A Theoretical Analysis of Random Forest Models for Imputation, Prediction and Variable Selection

Dortmund Data Science Center

Dortmund, November, 26, 2020
Burim Ramosaj, graduated at Ulm University and Syracuse University in Mathematics and Management

PhD-Studies at TU Dortmund University. Thesis title: *Analyzing Consistency and Statistical Inference in Random Forest Models*. Supervisor: Prof. Dr. Markus Pauly and Prof. Dr. Jörg Rahnenführer

Post-doc at the Institute of Mathematical Statistics and Applications in Industry within DFG project PA 2409/3-2.

From 2021 on: Own project on *Statistical Inference Analysis with Machine Learning* funded from the state of NRW.

Research focus of our group:

- Machine Learning (ML) Methods for Inference and Prediction
- Random Forest and Ensemble Methods
- Resampling
- Statistical Inference with Missing Values
- Survival Analysis and Time Series Analysis
My research work

Focus on **Random Forest Models (RF)** for the following purposes:

- Imputing missing values.
- Inference after imputation.
- Uncertainty quantification using Random Forests.
- Consistency and (un-)biasedness for RF based measures and estimators.
- Variable selection with Random Forest based importance measures:
  - Absolute Number of Selection Frequency.
  - Gini Importance.
  - Permutation Importance.
- Combination with Neural Networks (structural relations).
1. Multiple imputation and Random Forest

We focus on imputation procedures, specifically, **multiple imputation (MI)**. It has the following advantages:

- Easy to implement.
- Standard statistical analysis can be applied.
- Theoretical guarantees for its *correct* usage exists (Rubin, 2004).
- Reflects uncertainty of the data generating process.

Combination with research and application trend of involving ML-based methods, like the RF:

- No statistical modeling required.
- Treating simultaneously categorical and continuous variables.
- Low tuning efforts with comparably high predictive accuracy.
1. Multiple imputation and Random Forest

- Highly cited paper of Stekhoven and Bühlmann (2012) when using missForest for imputation (1,390 citations as of today).

We measured imputation accuracy for various techniques (Ramosaj et al., 2020):

- Normalized Root Mean Squared Error (NRMSE) for:
  (1) RFMI, (2) RFMICE, (3) PMM and (4) NORM
  - missForest performs best under the MI framework (RFMI).
  - Multiple Imputation Using Chained Equations under the Bayesian regression model with Gaussian assumption (NORM) performs worse.
1. Multiple imputation and Random Forest

Measuring **type-I error in paired data**

- $T_{ML}$ (---), RF MI (···), RF MICE (−·−), PMM (−), NORM (- - -) for $\chi^2$ distribution under $\rho = 0.1$ and $\Sigma_1$ for varying $k$ values multiplied to $(30, 10, 10)$, i.e. $(n_1, n_2, n_3) = k \cdot (30, 10, 10)$.
- RFMI (missForest) highly inflates type-I error.
- Reason: No multiple imputation; incorrect treatment of uncertainty.
- This is even provable (Ramosaj, 2020).
Future research work on imputation

- Influence of the often-used NRMSE (normalized root mean squared error) and PFC (proportion of false classification) measure on the prediction accuracy such as mis-classification and mean-squared error.

- Coverage of prediction intervals after imputation with ML-based imputation techniques.

- General issue of constructing correct and valid prediction intervals.
RF models are also used for selecting **informative** variables.

- Absolute Number of Selection Frequency
- Mean Decrease in Impurity
- Permutation Importance

Simulation-based evidence that RF variable selection is:

- biased (Strobl et al., 2007) exists. No **theoretical guarantees** so far!
2. Variable selection with Random Forest

We could show the following results (Ramosaj and Pauly, 2019):

- The permutation importance measure $I_{n,M}^{OOB}(j)$ for regression learning problems is asymptotically unbiased, i.e. there is a constant $C > 0$ such that
  \[
  \lim_{n \to \infty} \lim_{M \to \infty} \mathbb{E}[I_{n,M}(j)] = C \cdot 1\{j \in S\}
  \]
  and $S$ is the informative set.

- Under slightly stronger assumption, we can also show that
  \[
  I_{n,M}^{OOB}(j) \overset{\mathbb{P}}{\longrightarrow} C \cdot 1\{j \in S\}, \text{ as } (n, M) \overset{\text{seq}}{\to} \infty.
  \]
Future research work on variable selection

We aim to detect influential variables through formal testing of:

\[ H_0 : I_{n,M}^{OOB}(j) = 0 \text{ vs. } I_{n,M}^{OOB}(j) \neq 0 \]

The following approaches might be considered:

- Parametric approach, i.e. finding an appropriate limit distribution.
- Non-parametric approach by using permutation tests.

We received separate funding from the state of NRW for this specific research question.
Current collaborations

Thanks to my collaboration partners:

Prof. Dr. Markus Pauly
Faculty of Statistics, TUD

Prof. Dr. Jian-Jia Chen
Faculty of Informatics, TUD

Prof. Dr. Jörg Rahnenführer
Faculty of Statistics, TUD

Prof. Dr. Erwan Scornet
Ecole Polytechnique

Prof. Dr. Gérard Biau
Sorbonne Center for Artificial Intelligence, Paris

Prof. Dr. Lucas Mentch
Department of Statistics,
Pittsburgh University
Thank you for your attention!
Bibliography I


Main issue in statistical inference after imputation:

**Correct Uncertainty Quantification**

- **Use multiple imputation logic** as given in Rubin (2004): Interested in testing
  \[ H_0 : Q = Q_0 \quad \text{vs.} \quad H_0 : Q \neq Q_0 \]

- Generate \( m \in \mathbb{N} \) data sets \( D^{(1)}_n, \ldots, D^{(m)}_n \) from the initial data set.

- Calculate for every data the estimator \( Q_{n,t} \) as in standard complete procedures and estimate its variance \( U_{n,t} \) similarly.

- Aggregation of the estimator: \( \bar{Q}_{n,m} = \frac{1}{m} \sum_{t=1}^{m} Q_{n,t} \) und \( \bar{U}_{n,m} = \frac{1}{m} \sum_{t=1}^{m} U_{n,t} \).

- Estimate variance of \( \bar{Q}_{n,m} \) through \( \bar{U}_{n,m} + (1 + 1/m) B_{n,m} \),
  \[ B_{n,m} = \frac{1}{m-1} \sum_{t=1}^{m} (Q_{n,t} - \bar{Q}_{n,m})(Q_{n,t} - \bar{Q}_{n,m})^\top. \]
Appendix: Variable Selection

Random Forest models are also used in variable selection. Different measures do exists:

- **Absolute Number of Selection Frequency**

  $$ABS_{n,M}(j) = \frac{1}{M} \sum_{t=1}^{M} \sum_{k=1}^{t_n-1} 1 \{ \text{variable } j \text{ selected in tree } t \text{ at node } t_n \}$$

- **Mean Decrease in Impurity**

  $$MDI_{n,M}(j) = \frac{1}{M} \sum_{t=1}^{M} \sum_{k,s} \frac{|N_n(A^{(k)}_{n,s}(\Theta_t))|}{n} \cdot L^{(k)}_{n,s}(j, z_j) \cdot 1 \left\{ L^{(k)}_{n,s}[j, z_j] \geq L^{(k)}_{n,s}[\ell, z_\ell], \forall \ell \neq j, z_j, z_\ell \right\}$$

- **Permutation Importance**

  $$I^{OOB}_{n,M}(j) = \frac{1}{\gamma_n M} \sum_{t=1}^{M} \sum_{i \in D_{n-t}} \left\{ \psi \left( Y_i, m^{OOB}_{n,M}(X^\pi_{i,t}; \Theta_t) \right) - \psi \left( Y_i, m^{OOB}_{n,M}(X_i; \Theta_t) \right) \right\}$$

  \(\psi\) is some loss function, depending on the context.