Optimizing Decision Trees for Clustered and Hierarchical Data

Kevin McCoy[1,2] Christine B. Peterson[2]

[1]Department of Statistics, Rice University [2]The University of Texas MD Anderson Cancer Center

Introduction

- The Classification and Regression Tree (CART) [1] is perhaps the most popular decision tree algorithm used in machine learning. Though all decision trees are weak predictors by themselves, they have shown to perform incredibly well when used in an ensemble, as in a random forest or boosted tree model.
- Decision trees and their extensions are tolerant of missing data, interpretable, and robust against collinearity.
- Some work has been done to incorporate mixed effects or longitudinal data into CART [2, 3, 4]. However, existing methods have limitations in how they handle clustered or hierarchical data.
  - For example, many methods use a rigid additive structure for the mixed effects, similar to a linear mixed model, which takes away from the flexibility of decision trees.
  - Current methods are also only effective at predicting random effects for groups already seen in training data and resort to using the global mean for out-of-sample prediction.
- We propose a lightweight two stage model that first predicts the group membership, and then uses a combination of trees to weight the prediction towards similar training observations.
- Through simulation studies, we highlight the potential of our method in a variety of data settings.

Methods

Decision Tree Modification

- Consider grouped non-i.i.d. data with k groups, n samples per group, and p features. The tabular dataset is thus $(k \cdot n) \times p$, where each block of n samples is generated from one group.
- The proposed method is composed of three stages. See **Figure 1** for a visual explanation.

1. The first stage involves fitting a classification model, such as logistic regression or decision tree classifier, with the group factor as the output.
2. The second stage involves training a decision tree on each group seen in the training data.
3. In the third stage we apply a mixture of trees such that the final output is a linear combination of these group-specific trees with the group classification probabilities as the weights.

Generating Synthetic Test Data

- In order to generate synthetic data, we employ the matrix normal distribution:

$$\mathbf{X} \sim \mathcal{MN}_{k \times p}(\mathbf{M}, \mathbf{U}, \mathbf{V})$$

where $\mathbf{M}$ is a $(k \times p)$ location matrix, $\mathbf{U}$ is the $(k \times k)$ group covariance matrix, and $\mathbf{V}$ is the $(p \times p)$ feature covariance matrix. This allows us the freedom to specify a number of correlation structures between not only features, but also groups in hierarchical settings. We generate n samples.

- We construct a continuous output using:

$$y_{i,j} = f(\mathbf{x}_{i,j}) + \mathbf{Z}\boldsymbol{\alpha}_j + \epsilon$$

$$i \in \{1, \ldots, n\}, \quad j \in \{1, \ldots, k\}$$

where $\mathbf{Z}$ is the random effects design matrix, $\boldsymbol{\alpha}_j$ is a random slope and intercept shared by all observations within a group, and $\epsilon \sim \mathcal{N}(0, 0.5^2)$ is a shared noise term. We generate $\boldsymbol{\alpha}_j \sim \mathcal{MN}(\mathbf{0}, \sigma_\alpha * \mathcal{I})$, where $\mathcal{I}$ is the $(p+1 \times p+1)$ identity matrix. We employ Friedman's function [5], given below, for the fixed effects function f.

$$f(\mathbf{x}_{i,j}) = \sin(\pi x_1 x_2) + 2(x_3 - 0.5)^2 + x_4 + 0.5x_5$$

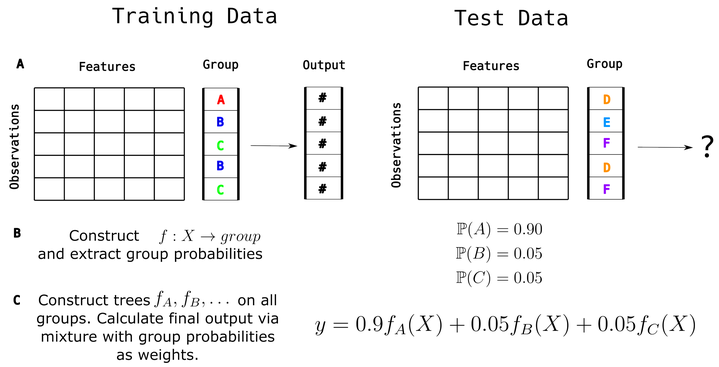

Figure 1. **(A)** Consider grouped data such that each training observation falls into some group (e.g. patients with multiple observations). The test data are similarly structured but come from new groups not seen in the training data. **(B)** We then construct a classifier that predicts the group that each new observation belongs to, and extract the output group probabilities. **(C)** Finally, we construct independent trees on each group seen in the training set. The final output is a linear combination of the predictions from all of these trees with the probabilities from (B) as weights.

Results

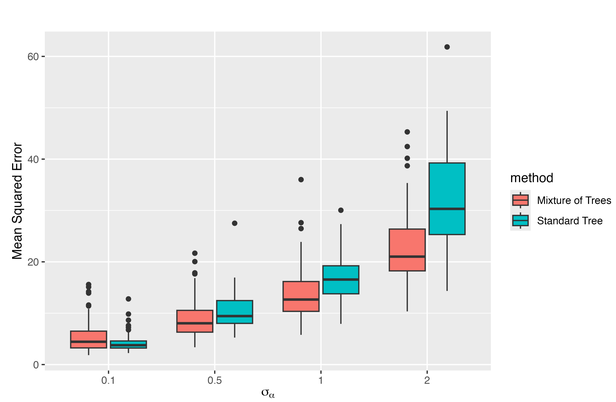

Figure 2. Box and whisker plot of MSE values over a range of σ_α. Tests were performed over N=50 iterations. Data generated with $n = 20$ and $k = 20$.

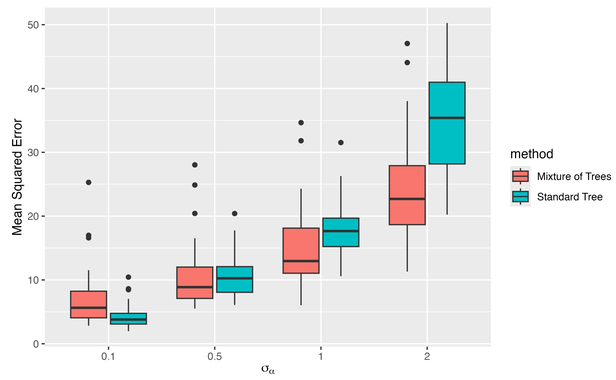

Figure 3. Box and whisker plot of MSE values over a range of σ_α. Tests were performed over N=50 iterations. Data generated with $n = 10$ and $k = 40$.

Results

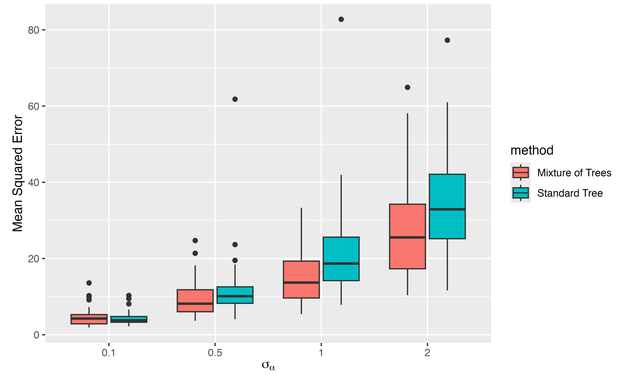

Figure 4. Box and whisker plot of MSE values over a range of σ_α. Tests were performed over N=50 iterations. Data generated with $n = 40$ and $k = 10$.

Discussion

- The mixture of trees method provides consistent improvements over standard trees when the random noise σ_α exceeds the common noise ϵ.
- Furthermore, this result is shared across high, low, and intermediate n and k settings.
- The mixture of trees method also performs as well as random forests, even though the method does not use any bootstrapping of observations or random sampling of features.
- Overall, our work shows that constructing decision trees and forests can be improved when significant random effects are present without the need for rigid model constraints.
- In the future, we plan on investigating how this method translates to categorical outputs as well as other fixed effects functions besides the Friedman function.
- Finally, this work opens the door into how the random forest algorithm can be altered to accommodate random effects. Initial findings suggest that this is possible, but only at the expense of high variance in the model.

References

[1] L. Breiman, ed., *Classification and regression trees*. Boca Raton, Fla.: Chapman & Hall/CRC, 1. crc press repr ed., 1998.

[2] A. Hajjem, F. Bellavance, and D. Larocque, "Mixed effects regression trees for clustered data," *Statistics & Probability Letters*, vol. 81, pp. 451–459, Apr. 2011.

[3] S. Lin and W. Luo, "A New Multilevel CART Algorithm for Multilevel Data with Binary Outcomes," *Multivariate Behavioral Research*, vol. 54, pp. 578–592, July 2019.

[4] J. L. Speiser, B. J. Wolf, D. Chung, C. J. Karvellas, D. G. Koch, and V. L. Durkalski, "BiMM tree: a decision tree method for modeling clustered and longitudinal binary outcomes," *Communications in Statistics - Simulation and Computation*, vol. 49, pp. 1004–1023, Apr. 2020.

[5] J. H. Friedman, "Multivariate Adaptive Regression Splines," *The Annals of Statistics*, vol. 19, Mar. 1991.

Acknowledgements

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. 1842494, as well as The Ken Kennedy Institute Computational Science and Engineering Recruiting Fellowship, funded by the Energy HPC Conference.