Creating Imputation Classes Using Classification Tree Methodology

Darryl V. Creel[1] and Karol Krotki[1]
[1]RTI International

Abstract

Virtually all surveys encounter some level of item nonresponse. To address this potential source of bias, practitioners often use imputation to replace missing values with valid values through some form of stochastic modeling. In order to improve the reliabilities of such models, imputation classes are formed to produce homogenous groups of respondents, where homogeneity is measured with respect to the item that will be imputed. A common method used to form imputation classes is Chi-squared Automatic Interaction Detection (CHAID) where the splitting rule is based on Chi-squared tests. This paper examines an alternative methodology used to form imputation classes, nonparametric classification trees where the splitting rules are based on the Gini index of impurity, which is one possible splitting rule used in Classification and Regression Trees (CART). In addition to a brief description of the two classification tree methodologies, we provide some comparative examples using simple generated data and real data. Finally, we use the imputation classes with three imputation procedures: mode value imputation, proportional random imputation, and weighted sequential hot-deck. To provide an additional comparison, we model the item nonresponse using logistic regression or polychotomous regression.

Keywords: Nonresponse, Imputation, Chi-squared Automatic Interaction Detection (CHAID), and Classification and Regression Trees (CART).

1. Introduction

Virtually all surveys encounter some level of item nonresponse. In order to address this item nonresponse, imputation is used to replace missing values. Often practitioners would like to form homogeneous imputation classes that restrict the donor pool to minimize the potential bias. When there are a large number of variables available to form the imputation classes, different methodologies are used to investigate the structure of the data and identify the variables useful in constructing the imputation class. Currently, one of the most common methodologies is Chi-squared Automatic Interaction Detection (CHAID) which creates parametric classification trees. We are proposing nonparametric classification trees based on the Gini index of impurity available in the Classification and Regression Tree methodology (CART) to form imputation classes. Section 2 provides a brief description and example of a classification tree. Section 3 introduces the imputation methodologies. Section 4 discussed the data used for the evaluation. Section 5 explains the evaluation methods. Section 6 describes the results. Finally, Section 7 provides a recommendation

2. Classification Trees

This section provides a brief description and example of a classification tree. The basic structure of the classification tree consists of root node which through a series of splits creates the terminal nodes. The main questions related to creating the classification tree are: How to (1) select the splits, (2) determine the terminal nodes, and (3) assign the terminal node a class? CHAID uses Chi-squared tests to select the splits and one option for CART is the Gini index of impurity as the impurity measure of node t, which is

$$i(t) = \sum_{i \neq j} p(i \mid t) p(j \mid t),$$

where $p(i|t)$ the is probability of class i in node t and $p(j|t)$ is the probability of class j in node t.[1] The Gini index of impurity is the splitting rule that will be used in this paper.

The terminal nodes are determined when CHAID can no longer find any statistically significant splits, CART can no longer find splits that lead to impurity reductions below a specified threshold, or the terminal nodes in CHAID or CART would create terminal nodes below a required minimum of observations. The assignment of a class value to the terminal nodes follows one of the imputation methods defined in Section IV, Imputation Methodologies. In this paper, we use the terminal node at the imputation class.

The following classification tree, Figure 1, is an example of a binary target variable with values 0 and 1 and two continuous on (0,1) predictor variables. The

[1] Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J. (1984). *Classification and Regression Trees*. Chapman & Hall. Page 38.

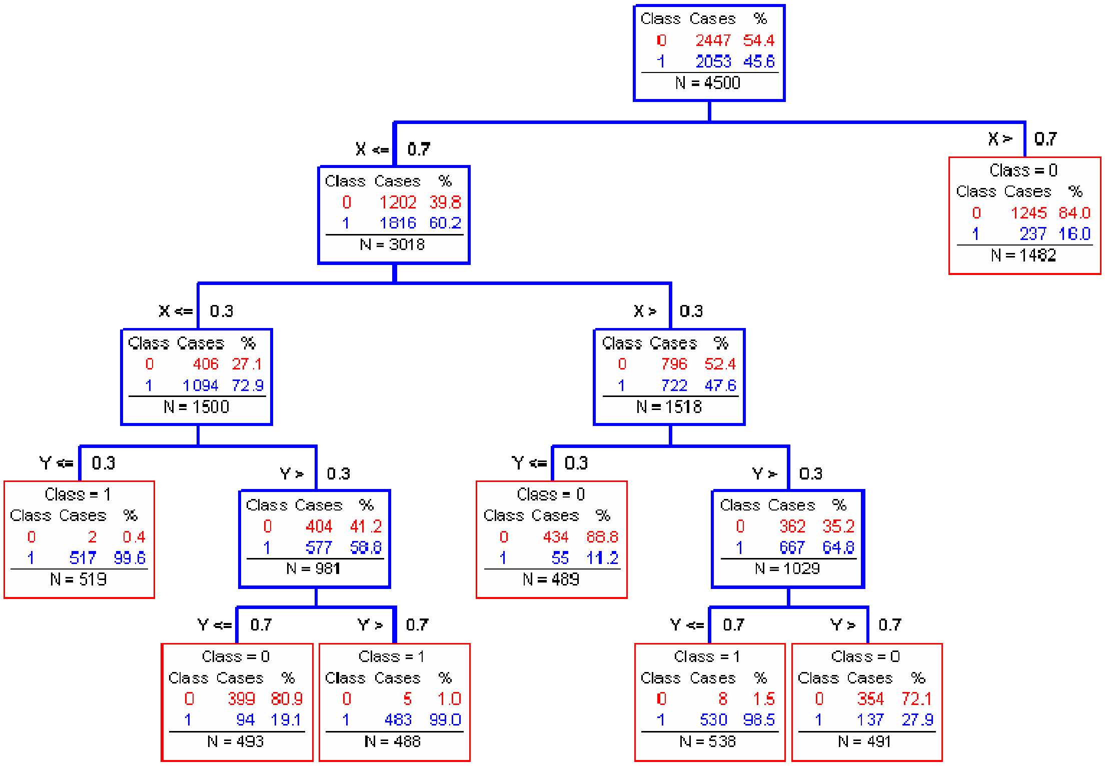

Figure 1. Example Classification Tree

root node has 4,500 observations: class 0 with 2,447 observations and class 1 with 2,053 observations. The first split is on the X variable at value 0.7. The left branch, X <= 0.7, contains 3,018 observations and the right branch, X > 0.7, contains 1,482 observations. The right branch is a terminal node with class 0 having 1,245 observations and class 1 having 237 observations. Consequently, this terminal node is assigned class value 0.

In this paper, we investigate three tree types. Two trees created by CHAID and a tree created by CART. The CHAID trees are a full tree and a tree that is restricted to a depth of three and the CART tree is a full tree. From this point forward, we will refer to the terminal nodes as imputation classes and will not use the assigned class values from CHAID or CART.

3. Imputation Methodologies

In conjunction with the imputation classes created by the three tree types, we used three imputation procedures: mode value imputation, proportional random imputation, and weighted sequential hot-deck. For mode value imputation, recipients were assigned an imputed value that was the mode value for the imputation class in which they fell. This is a simple imputation method for assigning the values. For proportional random imputation, recipients were randomly assigned an imputed value based on the value’s relative size in the empirical distribution of the donors in the imputation class. For weighted sequential hot-deck, recipients were assigned an imputed value based on the weights of the respondents, the weights of the recipients, the position of the recipients in the data file, and conditional sequential probabilities.

In order to have a more comprehensive evaluation of the imputation procedures, we modeled the item nonresponse using logistic regression and polychotomous regression where appropriate for some of the data.

4. Data

In order to evaluate how the different methodologies perform, two data sets with known characteristics were generated and two data sets of real data were used. Each of the four data sets was recreated with three levels of random missingness, 5%, 25%, and 50%, for a count of twelve data sets. For each of the twelve data sets nine combinations of tree type and imputation method were used. That is, each of the twelve data sets had all three tree types - CART, CHAID, and CHAID restricted to three levels – and all three imputation methods applied with in a tree type for a total of 108 data sets.

4.1 Generated Data

The first generated data set consists of points in two-dimensional space that have or do not have a characteristic of interest. In the data set, the points are represented by an indicator variable for the characteristic of interest, and continuous, on the interval zero to one, variables for the x-value and the y-value. The second generated data set consists of points in two-dimensional space that have one of three nominal values for a characteristic of interest. In the data set, the points are represented by a nominal variable with six values for the characteristic of interest, and continuous, on the interval zero to one, variables for the x-value and the y-value.

4.2 Real Data

The first real data set consists of a binomial dependent variable, or target variable, and numerous continuous and categorical independent variables. The second data set contains the same independent variables but has a three-level nominal variable.

5. Evaluation Methods

To evaluate which of the three tree types crossed with the three imputation methods produces the most accurate imputed vales, we used the misclassification rate and relative means square error.

Let t_i be the true value, m_i be the imputed value, and n' be the number of imputed observations. To calculate the misclassification rate, we construct an indicator variable to identify the misclassified imputed observations as follows

$$d_i = \begin{cases} 1 \;\; if \;\; t_i \neq m_i \\ 0 \;\; if \;\; t_i = m_i \end{cases}, where \; i = 1, \ldots, n'.$$

The number of misclassified observations is the sum, over all imputed observations, of the indicator variable

$$\sum_{i=1}^{n'} d_i \,.$$

Finally, the misclassification rate is the sum, over all imputed observations, of the indicator variable divided by the number of imputed observations

$$\frac{\sum_{i=1}^{n'} d_i}{n'} \,.$$

The following is a description of how we calculated the mean square error and relative mean square error.

First, we calculated an approximately unbiased estimate of the true mean (p) and the variance, σ^2 , of p . We kept a subset of the original data set that contains the observations that have a valid value for the dependent variable which we are using for the study. From this subset and assuming the missing values where missing at random, we calculated an approximately unbiased estimate of the true mean and the variance associated with this estimate.

Second, we created the imputed data sets. We deleted some portion of dependent variable from the data subset. We impute the missing values.

Third, we calculated the after imputation mean ($\hat{p}$) and after imputation variance ($\hat{\sigma}^2$). These values were calculated from the imputed values and the original values that were not deleted in the imputed data sets.

Finally, we calculated the mean square error (MSE) as the estimated bias squared plus the after imputation variance,

$$(\hat{p} - p)^2 + \hat{\sigma}^2 \,.$$

The relative mean square error (RMSE) is the mean square of the after imputation minus the variance of the approximately unbiased estimate of the true mean all of which is divided by the variance of the approximately unbiased estimate of the true mean. The formula is

$$RMSE = \frac{MSE(\hat{p}) - MSE(p)}{MSE(p)} =$$

$$\frac{\left[(\hat{p} - p)^2 + \hat{\sigma}^2\right] - \sigma^2}{\sigma^2}$$

and assuming that the p is approximately unbiased, we have

$$RMSE = \frac{MSE(\hat{p}) - MSE(p)}{MSE(p)} = \frac{(\hat{p} - p)^2 + Var(\hat{p}) - Var(p)}{Var(p)}$$

Finally, we use the combined error rate which is the product of the misclassification rate and the RMSE.

6. Results

We aggregated the information from the 108 different data sets using the actual values and the ranked values. The following table contains the information about the performance using the actual values.

Method	Aggregated Value
CART/WSHD	14.722
CHAID3/WSHD	15.962
CART/Random	16.122
CHAID/WSHD	16.194
CHAID/Random	17.776
CHAID3/Random	20.950
CART/Mode	224.878
CHAID/Mode	268.634
CHAID3/Mode	452.832

The smaller the aggregated value is the better the type of tree and imputation methodology preformed. So the CART tree methodology using the weighted sequential hot deck imputation methodology preformed the best and the mode imputation methodology preformed the worst no matter which tree methodology was used.

The following table contains the performance using the ranks.

Method	Aggregated Rank
CART/WSHD	1
CART/Random	2
CHAID3/WSHD	3
CHAID3/Random	4
CHAID/Random	5
CHAID/WSHD	6
CHAID3/Mode	7
CART/Mode	8
CHAID/Mode	9

It has generally the same patter as the table for the actual values but there are some slight difference is the middle of the rank based table when compare the actual value table.

7. Recommendation

The best method for these data sets is CART tree methodology for creating the imputation classes and then applying the weighted sequential hot deck imputation methodology within the imputation classes. The other tree methodologies combined with the random imputation methodologies have very similar values so a practitioner would not loose much predictive accuracy using any one those combinations. The worst were the mode imputations regardless of the tree methodology used.

References

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J. (1984). *Classification and Regression Trees*. Chapman & Hall.

Cox, B.G. (1980). The weighted sequential hot deck imputation procedure. *Proceedings of the American Statistical Association Section on Survey Research Methods*, 721-726.

Iannacchione, V.G. (1982). Weighted sequential hot deck imputation macros. *Proceedings of the Seventh Annual SAS User's Group International Conference*.

Kass, G.V. (1980). An exploratory technique for investigation large quantities of categorical data. *Applied Statistics*, Vol. 29, No. 2, 119-127.