HHS Public Access
Author manuscript
Multivariate Behav Res. Author manuscript; available in PMC 2018 January 18.

Published in final edited form as:
Multivariate Behav Res. 2017 ; 52(1): 117–. doi:10.1080/00273171.2016.1265433.

A Gradient Boosting Machine for Hierarchically Clustered Data

Patrick J. Miller, **Daniel B. McArtor**, and **Gitta H. Lubke**
University of Notre Dame

As increasingly larger data sets are collected (big data), exploratory data analysis to develop predictive models becomes more important but more challenging. Variable selection focuses on identifying important predictors and predictors with nonlinear effects. But simple correlation screens, stepwise regression, or ad-hoc model selection procedures can fail to detect these important nonlinear effects. When data are hierarchically clustered (e.g., observations on students within schools), the problem becomes even more challenging. When mixed effect models are used, model selection can produce misleading results when nonlinear effects are not included in the model. Additionally, mixed effect models can be difficult to estimate when there are a large number of predictors with group specific effects.

'Boosted decision trees' is an alternative approach to parametric modeling. It is an additive model of decision trees estimated by gradient descent (Friedman, 2001). This flexible approach can approximate complex and nonlinear functions of predictors without specifying their functional form beforehand; boasted decision trees are easily estimated (e.g., using the R package 'gbm', Ridgeway et al., 2015). However, it is unclear how to best use boosted decision trees with hierarchically clustered data. Including the clustering variable in the model makes the interpretation more difficult. Ignoring the clustering variable misses an opportunity for potentially improved prediction, potentially leading to biased variable selection.

We propose an extension to boosted decision trees that works by allowing the terminal node means of each tree in the model to vary by group at each iteration. We implement the approach using the popular R package 'lme4' (Bates et al., 2014). The method can be used with thousands of predictors, and allows the predictors to contain missing values. Initial results show that this procedure can improve prediction and variable selection performance by as much as 25% compared to including the grouping variable as a candidate for splitting in each tree.

Correspondence concerning this abstract should be addressed to Patrick J. Miller, 110 Haggar Hall, University of Notre Dame, Notre Dame, IN 46656. pmille13@nd.edu.

Conflict of Interest Disclosures: Each author signed a form for disclosure of potential conflicts of interest. No authors reported any financial or other conflicts of interest in relation to the work described.

Ethical Principles: The authors affirm having followed professional ethical guidelines in preparing this work. These guidelines include obtaining informed consent from human participants, maintaining ethical treatment and respect for the rights of human or animal participants, and ensuring the privacy of participants and their data, such as ensuring that individual participants cannot be identified in reported results or from publicly available original or archival data.

Acknowledgments

Funding: The third author is supported by NIDA R37 DA-018673. The computational work was done on clusters acquired through NSF MRI BCS-1229450.

Role of the Funders/Sponsors: None of the funders or sponsors of this research had any role in the design and conduct of the study; collection, management, analysis, and interpretation of data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

The primary author would like to express his appreciation to his advisor and SMEP sponsor, Gitta Lubke, and to Daniel McArtor for his critical insight that motivated the method. The ideas and opinions expressed herein are those of the authors alone, and endorsement by the authors' institution is not intended and should not be inferred.

References

Friedman J. Greedy function approximation: a gradient boosting machine. Annals of Statistics. 2001; 29(5):1189–1232.

Ridgeway, G. with contributions from others. gbm: Generalized Boosted Regression Models. R package version 2.1.1. 2015. https://CRAN.R-project.org/package=gbm

Bates D, Maechler M, Bolker B, Walker S. lme4: Linear mixed-effects models using Eigen and S4. R package version. 2014; 1(7)