Theta_M, \mathcal{D}_n)$ at the
query point $\mathbf{x}$ according to (1).
```

Figure : Breiman's random forest predicted value at x

Example of Random Forests



- A high-dimensional data of patients.
- n=349; p=500; qualitative response: 1-15 stands for different cancers; predictors: genes.

Introduction

Problem Setup

Main Results

Discussion and Future Work

Assumption 0 (H0)

- **(H0)** We study in this paper the property of the infinite forest estimate obtained as limit of (1) when the number of trees M grows to infinity:

$$m_n(x; D_n) = E_\Theta \left[m_n(x; \Theta, D_n) \right] = lim_{M \to \infty} m_{n,M}(x; \Theta_1, \ldots, \Theta_M, D_n)$$

- We use $m_n(x)$ to denote $m_n(x; D_n)$.

Assumption 1 (H1)

- **(H1)** The response Y follows Additive Regression Model:

$$Y = \sum_{j=1}^{p} m_j(X^{(j)}) + \epsilon,$$

where $X = (X^{(1)}, \ldots, X^{(p)})$ is uniformly distributed over $[0,1]^p$, ϵ is an independent centered Gaussian noise with finite variance $\sigma^2 > 0$, and each component m_j is continuous.

Theorem 3.1

- **Theorem 3.1.** Assume that **(H0)** and **(H1)** are satisfied. Then, provided $a_n \to \infty$ and $t_n (log\, a_n)^9 / a_n \to 0$, random forests are consistent, i.e.,

$$lim_{n \to \infty} E \left[m_n(X) - m(X) \right]^2 = 0.$$

- Interpretations:
  1. $t_n < a_n$: not fully developed trees.
  2. Theorem 3.1 holds for both sub-sampling($a_n < n$) and full sampling($a_n = n$).
  3. $log\, a_n$ is the complexity of the tree; $(log\, a_n)^9$ is used to control Gaussian error.
  4. $a_n / t_n \to 0$ is used to control the correlation between trees, eg. if $a_n = t_n = n$, all trees have the same points and splits, they are completely correlated.

The basic ideas to prove Theorem 3.1

- Step1: $H1$ ⇒Proposition 2; We get that within each cell, variation of $m(X)$ is small.
- Step2: ⇒ There exists as piecewise function f(X) : in each cell, $lim_{n\to\infty} \, inf \; E\left[f(X) - m(x)\right]^2 = 0$.
- Step3: With Theorem 5.1 (Gyorfi et. at. 2002), ⇒ $lim_{n\to\infty} E\left[T_{\beta_n} m_n(X, \Theta) - m(x)\right]^2 = 0$ where $T_{\beta_n} u = u \, if |u| < \beta_n, = sign(u)\beta_n \, if |u| > \beta_n$ is a threshold function; $\beta_n = \|m\|_\infty + \sigma\sqrt{2}(log \; a_n)^2$

The basic ideas to prove Theorem 3.1

- Step4:

$$
\begin{aligned}
\mathbb{E}\big[m_n(\mathbf{X}) - m(\mathbf{X})\big]^2 &= \mathbb{E}\big[\mathbb{E}_\Theta[m_n(\mathbf{X},\Theta)] - m(\mathbf{X})\big]^2 \\
&\le \mathbb{E}\big[m_n(\mathbf{X},\Theta) - m(\mathbf{X})\big]^2 \\
&\quad \text{(by Jensen's inequality)} \\
&\le \mathbb{E}\big[m_n(\mathbf{X},\Theta) - T_{\beta_n} m_n(\mathbf{X},\Theta)\big]^2 \\
&\qquad + \mathbb{E}\big[T_{\beta_n} m_n(\mathbf{X},\Theta) - m(\mathbf{X})\big]^2
\end{aligned}
$$

- Step5: right → 0 proved in step 3; for left, $T_\beta m_n(X,\Theta) = m_n(X,\Theta)\, if \, |m_n(X,\Theta)| < \beta_n$, for outside region, we can use Markov's Inquality to prove.

Assumption 2 (H2)

- **(H2)** Let $Z_{i,j} = (Z_i, Z_j')$. Then, one of the following two conditions holds:
- **(H2.1)** One has:

  $$lim_{n\to\infty}(log\ a_n)^{2p-2}(log\ n)^2 E\left[max_{i\neq j}|\psi_{i,j}(Y_i, Y_j) - \psi_{i,j}|\right]^2 = 0.$$

- **(H2.2)** There exist a constant $C > 0$ and a sequence $(\gamma_n)_n \to 0$ such that, almost surely,

  $$max_{l_1,l_2=0,1}\frac{\left|Corr(Y_i - m(X_i), 1_{Z_{i,j}=(l_1,l_2)}|X_i, X_j, Y_j)\right|}{P^{1/2}\left[Z_{i,j} = (l_1, l_2)|X_i, X_j, Y_j\right]} \leq \gamma_n,$$

  and

  $$max_{l_1=0,1}\frac{\left|Corr((Y_i - m(X_i))^2, 1_{Z_i=l_1}|X_i)\right|}{P^{1/2}\left[Z_i = l_1|X_i\right]} \leq C.$$

Notations of Assumption 2 (H2)

- $Z_i = 1_{X \overset{\Theta}{\leftrightarrow} X_i}$: the indicator that X_i falls in the same cell as X in the random tree designed with D_n and the random parameter Θ.
- $Z_j' = 1_{X \overset{\Theta'}{\leftrightarrow} X_j}$: Θ' is an independent copy of Θ.
- $\psi_{i,j}(Y_i, Y_j) = E\left[Z_i Z_j' | X, \Theta, \Theta', X_1, \ldots, X_n, Y_i, Y_j\right]$.
- $\psi_{i,j} = E\left[Z_i Z_j' | X, \Theta, \Theta', X_1, \ldots, X_n\right]$.

Assumption 2 (H2)

- Interpretations:

  1. **(H2.1)** means that the influence of two Y-values on the probability of connection of two couples of random points tends to be zero as $n \to \infty$.
  2. **(H2.2)** means that the correlation between the noise and the probability of connection of two couple of points vanishes fast enough as $n \to \infty$.
  3. In practice, both **(H2.1)** and **(H2.2)** may not be satisfied.

Theorem 3.2

- **Theorem 3.2.** Assume that **(H0)**, **(H1)** and **(H2)** are satisfied and let $t_n = a_n$. Then, provided $a_n \to \infty$ and $\frac{a_n log\ n}{n} \to 0$, random forests are consistent, i.e.,

$$lim_{n\to\infty} E\left[m_n(X) - m(X)\right]^2 = 0.$$

- Interpretations:
  1. $t_n = a_n$ means fully developed trees.
  2. $a_n < n$ means sub-sampling.
  3. $a_n/n \to 0$ is used to control the correlation between trees, as each observation will not show up in too many trees.

Proposition 1

- **Proposition 1.** Assume that **(H0)** and **(H1)** are satisfied. Let $k \in N^*$ and $\xi > 0$. Assume that there is no interval [a,b] and no $j \in \{1, \dots, S\}$ such that m_j is constant on [a,b], i.e. $Y = \sum_{j=1}^{S} m_j(X^{(j)}) + \epsilon$. Then, with probability $1 - \xi$, for all n large enough, we have, for all $1 \leq q \leq k$,

$$j_{q,n}(X) \in \{1, \dots, S\} .$$

- Interpretations:

  1. Proposition 1 proves the consistency of random forests in high-dimensional settings.
  2. For sparsity, when n is large enough, random forests will pick up the S informative predictors.
  3. This prosition requires $p \leq n$. It doesn't cover the case: $p > n$.

Proposition 2

- **Proposition 2.** Assume that **(H0)** and **(H1)** hold. Then, for all $\rho, \xi > 0$, there exists $N \in N^*(\mathit{num\,of\,cuts})$ such that, for all $n > N$,

$$P\left[\Delta(m, A_n(X, \Theta)) \leq \xi\right] \geq 1 - \rho.$$

- Interpretations:
  1. Proposition 2 states that under H1, the variation of the regression function m within a cell of a random tree is small provided n is large enough.
  2. This proposition forces the approximation error of the forests to asymptotically approach 0.

Introduction

Problem Setup

Main Results

Discussion and Future Work

Discussion

- **Advantages:**

  1. Proved consistency of random forests in subsampling.

  2. Proved consistency of random forests in high-dimensional settings (Sparsity).

- **Disadvantages:**

  1. Predictors are limited in $X \in [0,1]^p$.

  2. Random forests are created without replacement which is different from Bootstrap.

  3. Only the regression model is studied.

  4. The high-dimensional setting doesn't include the situation when $p > n$.

Future Work

- Prove consistency of ramdom forests with predictors $X \in R^p$.
- Prove consistency of random forests by Bootstrap with replacement .
- Prove consistency of ramdom forests of classification.
- Prove consistency of ramdom forests in high-dimensional settings when $p > n$.

Reference

- Breiman, L., *Random Forests*. Machine Learning, 45:5-32,2001.
- James, G., Witten, D., Hastie, T., and Tibshirani, R.*An Introduction to Statistical Learning*. First Edition. Springer,New York,2015.
- Scornet, E., Biau, G., and Vert, J. *Consistency of Random Forests*. Annals of Statistics, 43:1716-1741,2015.
- Gyorfi, M., Krzyzak, A., and Walk, H. *A Distribution-Free Theory of Nonparametric Regression*. Springer, New York, 2002.

THANK YOU!