TREE-BASED MULTIPLE IMPUTATION METHODS

Michael Dellermann, Anatol Sluchych, and Jonah Wermter

Freie Universität Berlin

1. Motivation

- Standard MICE approach: conditional models to be specified for *all* variables with missing data
- Still may fail to capture interactive and nonlinear relations among variables as well as non-standard distributions
- Classification and regression trees (CART) *automatically* capture interactions, nonlinear relations, and complex distributions with no parametric assumptions or data transformations needed (Burgette & Reiter 2010)
- Implementation in R: *mice* and *miceranger* packages

2. Tree-based methods

Description of MICE approach? Detailed description of trees?

CART:

- seek to approximate the conditional distribution of a univariate outcome from multiple predictors
- partition the predictor space so that subsets of units formed by the partitions have relatively homogeneous outcomes
- partitions are found by recursive binary splits of the predictors
- series of splits can be effectively represented by a tree structure, with leaves corresponding to the subsets of units
- values in each leaf represent the conditional distribution of the outcome for units in the data with predictors that satisfy the partitioning criteria that define the leaf

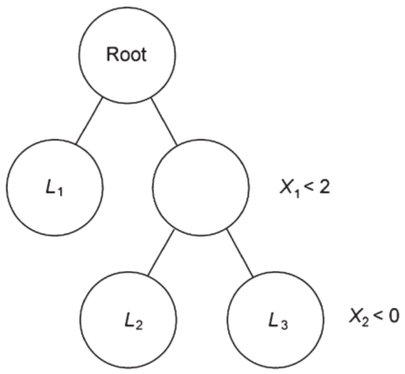

Fig. 1: Example of a tree structure. Source: Burgette & Reiter (2010)

Disadvantages relative to parametric models:

- decreased efficiency when the parametric models are adequate
- discontinuities at partition boundaries
- categorical predictors with many levels can cause computational difficulties

3. Imputation algorithm

Let Y be $n \times p$ the data matrix arranged as $Y = (Y_p, Y_c)$, where

- Y_p consists of p_1 *partially observed* columns, such that moving from left to right, the number of missing elements in each column is nondecreasing
- Y_c remaining completely observed columns
- Y_{obs} set of observed and Y_{mis} set of missing elements

4-steps algorithm:

1. Initial values for the missing values filled in as follows:
   (a) Define a matrix Z equal to Y_c
   (b) Impute missing values in Y_i, where $i = 1, ...p_1$, using CART on Z and append the completed version of Y_i to Z prior to incrementing i
2. Replace the originally missing values of Y_i, where $i = 1, ...p_1$, with CART on Y_{-i}
3. Repeat l times step 2
4. Repeat steps 1–3 m times and obtain m imputed sets.

- sequential CART imputation algorithm
- order the variables from least amount to largest amount of missing data
- minimum leaf size of 5 and the splitting criteria of a deviance greater than 0.0001
- trees are not pruned to minimize bias
- size of trees modulated by requiring a minimum number of observations in each leaf and by controlling the minimum heterogeneity in the values in the leaf in order to consider it for further splitting
- We take draws from the predictive distribution by sampling elements from the leaf that corresponds to the covariate values of the record of interest
- actually perform a Bayesian bootstrap within each leaf before sampling.

4. Simulation study

- interactions among the variables in these domains, rather than main effects alone, are likely to be predictors
- nature of these interactions is not known a prior
- imputations of missing data must be flexible enough to capture the most important interactions in the data
- check the plausibility of our imputation models using posterior predictive checks

5. Results

6. Conclusion

- MICE by CART imputation can result in more reliable inferences compared with naive applications of MICE based on main-effects generalized linear models
- For the quadratic and interaction terms, CART-based MICE results in notably lower mean-squared errors and biases

7. Next steps

- random forests
- neural networks
- Bayesian additive regression trees

References

Burgette, Lane F, and Jerome P Reiter. "Multiple imputation for missing data via sequential regression trees". *American journal of epidemiology* 172, no. 9 (2010): 1070–1076.

Hastie, Trevor, et al. *The elements of statistical learning: data mining, inference, and prediction.* Vol. 2. Springer, 2009.