Econometrics October 3, 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 $\mathbf{x}_i \in [0,1]^p$, $y_i \in \mathbb{R}$, $i = 1, \dots, n$, from the model

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

where X, x_i , ϵ_i , $i = 1, \dots, n$ are independent, and $\{x_i\}$ and $\{\epsilon_i\}$ are two sequences of identically distributed random variables. x_i is distributed identically as 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 $\Theta_{k,s} \subset \{1, \cdots, p\}$.

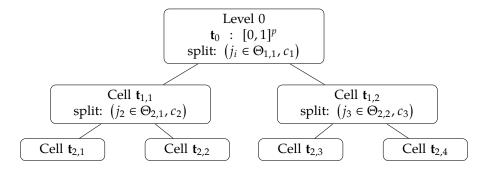


Figure 20.1: Level-2 Tree Example

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

20-2 Week 20: Random Forest

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

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

where $X_n := \{x_i, y_i\}_{i=1}^n$, 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 $1, \dots, p$, with $[\cdot]$ the ceiling function for some $0 < \gamma_0 \le 1$. γ_0 is the predetermined constant parameter of column subsampling. Introduce the boldface random mappings $\Theta_{1:k}$, 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 X_n is given by

$$\mathbb{E}\left(\hat{m}_{T(\boldsymbol{\Theta}_{1:k})}\left(\mathbf{c}, X_{n}\right) \mid X_{n}\right) = \sum_{\boldsymbol{\Theta}_{1:k}} \mathbb{P}\left(\bigcap_{s=1}^{k} \left\{\boldsymbol{\Theta}_{s} = \boldsymbol{\Theta}_{s}\right\}\right) \hat{m}_{T(\boldsymbol{\Theta}_{1:k})}\left(\mathbf{c}, X_{n}\right)$$

The expectation is taken over sets of available features.

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

Chi et al. (2022)

Week 20: Random Forest 20-3

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.