ARTICLE IN PRESS

Computational Statistics and Data Analysis xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Should we impute or should we weight? Examining the performance of two CART-based techniques for addressing missing data in small sample research with nonnormal variables

Timothy Hayes *, John J. McArdle

University of Southern California, United States

ARTICLE INFO

Article history:
Received 23 September 2016
Received in revised form 6 May 2017
Accepted 18 May 2017
Available online xxxx

Keywords:
CART
Classification and regression trees
Random forests
Small samples
Missing data
Nonnormality

ABSTRACT

Recently, researchers have proposed a variety of new methods for employing exploratory data mining algorithms to address missing data. Two promising classes of missing data methods take advantage of classification and regression trees and random forests. A first method uses the predicted probabilities of response (vs. non-response) generated by a CART analysis to create inverse probability weights. This method has been shown to perform well in prior simulations when nonresponse was generated by tree-based structures, even under low sample sizes. A second method uses the values falling in terminal nodes of CART trees to generate multiple imputations. In prior studies, these methods performed well at estimating main effects and interactions in regression models when sample sizes were large ($N = 1000$), but their performance was not evaluated under small sample conditions. In the present research, we assess the performance of CART-based weights and CART-based imputations under low sample sizes ($N = 125$ or 250) and nonnormality when missing data are generated by smooth functions (linear, quadratic, cubic, interactive). Results suggest that random forest weights excel under low sample sizes, regardless of nonnormality, whereas CART multiple imputation is more efficient with larger samples ($N = 500$ or 1000).

1. Introduction

Missing data are a prevalent problem in applied research. One challenge of addressing missing data is that, unlike variables of substantive interest, researchers often lack strong a priori hypotheses about the factors that may lead to nonresponse in their datasets. Another potential challenge is that the relations between covariates in the datasets and participants' probabilities of nonresponse may be nonlinear and interactive.

One promising new approach to help address these challenges involves employing exploratory data mining (or *machine learning*) algorithms to help model potentially complex relationships between observed covariates and missing data (cf. Hastie et al., 2009 for a comprehensive overview of statistical learning). In this paper, we focus on two broad approaches to utilizing data mining methods to address missing data. The first method uses data mining techniques to predict participants' probabilities of nonresponse and form inverse probability weights. The second method uses data mining to generate predicted values to be used as imputations to fill in missing cases.

* Correspondence to: University of Southern California, 3620 South McClintock Ave, Seeley G. Mudd (SGM) 501, Los Angeles, CA 90089-1061, United States.
E-mail address: hayest@usc.edu (T. Hayes).

http://dx.doi.org/10.1016/j.csda.2017.05.006
0167-9473/

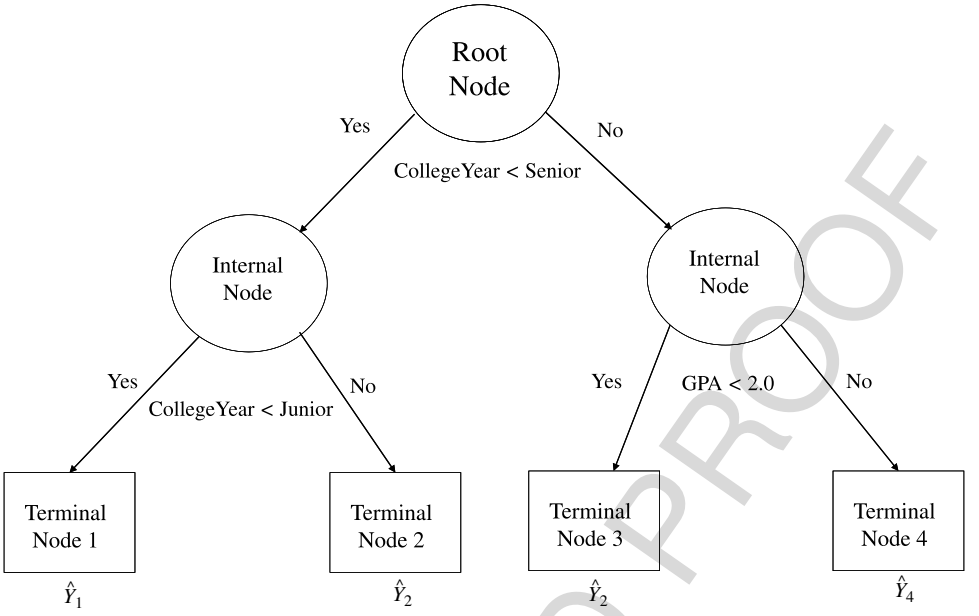

Fig. 1. Example tree diagram.

Although each of these methods performed well in initial studies (Doove et al., 2014; Hayes et al., 2015; Shah et al., 2014), so far we are unaware of any studies that compare their performance. Furthermore, prior research has not systematically assessed the performance of these methods in small samples with nonnormal data. In this paper, we first describe the exploratory data mining techniques that form the basis for the missing data methods under study. Then, we describe how these data mining techniques can be applied to missing data problems using weighting and imputation methods. Next, we highlight important findings and questions from past research on these methods. Finally, we report and discuss a simulation study designed to assess the performance of these data mining methods in addressing missing data in small sample experimental studies, under varying degrees of nonnormality and rates of nonresponse, when missing data are related to a set of observed covariates through a variety of nonlinear and interactive Missing At Random (MAR; Rubin, 1976) missing data mechanisms.

2. Overview of CART, bagging, and random forests

Classification and regression trees (CART; Breiman et al., 1984) is a machine learning algorithm that uses the values of a set of observed predictor variables to split a dataset into homogeneous subgroups with respect to a categorical or continuous dependent variable. A homogeneous group in the categorical case is one in which the group members share the same class membership. In the continuous case, a homogeneous group is one in which the group members share similar values on the continuous outcome, such that the values are closely centered around a common group mean.

Because each split in a CART tree is contingent upon the splits that came before it, CART trees are inherently *conditional*. Thus, when CART splits the dataset on the basis of more than one predictor variable, these variables *interact* in determining the final subgroups in the dataset. Likewise, when CART creates multiple splits at different cutpoints on a single variable, this represents a type of *nonlinearity*, in which the prediction for that variable is not constant across levels or values of the variable (as it would in linear regression, where *any* single-unit change in x at any point in the variable would produce a corresponding change in y).

As an illustration of these concepts, Fig. 1 depicts a tree diagram resulting from a hypothetical CART analysis. On the right-hand side of the diagram, two variables, CollegeYear and GPA, interact in predicting y, indicating that college seniors' scores on the outcome depend upon whether or not their GPAs are less than 2.0. On the left-hand side of the diagram, there is a non-linear prediction on the variable CollegeYear, indicating that participants' ultimate grouping in the analysis upon whether they are juniors or underclassmen (freshmen + sophomores).

One of the main results produced by CART analyses is a set of predicted values of the dependent variable. In the case of a continuous dependent variable, the predicted value for any given case is set equal to the mean of the terminal node (final subgroup) in which it falls. In the case of a categorical dependent variable, CART produces two kinds of predicted values: the *predicted probability* of membership in a given class (e.g., of being classified as a 1 rather than a 0 on a binary outcome) is equal to the proportion of cases in the node who are members of that class, whereas the *predicted class*

Please cite this article in press as: Hayes, T., McArdle, J.J., Should we impute or should we weight? Examining the performance of two CART-based techniques for addressing missing data in small sample research with nonnormal variables. Computational Statistics and Data Analysis (2017), http://dx.doi.org/10.1016/j.csda.2017.05.006.