Econometrics **October 7, 2023**

Topic 20: Random Forest

by Sai Zhang

Key points: .

Disclaimer: *The note is built on Prof. Jinchi Lv's lectures of the course at USC, DSO 607, High-Dimensional Statistics and Big Data Problems.*

20.1 Motivation

Denote by $m(\mathbf{X})$ the measurable nonparametric regression function with p-dimensional random vector $\mathbf{X}$ taking values in $[0,1]^p$. The Random Forest algorithm aims to learn the regression function in a nonparametric way based on the observations $\mathbf{x}_i \in [0,1]^p$, $y_i \in \mathbb{R}$, $i = 1, \cdots, n$, from the model

$$y_i = m(\mathbf{x}_i) + \epsilon_i$$

where $\mathbf{X}, \mathbf{x}_i, \epsilon_i, i = 1, \cdots, n$ are independent, and $\{\mathbf{x}_i\}$ and $\{\epsilon_i\}$ are two sequences of identically distributed random variables. $\mathbf{x}_i$ is distributed identically as $\mathbf{X}$.

Why Random Forest (RF)? **RF** has gained significant popularity due to its

- **High accuracy**: **RF** consistently rank among the top performer, often surpassing more complex models
- **Robustness**: **RF** are less subject to overfitting due to the ensemble nature leveraging multiple decision trees
- **Interpretability**: **RF** provide rankings of feature importance

As illustrated in Figure 20.1, in a level-2 tree, each node (cell) defines the point where the current cell split and new cells are produced. The sets of features eligible for splitting cells at level $k-1$ are denoted as $\Theta_k := \{\Theta_{k,1}, \cdots, \Theta_{k,2^{k-1}}\}$, where $\Theta_{k,s} \subset \{1, \cdots, p\}$.

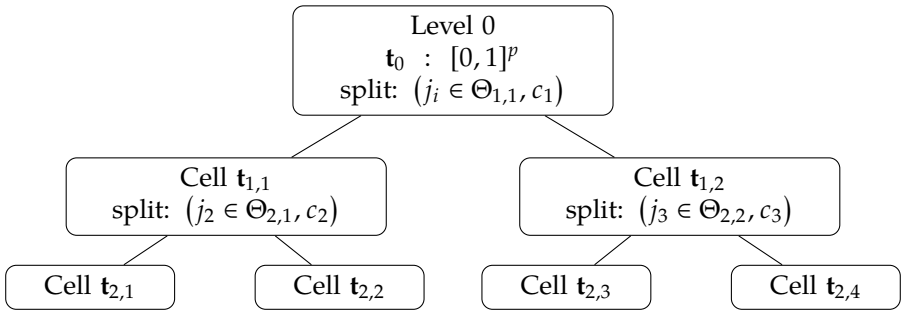

Figure 20.1: Level-2 Tree Example

Given any T (and the associated splitting criterion) and $\Theta_{1:k}$, the tree estimate denoted as $\hat{m}_{T(\Theta_{1:k})}$ for a test

point $\mathbf{c} \in [0,1]^p$ is defined as

$$\hat{m}_{T(\Theta_{1:k})}(\mathbf{c}, \mathcal{X}_n) := \sum_{(\mathbf{t}_1, \cdots, \mathbf{t}_k) \in T(\Theta_{1:k})} \mathbf{1}_{\mathbf{c} \in \mathbf{t}_k} \left(\frac{\sum_{i \in \{i: \mathbf{x}_i \in \mathbf{t}_k\}} y_i}{\#\{i : \mathbf{x}_i \in \mathbf{t}_k\}} \right)$$

where $\mathcal{X}_n := \{\mathbf{x}_i, y_i\}_{i=1}^n$, the fraction is defined as 0 when no sample is in the cell $\mathbf{t}_k$, and $\mathbf{1}_{\mathbf{c} \in \mathbf{t}_k}$ is an indicator function = 1 if $\mathbf{c} \in \mathbf{t}_k$ and = 0 otherwise.

20.2 Chi et al. (2022): High Dimensional RFs

For a RF model where

- a sequence of distinct $\Theta_{1:k}$ results in a distinct tree
- every set of available features $\Theta_{l,s}$, $l = 1, \cdots, k$; $s = 1, \cdots, 2^{l-1}$

Column subsampling Define a **column subsampling** procedure: $\Theta_{l,s}, \forall l, s$ has $[\gamma_0 p]$ distinct integers among $1, \cdots, p$, with $[\cdot]$ the ceiling function for some $0 < \gamma_0 \leq 1$. γ_0 is the predetermined constant parameter of column subsampling. Introduce the boldface random mappings $\mathbf{\Theta}_{1:k}$, which are independent and uniformly distributed over all possible $\Theta_{1:k}$ for all integer k. Then random forests estimate for $\mathbf{c}$ with observations $\mathcal{X}_n$ is given by

$$\mathbb{E}\left(\hat{m}_{T(\mathbf{\Theta}_{1:k})}(\mathbf{c}, \mathcal{X}_n) \mid \mathcal{X}_n\right) = \sum_{\Theta_{1:k}} \mathbb{P}\left(\bigcap_{s=1}^{k} \{\mathbf{\Theta}_s = \Theta_s\}\right) \hat{m}_{T(\Theta_{1:k})}(\mathbf{c}, \mathcal{X}_n)$$

The expectation is taken over sets of available features.

Observation resampling Let $A = \{a_1, \cdots, a_B\}$ be a set of subsamples with each a_i consisting of $\lceil bn \rceil$ observations (indices) drawn without replacement from $\{1, \cdots, n\}$ for some positive integer B and $0 < b \leq 1$; in addition, each a_i is independent of model training. The default values of B and b are 500 and 0.632[1]. Then the tree estimate using subsample a is define as

$$\hat{m}_{T(\Theta_{1:k}),a}(\mathbf{c}, \mathcal{X}_n) := \sum_{(\mathbf{t}_1, \cdots, \mathbf{t}_k) \in T(\Theta_{1:k})} \mathbf{1}_{\mathbf{c} \in \mathbf{t}_k} \left(\frac{\sum_{i \in a \cap \{i: \mathbf{x}_i \in \mathbf{t}_k\}} y_i}{\#(a \cap \{i : \mathbf{x}_i \in \mathbf{t}_k\})} \right)$$

the random forests estimate given A is then

$$B^{-1} \sum_{a \in A} \mathbb{E}\left[\hat{m}_{T,a}(\mathbf{\Theta}_{1:k}, \mathbf{c}, \mathcal{X}_n) \mid \mathcal{X}_n\right] := B^{-1} \sum_{a \in A} \mathbb{E}\left[\hat{m}_{T(\mathbf{\Theta_{1:k}}),a}(\mathbf{c}, \mathcal{X}_n) \mid \mathcal{X}_n\right]$$

CART-split criterion Given a cell $\mathbf{t}$, a subset of observation indices a and a set of available features $\Theta \subset \{1, \cdots, p\}$, the CART-split is defined as

$$(\hat{j}, \hat{c}) = \arg \min_{j \in \Theta, c \in \{x_{ij}: \mathbf{x}_i \in \mathbf{t}, i \in a\}} \left[\sum_{i \in a \cap P_L} (\overline{y}_L - y_i)^2 + \sum_{i \in a \cap P_R} (\overline{y}_R - y_i)^2 \right] \tag{20.1}$$

[1] Or, $b = 1$ but observations are drawn with replacement.

where

$$P_L \coloneqq \left\{ i : \mathbf{x}_i \in \mathbf{t}, x_{ij} < c \right\} \qquad P_R \coloneqq \left\{ i : \mathbf{x}_i \in \mathbf{t}, x_{ij} \geq c \right\}$$
$$\overline{y}_L \coloneqq \sum_{i \in a \cap P_L} \frac{y_i}{\#(a \cap P_L)} \qquad \overline{y}_R \coloneqq \sum_{i \in a \cap P_R} \frac{y_i}{\#(a \cap P_R)}$$

The CART-split criterion conditional on the sample is a deterministic splitting criterion; conditioning on another sample leads to another deterministic splitting criterion. Define $\hat{T}_a$ as the sample tree growing rule that is associated with a splitting criterion following Eq. (20.1), the tree estimates using $\hat{T}_a$ can be similarly defined as

$$\hat{m}_{\hat{T}_a(\Theta_{1:k})}(\mathbf{c}, \mathcal{X}_n) \coloneqq \sum_{(\mathbf{t}_1, \cdots, \mathbf{t}_k) \in \hat{T}_a(\Theta_{1:k})} \mathbf{1}_{\mathbf{c} \in \mathbf{t}_k} \left(\frac{\sum_{i \in \{i : \mathbf{x}_i \in \mathbf{t}_k\}} y_i}{\#\{i : \mathbf{x}_i \in \mathbf{t}_k\}} \right)$$

the definition is the same for $\hat{m}_{\hat{T}_a, a}$. Then the random forests estimate for a test point $\mathbf{c} \in [0,1]^p$ is given by

$$B^{-1} \sum_{a \in A} \mathbb{E}\left(\hat{m}_{\hat{T}_a, a}(\mathbf{\Theta}_{1:k}, \mathbf{c}, \mathcal{X}_n) \mid \mathcal{X}_n \right)$$

where the average and conditional expectation correspond to the sample and column subsamplings respectively, and they are interchangeable.

Bias-variance decomposition For a tree growing rule T and $\Theta_{1:k}$, the population version is defined as

$$m^*_{T(\Theta_{1:k})}(\mathbf{c}) \coloneqq \sum_{(\mathbf{t}_1, \cdots, \mathbf{t}_k) \in T(\Theta_{1:k})} \mathbf{1}_{\mathbf{c} \in \mathbf{t}_k} \mathbb{E}\left(m(\mathbf{X}) \mid \mathbf{X} \in \mathbf{t}_k \right) \tag{20.2}$$

for each test point $\mathbf{c} \in [0,1]^p$. And the $\mathbb{L}^2$ prediction loss for random forests is defined as

$$\mathbb{E}\left[m(\mathbf{X}) - B^{-1} \sum_{a \in A} \mathbb{E}\left(\hat{m}_{\hat{T}_a, a}(\mathbf{\Theta}_{1:k}, \mathbf{X}, \mathcal{X}_n) \mid \mathbf{X}, \mathcal{X}_n \right) \right]^2 \tag{20.3}$$

if we use the full sample $a = \{1, \cdots, n\}$, and denote $\hat{T}_a$ and $\hat{m}_{\hat{T}_a, a}$ as $\hat{T}$ and $\hat{m}_{\hat{T}}$, the sample subsampling and average $B^{-1} \sum_{a \in A}(\cdot)$ in the random forests estimate are no longer needed, then Eq.(20.3) can be simplified as

Chi et al. (2022)

References

Chien-Ming Chi, Patrick Vossler, Yingying Fan, and Jinchi Lv. Asymptotic properties of high-dimensional random forests. *The Annals of Statistics*, 50(6):3415–3438, 2022.