Journal of Machine Learning Research 11 (2010) 131-170 Submitted 4/09; Revised 12/09; Published 1/10

An Investigation of Missing Data Methods for Classification Trees Applied to Binary Response Data

Yufeng Ding YUFENG.DING@MOODYS.COM
Moody's Investors Service
250 Greenwich Street
New York, NY 10007

Jeffrey S. Simonoff JSIMONOF@STERN.NYU.EDU
New York University, Stern School of Business
44 West 4th Street
New York, NY 10012

Editor: Charles Elkan

Abstract

There are many different methods used by classification tree algorithms when missing data occur in the predictors, but few studies have been done comparing their appropriateness and performance. This paper provides both analytic and Monte Carlo evidence regarding the effectiveness of six popular missing data methods for classification trees applied to binary response data. We show that in the context of classification trees, the relationship between the missingness and the dependent variable, as well as the existence or non-existence of missing values in the testing data, are the most helpful criteria to distinguish different missing data methods. In particular, separate class is clearly the best method to use when the testing set has missing values and the missingness is related to the response variable. A real data set related to modeling bankruptcy of a firm is then analyzed. The paper concludes with discussion of adaptation of these results to logistic regression, and other potential generalizations.

Keywords: classification tree, missing data, separate class, RPART, C4.5, CART

1. Classification Trees and the Problem of Missing Data

Classification trees are a supervised learning method appropriate for data where the response variable is categorical. The simple methodology behind classification trees is to recursively split data based upon the predictors that best distinguish the response variable classes. There are, of course, many subtleties, such as the choice of criterion function used to pick the best split variable, stopping rules, pruning rules, and so on. In this study, we mostly rely on the built-in features of the tree algorithms C4.5 and RPART to implement tree methods. Details about classification trees can be found in various references, for example, Breiman, Friedman, Olshen, and Stone (1998) and Quinlan (1993). Classification trees are computationally efficient, can handle mixed variables (continuous and discrete) easily and the rules generated by them are relatively easy to interpret and understand. Classification trees are highly flexible, and naturally uncover interaction effects among the independent variables. Classification trees are also popular because they can easily be incorporated into learning ensembles or larger learning systems as base learners.

Like most statistics or machine learning methods, "base form" classification trees are designed assuming that data are complete. That is, all of the values in the data matrix, with the rows being the observations (instances) and the columns being the variables (attributes), are observed. However, missing data (meaning that some of the values in the data matrix are not observed) is a very common problem, and for this reason classification trees have to, and do, have ways of dealing with missing data in the predictors. (In supervised learning, an observation with missing response value has no information about the underlying relationship, and must be omitted. There is, however, research in the field of semi-supervised learning methods that tries to handle the situation where the response value is missing, for example, Wang and Shen 2007.)

Although there are many different ways of dealing with missing data in classification trees, there are relatively few studies in the literature about the appropriateness and performance of these missing data methods. Moreover, most of these studies limited their coverage to the simplest missing data scenario, namely, missing completely at random (MCAR), while our study shows that the missing data generating process is one of the two crucial criteria in determining the best missing data method. The other crucial criterion is whether or not the testing set is complete. The following two subsections describe in more detail these two criteria.

1.1 Different Types of Missing Data Generating Process

Data originate according to the data generating process (DGP) under which the data matrix is "generated" according to the probabilistic relationships between the variables. We can think of the missingness itself as a random variable, realized as the matrix of the missingness indicator I_m. I_m is generated according to the missingness generating process (MGP), which governs the relationship between I_m and the variables in the data matrix. I_m has the same dimension as the original data matrix, with each entry equal to 0 if the corresponding original data value is observed and 1 if the corresponding original data value is not observed (missing). Note that an I_m value not only can be related to its corresponding original data value, but can also be related to other variables of the same observation.

Depending on the relationship between I_m and the original data, Rubin (1976) and Little and Rubin (2002) categorize the missingness into three different types. If I_m is dependent upon the missing values (the unobserved original data values), then the missingness pattern is called "not missing at random" (NMAR). Otherwise, the missingness pattern is called "missing at random" (MAR). As a special case of MAR, when the missingness is also not dependent on the observed values (that is, is independent of all data values), the missingness pattern is called "missing completely at random" (MCAR). The definition of MCAR is rather restrictive, which makes MCAR unlikely in reality. For example, in the bankruptcy data discussed later in the paper, there is evidence that after the Enron scandal in 2001, when both government and the public became more wary about financial reporting misconduct, missingness of values in financial statement data was related to the well-being of the company, and thus other values in the data. This makes intuitive sense because when scrutinized, a company is more likely to have trouble reporting their financial data if there were problems. Thus, focusing on the MCAR case is a major limitation that will be avoided in this paper. In fact, this paper shows that the categorization of MCAR, MAR and NMAR itself is not appropriate for the missing data problem in classification trees, as well as in another supervised learning context (at least with respect to prediction), although it has been shown to be helpful with likelihood-based or Bayesian analysis.

	Missingness is related to				
	Missing values	Observed Predictors	Response Variable	LR	Three-Letter
1	No	No	No	MCAR	− − −
2	No	Yes	No	MAR	−X−
3	Yes	No	No	NMAR	M − −
4	Yes	Yes	No	NMAR	M X−
5	No	No	Yes	MAR	− − Y
6	No	Yes	Yes	MAR	−X Y
7	Yes	No	Yes	NMAR	M − Y
8	Yes	Yes	Yes	NMAR	M X Y

Table 1: Eight missingness patterns investigated in this study and their correspondence to the categorization MCAR, MAR and NMAR defined by Rubin (1976) and Little and Rubin (2002) (the LR column). The column Three-Letter shows the notation that is used in this paper.

In this paper, we investigate eight different missingness patterns, depending on the relationship between the missingness and three types of variables, the observed predictors, the unobserved predictors (the missing values) and the response variable. The relationship is conditional upon other factors, for example, missingness is not dependent upon the missing values means that the missingness is conditionally independent of the missing values given the observed predictors and/or the response variable. Table 1 shows their correspondence with the MCAR/MAR/NMAR categorization as well as the three-letter notation we use in this paper. The three letters indicate if the missingness is conditionally dependent on the missing values (M), on other predictors (X) and on the response variable (Y), respectively. As will be shown, the dependence of the missingness on the response variable (the letter Y) is the one that affects the choice of best missingness data method. Later in the paper, some derived notations are also used. For example, $*X*$ means the union of $-X-$, $-XY$, $MX-$ and MXY, that is, the missingness is dependent upon the observed predictors, and it may or may not be related to the missing values and/or the response variable.

1.2 Scenarios Where the Testing Data May or May Not Be Complete

There are essentially two stages of applying classification trees, the training phase where the historical data (training set) are used to construct the tree, and the testing phase where the tree is put into use and applied to testing data. Similar to most other studies, this study deals with the scenario where missing data occur in the training set, but the testing set may or may not have missing values. One basic assumption is, of course, that the DGP (as well as MGP if the testing set also contains missing values) is the same for both the training set and the testing set.

While it would probably typically be the case that the testing data would also have missing values (generated by the same process that generated them in the training set), it should be noted that in certain circumstances a testing set without missing values could be expected. For example, consider a problem involving prediction of bankruptcy from various financial ratios. If the training set comes from a publicly available database, there could be missing values corresponding to information that was not supplied by various companies. If the goal is to use these publicly available data to try

to predict bankruptcy from ratios from one's own company, it would be expected that all of the necessary information for prediction would be available, and thus the test set would be complete.

This study shows that when the missingness is dependent upon the response variable and the test set has missing values, separate class is the best missing data method to use. In other situations, the choice is not as clear, but some insights on effective choices are provided. The rest of paper provides detailed theoretical and empirical analysis and is organized as follows. Section 2 gives a brief introduction to the previous research on this topic. This is followed by discussion of the design of this study and findings in Section 3. The generality of the results are then tested on real data sets in Section 4. A brief extension of the results to logistic regression is presented in Section 5. We conclude with discussion of these results and future work in Section 6.

2. Previous Research

There have been several studies of missing data and classification trees in the literature. Liu, White, Thompson, and Bramer (1997) gave a general description of the problem, but did not discuss solutions. Saar-Tsechansky and Provost (2007) discussed various missing data methods in classification trees and proposed a cost-sensitive approach to the missing data problem for the scenario when missing data occur only at the testing phase, which is different from the problem studied here (where missing values occur in the training phase).

Kim and Yates (2003) conducted a simulation study of seven popular missing value methods but did not find any dominant method. Feelders (1999) compared the performance of surrogate split and imputation and found the imputation methods to work better. (These methods, and the methods described below, are described more fully in the next section.) Batista and Monard (2003) compared four different missing data methods, and found that 10 nearest neighbor imputation outperformed other methods in most cases. In the context of cost sensitive classification trees, Zhang, Qin, Ling, and Sheng (2005) studied four different missing data methods based on their performances on five data sets with artificially generated random missing values. They concluded that the internal node method (the decision rules for the observations with the next split variable missing will be made at the (internal) node) is better than the other three methods examined. Fujikawa and Ho (2002) compared several imputation methods based on preliminary clustering algorithms to probabilistic split on simulations based on several real data sets and found comparable performance. A weakness of all of the above studies is that they focused only on the restrictive MCAR situation.

Other studies examined both MAR and NMAR missingness. Kalousis and Hilario (2000) used simulations from real data sets to examine the properties of seven algorithms: two rule inducers, a nearest neighbor method, two decision tree inducers, a naive Bayes inducer, and linear discriminant analysis. They found that the naive Bayes method was by far most resilient to missing data, in the sense that its properties changed the least when the missing rate was increased (note that this resilience is related to, but not the same as, its overall predictive performance). They also found that the deleterious effects of missing data are more serious if a given amount of missing values are spread over several variables, rather than concentrated in a few.

Twala (2009) used computer simulations based on real data sets to compare the properties of different missing value methods, including using complete cases, single imputation of missing values, likelihood-based multiple imputation (where missing values are imputed several times, and the results of fitting trees to the different generated data sets are combined), probabilistic split, and surrogate split. He studied MAR, MCAR, and NMAR missingness generating processes, although

dependence of missingness on the response variable was not examined. Multiple imputation was found to be most effective, with probabilistic split also performing reasonably well, although little difference was found between methods when the proportion of missing values was low. As would be expected, MCAR missingness caused the least problems for methods, while NMAR missingness caused the most, and as was also found by Kalousis and Hilario (2000), missingness spread over several predictors is more serious than if it is concentrated in only one. Twala, Jones, and Hand (2008) proposed a method closely related to creating a separate class for missing values, and found that its performance was competitive with that of likelihood-based multiple imputation.

The study described in the next section extends these previous studies in several ways. First, theoretical analyses are provided for simple situations that help explain observed empirical performance. We then extend these analyses to more complex situations and data sets (including large ones) using Monte Carlo simulations based on generated and real data sets. The importance of whether missing is dependent on the response variable, which has been ignored in previous studies on classification trees yet turns out to be of crucial importance, is a fundamental aspect of these results. The generality of the conclusions is finally tested using real data sets and application to logistic regression.

3. The Effectiveness of Missing Data Methods

The recursive nature of classification trees makes them almost impossible to analyze analytically in the general case beyond 2×2 tables (where there is only one binary predictor and a binary response variable). On the other hand, trees built on 2×2 tables, which can be thought of as "stumps" with a binary split, can be considered as degenerate classification trees, with a classification tree being built (recursively) as a hierarchy of these degenerate trees. Therefore, analyzing 2×2 tables can result in important insights for more general cases. We then build on the 2×2 analyses using Monte Carlo simulation, where factors that might have impact on performance are incrementally added, in order to see the effect of each factor. The factors include variation in both the data generating process (DGP) and the missing data generating process (MGP), the number and type of predictors in the data, the number of predictors that contain missing values, and the number of observations with missing data.

This study examines six different missing data methods: probabilistic split, complete case method, grand mode/mean imputation, separate class, surrogate split, and complete variable method. Probabilistic split is the default method of C4.5 (Quinlan, 1993). In the training phase, observations with values observed on the split variable are split first. The ones with missing values are then put into each of the child nodes with a weight given as the proportion of non-missing instances in the child. In the testing phase, an observation with a missing value on a split variable will be associated with all of the children using probabilities, which are the weights recorded in the training phase. The complete case method deletes all observations that contain missing values in any of the predictors in the training phase. If the testing set also contains missing values, the complete case method is not applicable and thus some other method has to be used. In the simulations, we use C4.5 to realize the complete case method. In the training phase, we manually delete all of the observations with missing values and then run C4.5 on the pre-processed remaining complete data. In the testing phase, the default missing data method, probabilistic split, is used. Grand mode imputation imputes the missing value with the grand mode of that variable if it is categorical. Grand mean is used if the variable is continuous. The separate class method treats the missing values as a new class

(category) of the predictor. This is trivial to apply when the original variable is categorical, where we can create a new category called "missing". To apply the separate class method to a numerical variable, we give all of the missing values a single extremely large value that is obviously outside of the original data range. This creates the needed separation between the nonmissing values and the missing values, implying that any split that involves the variable with missing values will put all of the missing observations into the same branch of the tree. Surrogate split is the default method of CART (realized using RPART in this study; Breiman et al. 1998 and Therneau and Atkinson 1997). It finds and uses a surrogate variable (or several surrogates in order) within a node if the variable for the next split contains missing values. In the testing phase, if a split variable contains missing values, the surrogate variables in the training phase are used instead. The complete variable method simply deletes all variables that contain missing values.

Before we start presenting results, we define a performance measure that is appropriate for measuring the impact of missing data. Accuracy, calculated as the percentage of correctly classified observations, is often used to measure the performance of classification trees. Since it can be affected by both the data structure (some data are intrinsically easier to classify than others) and by the missing data, this is not necessarily a good summary of the impact of missing data. In this study, we define a measure called *relative accuracy* (*RelAcc*), calculated as

$$RelAcc = \frac{\text{Accuracy with missing data}}{\text{Accuracy with original full data}}.$$

This can be thought of as a standardized accuracy, as *RelAcc* measures the accuracy achievable with missing values relative to that achievable with the original full data.

3.1 Analytical Results

In the following consistency theorems, the data are assumed to reflect the DGP exactly, and therefore the training set and the testing set are exactly the same. Several of the theorems are for 2×2 tables, and in those cases stopping and pruning rules are not relevant, since the only question is whether or not the one possible split is made. The proofs are thus dependent on the underlying parameters of the DGP and MGP, rather than on data randomly generated from them. It is important to recognize that these results are only designed to be illustrative of the results found in the much more realistic simulation analyses to follow. Proofs of all of the results are given in the appendix.

Before presenting the theorems, we define some terms to avoid possible confusion. First, a partition of the data refers to the grouping of the observations defined by the classification tree's splitting rules. Note that it is possible for two different trees on the same data set to define the same partition. For example, suppose that there are only two binary explanatory variables, X_1 and X_2, and one tree splits on X_1 then X_2 while another tree splits on X_2 then X_1. In this case, these two trees have different structures, but they can lead to the same partition of the data. Secondly, the set of rules defined by a classification tree consists of the rules defined by the tree leaves on each of the groups (the partition) of the data.

3.1.1 WHEN THE TEST SET IS FULLY OBSERVED WITH NO MISSING VALUES

We start with Theorems 1 to 3 that apply to the complete case method. Theorems 4 and 5 apply to probabilistic split and mode imputation, respectively. Proofs of the theorems can be found in the appendix.

Theorem 1 *Complete Case Method: If the MGP is conditionally independent of Y given X, then the tree built on the data containing missing values using the complete case method gives the same set of rules as the tree built on the original full data set.*

Theorem 2 *Complete Case Method: If the partition of the data defined by the tree built on the incomplete data is not changed from the one defined by the tree built on the original full data, the loss in accuracy when the testing set is complete is bounded above by P_M, where P_M is the missing rate, defined as the percentage of observations that contain missing values.*

Theorem 3 *Complete Case Method: If the partition of the data defined by the tree built on the incomplete data is not changed from the one defined by the tree built on the original full data, the relative accuracy when the testing set is complete is bounded below by*

$$RelAcc_{min} = \frac{1-P_M}{1+P_M},$$

where P_M is the missing rate. Notice that the tree structure itself could change as long as it gives the same final partition of the data.

There are similar results in regression analyses as in Theorem 1. In regression analyses, when the missingness is independent of the response variable, by using only the complete observations, the parameter estimators are all unbiased (Allison, 2001). This implies that in theory, when the missingness is independent of the response variable, using complete cases only is not a bad approach on average. However, in practice, as will be seen later, deleting observations with missing values can cause severe loss in information, and thus has generally poor performance.

Theorem 4 *Probabilistic Split: In a 2×2 data table, if the MGP is independent of either Y or X, given the other variable, then the following results hold for probabilistic split.*

1. *If X is not informative in terms of classification, that is, the majority classes of Y for different X values are the same, then probabilistic split will give the same rule as the one that would be obtained from the original full data;*
2. *If probabilistic split shows that X is informative in terms of classification, that is, the majority classes of Y for different X values are different, then it finds the same rule as the one that would be obtained from the original full data;*
3. *The absolute accuracy when the testing set is complete is bounded below by 0.5. Since the original full data accuracy is at most 1, the relative accuracy is also bounded below by 0.5.*

Theorem 5 *Mode Imputation: If the MGP is independent of Y, given X, then the same results hold for mode imputation as for probabilistic split under the conditions of Theorem 4.*

Theorems 1, 2 and 3 (for the complete case method) are true for general data sets. Theorems 4 and 5 are for 2×2 tables only but they imply that probabilistic split and mode imputation have advantages over the complete case method, which can have very poor performance (as will be shown in Figure 1).

Moreover, with 2×2 tables, the complete variable method will always have a higher than 0.5 accuracy since by ignoring the only predictor, we will always classify all of the data to the overall majority class and achieve at least 0.5 accuracy, and thus at least 0.5 relative accuracy. Together with Theorems 4 and 5, as well as the evidence to be shown in Figure 1, this is an indication that classification trees tend not to be hurt much by missing values, since trees built on 2×2 tables can be considered as degenerate classification trees and more complex trees are composites of these degenerate trees. The performance of a classification tree is the average (weighted by the number of observations at each leaf) over the degenerate trees at the leaf level, and, as will be seen later in the simulations, can often be quite good.

Surrogate split is not applicable to 2×2 tables because there are no other predictors. For 2×2 table problems with a complete testing set, separate class is essentially the same as the complete case method, because as long as the data are split according to the predictor (and it is very likely that this will be so), the separate class method builds separate rules for the observations with missing values; when the testing set is complete, the rules that are used in the testing phase are exactly the ones built on the complete observations. When there is more than one predictor, however, the creation of the "separate class" will save the observations with missing values from being deleted and affect the tree building process. It will very likely lead to a change in the tree structure. This, as will be seen, tends to have a favorable impact on the performance accuracy.

Figure 1 illustrates the lower bound calculated in Theorem 3. The illustration is achieved by Monte Carlo simulation of 2×2 tables. A 2×2 table with missing values has only eight cells, that is, eight different value combinations of the binary variables X, Y and M, where M is the missingness indicator such that $M = 0$ if X is observed and $M = 1$ if X is missing. There is one constraint, that the sum of the eight cell probabilities must equal one. Therefore, this table is determined by seven parameters. In the simulation, for each 2×2 table, the following seven parameters (probabilities) are randomly and independently generated from a uniform distribution between $(0,1)$: (1)$P(X = 1)$, (2)$P(Y = 1|X = 0)$, (3)$P(Y = 1|X = 1)$, (4)$P(M = 1|X = 0,Y = 0)$, (5)$P(M = 1|X = 0,Y = 1)$, (6)$P(M = 1|X = 1,Y = 0)$ and (7)$P(M = 1|X = 1,Y = 1)$. Here we assume the data tables reflect the true underlying DGP and MGP without random variation, and thus the expected performance of the classification trees can be derived using the parameters. In this simulation, sets of the seven parameters are generated (but no data sets are generated using these parameters) repeatedly, and the relative accuracy of each missing data method on each parameter set is determined. One million sets of parameters are generated for each missingness pattern.

In Figure 1, the plot on the left is a scatter plot of relative accuracy versus missing rate for each Monte Carlo replication for the complete case method when the MGP depends on the response variable. The lower bound is clearly shown. We can see that when the missing rate is high, the lower bound can reduce to almost zero (implying that not only relative accuracy, but accuracy itself, can approach zero). This perhaps somewhat counterintuitive result can occur in the following way. Imagine the extreme case where almost all cases are positive and (virtually) all of the positive cases have missing predictor value at the training phase; in this situation the resultant rule will be to classify everything as negative. When this rule is applied to a complete testing set with almost all positive cases, the accuracy will be almost zero. The graph on the right is the quantile version of the scatter plot on the left. The lines shown in the quantile plot are the theoretical lower bound, the 10th, 20th, 30th, 40th and 50th percentile lines from the lowest to the highest. Higher percentile lines are the same as the 50th percentile (median) line, which is already the horizontal line at $RelAcc = 1$. The percentile lines are constructed by connecting the corresponding percentiles in a moving window

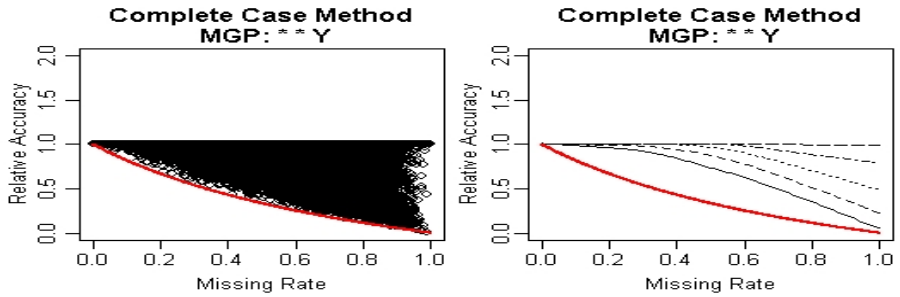

Figure 1: Scatter plot and the corresponding quantile plot of the complete testing set *RelAcc* vs. missing rate of the complete case method when the MGP is dependent on the response variable. Recall that "$**$Y" means the MGP is conditionally dependent on the response variable but no restriction on the relationship between the MGP and other variables, missing or observed, is assumed. Each point in the scatter plot represents the result on one of the simulated data tables.

of data from the left to the right. Due to space limitations, we do not show quantile plots of other missing data methods and/or under different scenarios, but in all of the other plots, the quantile lines are all higher (that is, the quantile plot in Figure 1 shows the worst case scenario). The plots show that the missing data problem, when the missing rate is not too high, may not be as serious as we might have thought. For example, when 40% of the observations contain missing data, 80% of the time the expected relative accuracy is higher than 90%, and 90% of the time the expected relative accuracy is higher than 80%.

3.1.2 WHEN THE TEST SET HAS MISSING VALUES

Theorem 6 *Separate Class: In 2×2 data tables, if missing values occur in both the training set and the testing set, then the separate class method achieves the best possible performance.*

In the Monte Carlo simulation of the 2×2 tables, the head-to-head comparison between the separate class method and other missing data methods confirmed the uniform dominance of the separate class when the test set also contains missing values, regardless whether the MGP is dependent on the response variable or not. However, as shown in Figure 2, when the MGP is independent of the response variable, separate class never performances better than the performance on the original full data, indicated by relative accuracies less than one. This means that separate class is not gaining from the missingness. On the other hand, when the MGP is dependent on the response variable, a fairly large percentage of the time the relative accuracy of the separate class method is larger than one (the quantiles shown are from the 10th to the 90th percentile with increment 10 percent). This means that trees based on the separate class method can improve on predictive performance compared to the situation where there are no missing data. Our simulations show that other methods can also gain from the missingness when the MGP is dependent on the response variable, but not as frequently as the separate class method and the gains are in general not as large. We follow up on this behavior in more detail in the next section, but the simple explanation is that since missingness depends on the response variable, the tree algorithm can use the presence of missing data in an observation to improve prediction of the response for that observation. Duda, Hart, and Stork (2001) and Hand (1997) briefly mentioned this possibility in the classification context, but did not give any

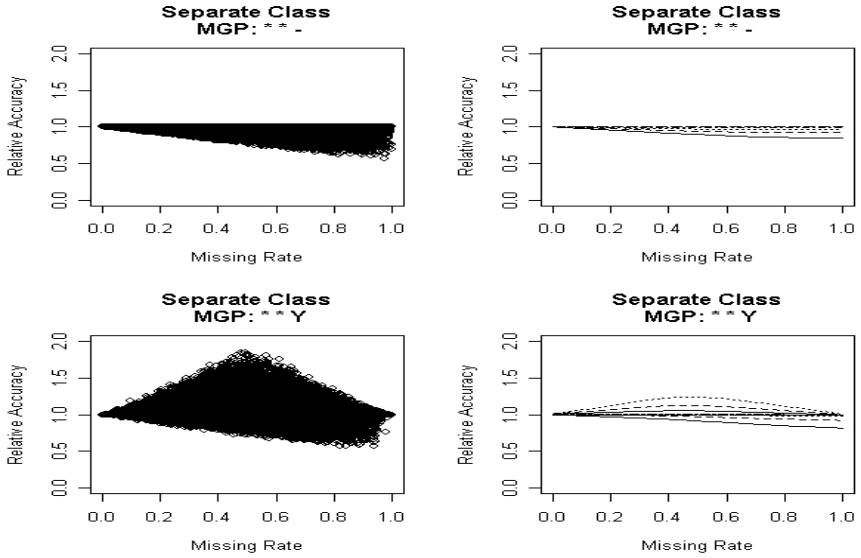

Figure 2: Scatter plot of the separate class method with incomplete testing set. Each point in the scatter plot represents the result on one of the simulated data tables.

supporting evidence. Theorem 6 makes a fairly strong statement in the simple situation, and it will be seen to be strongly indicative of the results in more general cases.

3.2 Monte Carlo Simulations of General Data Sets

In this section extensions of the simulations in the last section are summarized.

3.2.1 AN OVERVIEW OF THE SIMULATION

The following simulations are carried out.

1. 2×2 tables, missing values occur in the only predictor.
2. Up to seven binary predictors, missing values occur in only one predictor.
3. Eight binary predictors, missing values occur in two of them.
4. Twelve binary predictors, missing values occur in six of them.
5. Eight continuous predictors, missing values occur in two of them.
6. Twelve continuous predictors, missing values occur in six of them.

Two different scenarios of each of the last four simulations listed above were performed. In the first scenario, the six complete predictors are all independent of the missing ones, while in the second scenario three of the six complete predictors are related to the missing ones. Therefore, ten simulations were done in total.

In each of the simulations, 5000 sets of DGPs are simulated in order to cover a wide range of different-structured data sets so that a generalizable inference from the simulation is possible. For

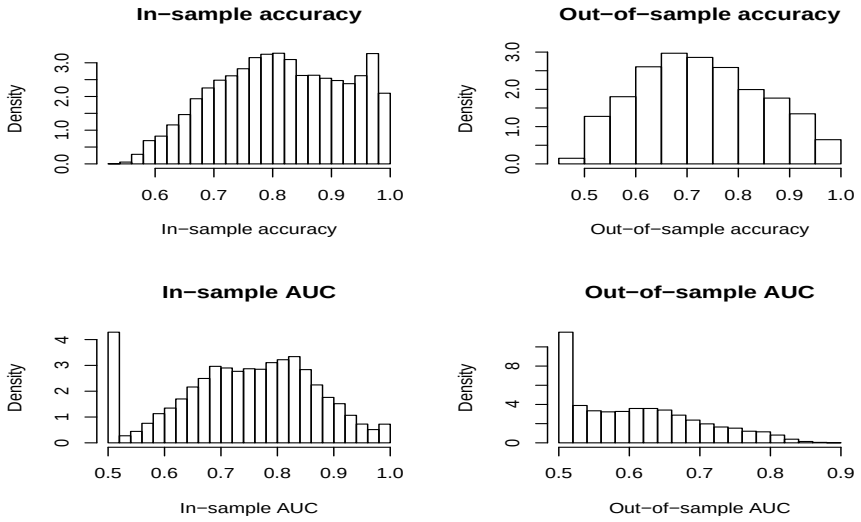

Figure 3: A summary of the tree performance on the simulated original full data.

each DGP, eight different MGPs are simulated to cover different types of missingness patterns. For each data set, the variables are generated sequentially in the order of the predictors, the response and the missingness. The probabilities associated with the binary response variable and the binary missingness variable are generated using conditional logit functions. The predictors may or may not be correlated with each other. Details about the simulations implementation can be found in Ding and Simonoff (2008). For each set of DGP/MGP, several different sample sizes are simulated to see any possible learning curve effect, since it was shown by Perlich, Provost, and Simonoff (2003) that sample size is an important factor in the effectiveness of classification trees. Figure 3 shows the distribution of the tree performance on the simulated original full data, as measured by accuracy and area under the ROC curve (AUC). As we can see, there is broad coverage of the entire range of strength of the underlying relationship. Also, as expected, the out-of-sample performance (on the test set) is generally worse than the in-sample performance (on the training set). When the in-sample AUC is close to 0.5, a tree is likely to not split and as a result, any missing data method will not actually be applied, resulting in equivalent performance over all of them. To make the comparisons more meaningful, we exclude the cases where the in-sample AUC is below 0.7. Lower thresholds for exclusion (0.55 and 0.6) yield very similar results.

Of the six missing data methods covered by this study, five of them, namely, complete case method, probabilistic split, separate class, imputation and complete variable method, are realized using C4.5. These methods are always comparable. However, surrogate split is carried out using RPART, which makes it less comparable to the other methods because of differences between RPART and C4.5 other than the missing data methods. To remedy this problem, we tuned the RPART parameters (primarily the parameter "cp") so that it gives balanced results compared to C4.5 when applied to the original full data (i.e., each has a similar probability of outperforming the other), and special attention is given when comparing RPART with other methods. The out-of-sample performances of each pair of missing data methods were compared based on both *t*-tests and nonparametric tests; each difference discussed in the following sections was strongly statistically significant.

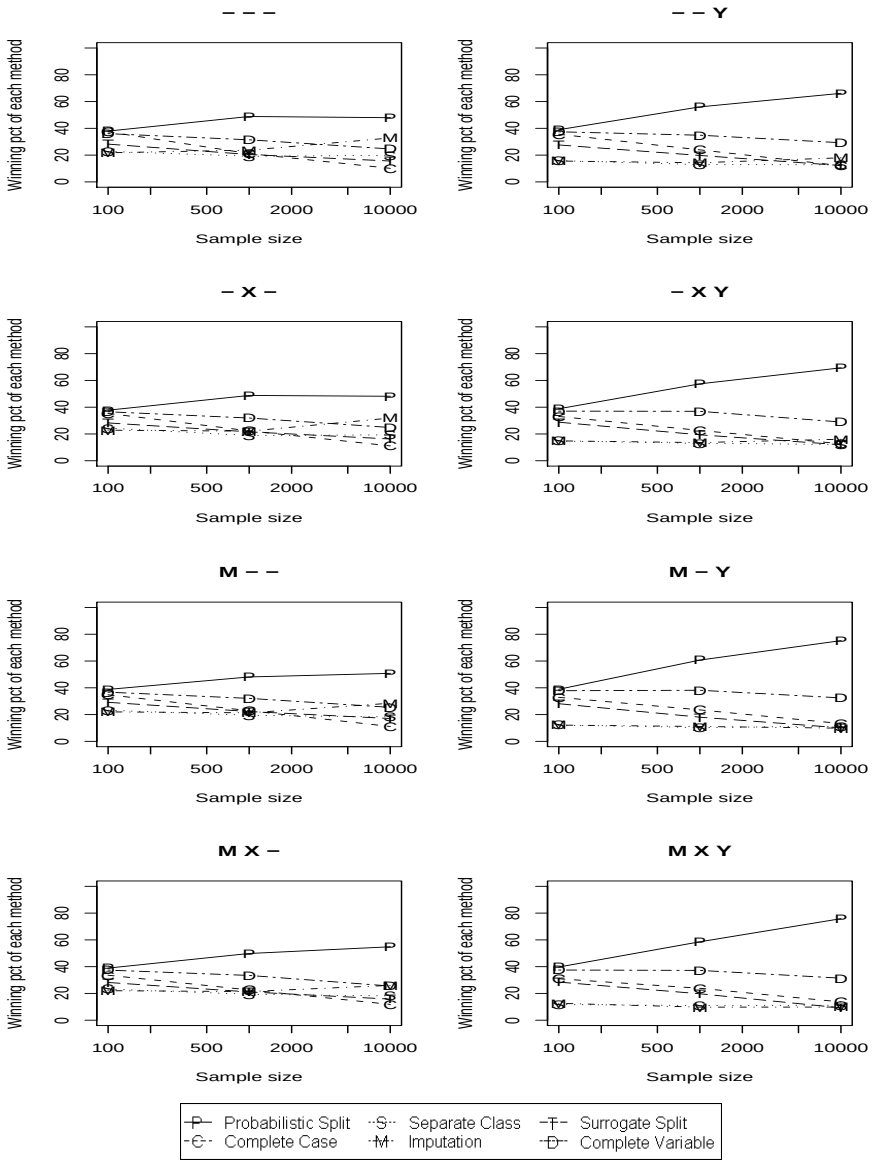

Figure 4: A summary of the order of six missing data methods when tested on a new complete testing set. The Y axis is the percentage of times each method is the best (including being tied with other methods; therefore the percentages do not sum up to one).

3.2.2 THE TWO FACTORS THAT DETERMINE THE PERFORMANCE OF DIFFERENT MISSING DATA METHODS

The simulations make clear that the dependence relationship between the missingness and the response variable is the most informative factor in differentiating different missing data methods, and thus is most helpful in determining the appropriateness of the methods. This can be clearly seen in Figures 4 and 5 (these figures refer to the case with twelve continuous predictors, six of which are subject to missing values, but results for other situations were broadly similar). The left column in

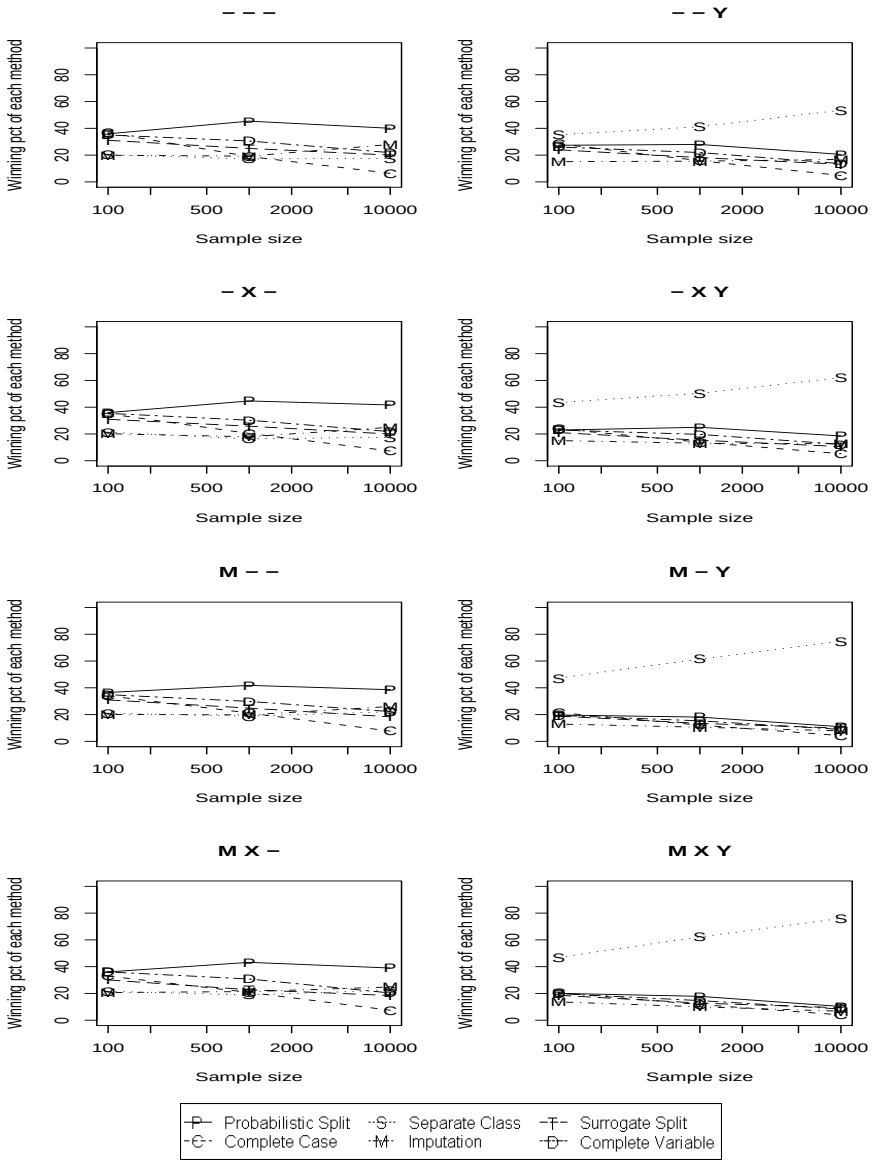

Figure 5: A summary of the order of six missing data methods when tested on a new incomplete testing set. The Y axis is the percentage of times each method is the best (including being tied with other methods).

the pictures shows the results when the missingness is independent of the response variable and the right column shows the results when the missingness is dependent on the response variable. We can see that there are clear differences between the two columns, but within each column there is essentially no difference. This also says the categorization of MCAR/MAR/NMAR (which is based upon the dependence relationship between the missingness and missing values, and does not distinguish the dependence of the missingness on other Xs and on Y) is not helpful in this context.

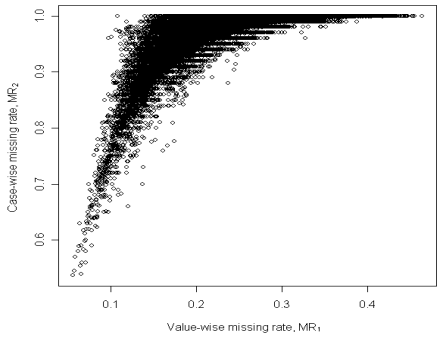

Figure 6: Plot of the case-wise missing rate MR_2 versus the value-wise missing rate MR_1 in the simulations using the 36 real data sets.

Comparison of the right columns of Figures 4 and 5 shows that whether or not there are missing values in the testing set is the second important criterion in differentiating between the methods. The separate class method is strongly dominant when the testing set contains missing values and the missingness is related to the response variable. The reason for this is that when missing data exist in both the training phase and the testing phase, they become part of the data and the MGP becomes an essential part of the DGP. This, of course, requires the assumption that the MGP (as well as the DGP) is the same in both the training phase and the testing phase. Under this scenario, if the missingness is related to the response variable, then there is information about the response variable in the missingness, which should be helpful when making predictions. Separate class, by taking the missingness directly as an "observed" variable, uses the information in the missingness about the response variable most effectively and thus is the best method to use. As a matter of fact, as can be seen in the bottom rows of Figures 7 and 8 (which give average relative accuracies separated by missing rate), the average relative accuracy of separate class under this situation is larger than one, indicating, on average, a better performance than with the original full data.

On the other hand, when the missing data only occur in the training phase and the testing set does not have missing values, or when the missingness is not related to and carries no information about the response variable, the existence of missing values is a nuisance. Its only effect is to obscure the underlying DGP and thus would most likely reduce a tree's performance. In this case, simulations show probabilistic split to be the dominantly best method. However, we don't see this dominance later in results based on real data sets. More discussion of this point will follow in Section 4.

3.2.3 MISSING RATE EFFECT

There are two ways of defining the missing rate: the percentage of predictor values that are missing from the data set (the value-wise missing rate, termed here MR_1), and the percentage of observations that contain missing values (the case-wise missing rate, termed here MR_2). If there is only one predictor, as is the case with 2×2 tables, then the two definitions are the same. We have seen earlier in the theoretical analyses that the missing rate has a clear impact on the performance of the missing data methods. In the simulations, there is also evidence of a relationship between relative performance and missing rate, whichever definition is used to define the missing rate.

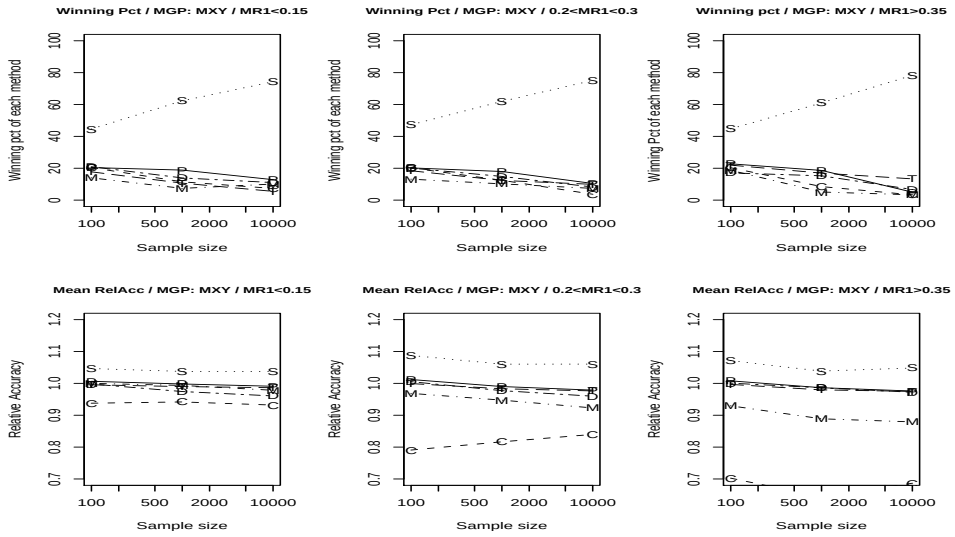

Figure 7: A comparison of the low, median and high missing rate situations. The top row shows the comparison in terms of winning percentage and the bottom row shows the comparison of the absolute performance of each missing data method.

Figure 6 shows the relationship between MR_1 and MR_2 in the simulations with 12 continuous predictors and 6 of them with missing values. Notice that in this setting, MR_1 is naturally between 0 and 0.5 (since half of the predictors can have missing values). MR_2 values are considerably larger than MR_1 values, as would be expected.

The simulations clearly show that the relative performance of different missing data methods is very consistent regardless of the missing rate (see the top row of Figure 7). However, the bottom row of Figure 7 shows that the absolute performance of the complete case method and the mean imputation method deteriorate as the missing rate gets higher. It also shows that separate class method performs best when the missing rate is neither too high or too low, although this effect is relatively small. Interestingly, the relative accuracy of the other missing data methods is very close to one regardless of the missing rate, indicating that they can almost achieve the same accuracy as if the data are complete without missing values.

A final effect connected to missing rate relates to results in earlier papers (Kalousis and Hilario, 2000; Twala, 2009) that suggested that missingness over several predictors is more problematic than missingness concentrated in a few predictors. This pattern was not evident here (e.g., in comparing the results for 8 predictors with 2 having missing values to those for 12 predictors with 6 having missing values), but it should be noted that the comparisons here are based on relative performance between methods, not absolute performance. That is, even if absolute performance deteriorates in the presence of missingness over multiple predictors, this is less important to the data analyst than is relative performance between methods (since a method must be chosen), and with respect to the latter criterion the observed patterns are reasonably stable.

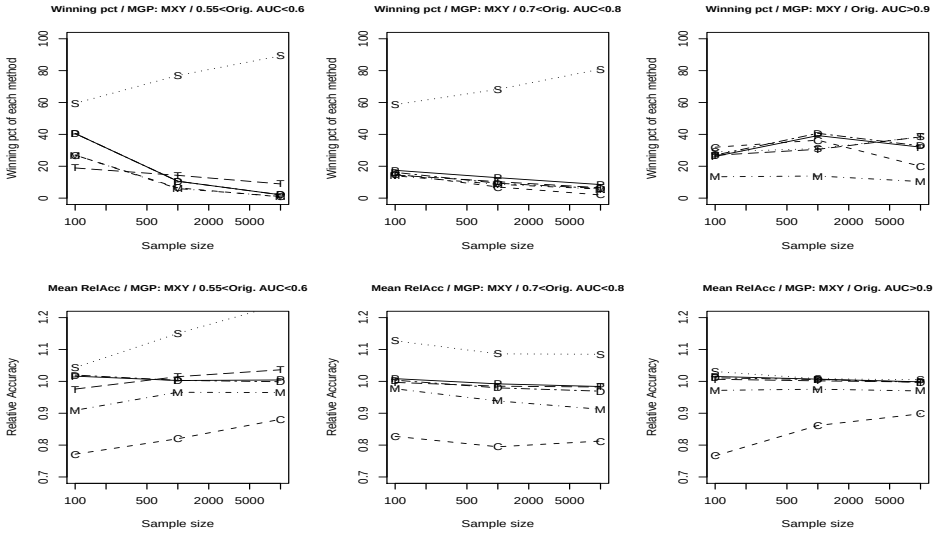

Figure 8: A comparison of the low, median and high original full data AUC situations. The top row shows the comparison in terms of winning percentage and the bottom row shows the comparison of the absolute performance of each missing data method.

3.2.4 THE IMPACT OF THE ORIGINAL FULL DATA AUC

Figure 8 shows that the original full data AUC primarily has an impact on the performance of separate class method. When the original full data AUC is higher, the loss in information due to missing values is less likely to be compensated by the information in the missingness, and thus separate class method deteriorates in performance (see the bottom row of Figure 8). When the original AUC is very high, although separate class still does a little better on average, it loses the dominance over the other methods.

Another observation is that the missing data methods other than separate class have fairly stable relative accuracy, with complete case and mean imputation consistently being the poorest performers (see the graphs in the bottom rows of both Figure 7 and Figure 8). This is true regardless of the AUC or the missing rate, even when the missingness does not depend on the response variable and there are no missing data in the testing set where, in theory, the complete case method can eventually recover the DGP.

4. Performance On Real Data Sets

In this section, we show that most of the previously described results hold when using real data sets. Moreover, we propose a method of determining the best missing data method to use when analyzing a real data set. Unlike in the previous sections, in these simulations based on real data, default settings of C4.5 are used and RPART is tuned (primarily using its parameter "cp") to get similar performance on the original full data as C4.5. Therefore, in particular, the effect of pruning is present. In Section 4.1, we show the results on 36 data sets that were originally complete. In Section 4.2, we propose a way to determine the best missing data method to use when facing real

Missingness is related to				
Missing values	Observed Predictors	Response Variable	LR	Three-Letter
No	No	No	MCAR	– – –
No	No	Yes	MAR	– – Y
Yes	No	No	NMAR	M – –

Table 2: Three missingness patterns used in simulations based on real data sets. The LR column shows the categorization according to Rubin (1976) and Little and Rubin (2002). The Three-Letter column shows the categorization used in this paper.

data sets that contain missing values (since in that case the true missingness generating process is not known by the data analyst).

4.1 Results on Real Data Sets with Simulated Missing Values

The same 36 data sets as in Perlich, Provost, and Simonoff (2003) are used here (except for Covertype and Patent, which are too big for RPART to handle; in those cases a random subset of 100,000 observations for each of them was used as the "true" underlying data set). They are either complete or were made complete by Perlich et al. (2003). Missing values with different missingness patterns were generated for the purpose of this study. According to the earlier results, the only important factor in the missingness generating process is the relationship between the missingness and the response variable. Therefore, two missingness patterns are included. In one of them, missingness is independent of all of the variables (including the response variable). In the other one, missingness is related to the response variable, but independent of all of the predictors. These two missingness patterns can be categorized as missing completely at random (MCAR) and missing at random (MAR), respectively. To account for this categorization of MGPs, the third type of missingness, not missing at random (NMAR), is also included. In the NMAR case, missingness is made dependent upon the missing values but not on the response variable (see Table 2). To maximize the possible effect of missing values, the first split variable of the original full data is chosen as the variable that contains missing values. It can be either numeric or categorical (binary or multi-categorical). Ten new data sets with missing values are generated for each combination of data set, training set size, and missingness pattern combination, with the missing rate chosen randomly for each. The performance of the missing data methods is measured out-of-sample, on a hold out test sample.

The same six missing data methods, namely, the complete case method, the complete variable method, probabilistic split, grand mode/mean imputation, surrogate split and the separate class method are applied. All of them are realized using C4.5 except for surrogate split, which is realized using RPART. C4.5 is run with its default settings. To make surrogate split comparable to the other missing data methods, the RPART parameters are tuned for each data set and each sample size so that RPART and C4.5 have comparable in sample performance on the original full data (by comparable performance we mean the average in sample original full data accuracies are similar to each other).

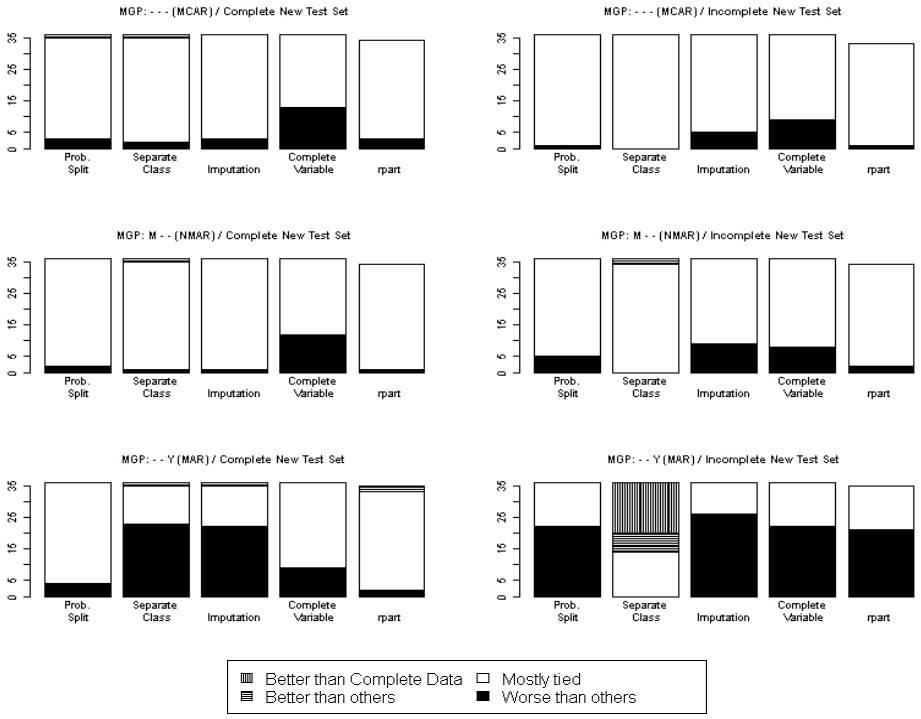

Figure 9: A tally of the relative out-of-sample performance measured in accuracy of all of the missing data methods on the 36 data sets.

4.1.1 THE TWO FACTORS AND THE BEST MISSING DATA METHOD

Consistent with the earlier results, the two factors that differentiate the performance of different missing data methods are whether the testing set is complete and whether the missingness is dependent upon the response variable. Figure 9 summarizes the relative out-of-sample performance in terms of accuracy of all of the missing data methods under different situations. In the graph, each bar represents one missing data method. Since the complete case method is consistently the worst method, it is omitted in the comparisons. Within each bar, the blank block shows the frequency that the missing data method has comparable performance with others. The shadowed block on the bottom shows the frequency that the missing data method has worse performance than others. The line-shadowed blocks on the top show the frequency that the missing data method has better performance than others, with the vertically line-shadowed block further indicating that the missing data method has better performance than with the original full data.

As was seen in the previous section, when missingness is related to the response variable and the test set contains missing values (the graph at the bottom right corner of Figure 9), the separate class method is dominant and in almost half of the cases, its performance is even better than the full data performance. Interestingly, the middle plot on the right shows that the separate class method still has an edge over the others (sometimes even over the original full data) when the test set contains missing values and the missingness is dependent upon the predictor but conditionally independent of the response variable. This is probably due to the indirect relationship between the missingness

and the response variable because both the missingness and the response variable are related to the predictor.

However, the dominance of probabilistic split is not observed in these real data sets. One possible reason could be the effect of pruning, which is used in these real data sets. The other two methods realized using C4.5 (imputation and separate class) both work with "filled-in" data sets, while probabilistic split takes the missing values as-is. Given this, we speculate that the branches with missing values are more likely to be pruned under probabilistic split, which causes it to lose predictive power. Another possible reason could be the competition from surrogate split, which is realized using RPART. Although we tried to tune RPART for each data set and each sample size, RPART and C4.5 are still two different algorithms. Different features of RPART and C4.5, other than the missing data methods, may cause RPART to outperform C4.5. Complete variable method performs a bit worse than the others, presumably because in these simulations the initial split variable on the full data was used as the variable with missing values.

In addition to accuracy, AUC was also tested as an alternative performance measure. We also examined the use of bagging (bootstrap aggregating) to reduce the variability of classification trees (discussion of bagging can be found in many sources, for example, Hastie et al. 2001). The learning curve effect (that is, the relationship between effectiveness and sample size) is also examined. We see patterns consistent with those in the simulated data sets. That is, the relative performance of the missing data methods is fairly consistent across different sample sizes.

4.1.2 THE EFFECT OF MISSING RATE

Figure 10 shows the distribution of the generated missing rates in these simulations. Recall that missing values occur in one variable, so this missing rate is the percentage of observations that have missing values, that is, MR_2 as defined earlier. Figure 11 shows a comparison between the case when the missing rate is low ($MR_2 < 0.2$) and the case when the missing rate is high ($MR_2 > 0.8$). For brevity, only the result when the MGP is dependent on the response variable is shown; differences between the low and high missing rate situations for other MGP's are similar. Since the missing rate is chosen at random, some of the original data sets do not have any generated data sets with simulated missing values with low missing rate, while for others we do not have any with high missing rate, which accounts for the "no data" category in the figures. Also, when the missing rate is high, the complete case method is obviously much worse than other missing data methods, and is therefore omitted from the comparison in that situation.

By comparing the graphs in Figures 11 with the corresponding ones in Figure 9, we can see some of the effects of missing rate. First, when the missing rate is lower than 0.2, the complete case method has comparable performance to other methods other than the complete variable method. This is unsurprising, as in this situation the complete case method does not lose much information from omitted observations. Secondly, the complete variable method has the worst performance when the missing rate is low, presumably (as noted earlier) because the complete variable method omits the most important explanatory variable in these simulations.

Moreover, in both the low and high missing rate cases, when the missingness depends on the response and the testing set is incomplete, the dominance of the separate class is not as strong as it is in Figure 9. This indicates that separate class works best when the missing rate is moderate. If the missing rate is too low, there might not be enough observations in the category of "missing" for the separate class method to be as effective. On the other hand, if the missing rate is very high, the

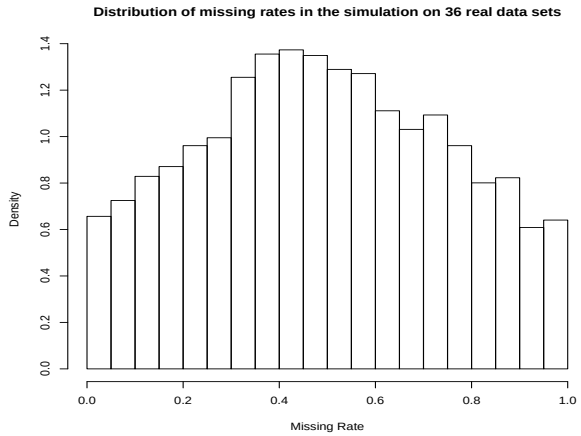

Figure 10: The distribution of missing rate in the simulation on 36 real data sets.

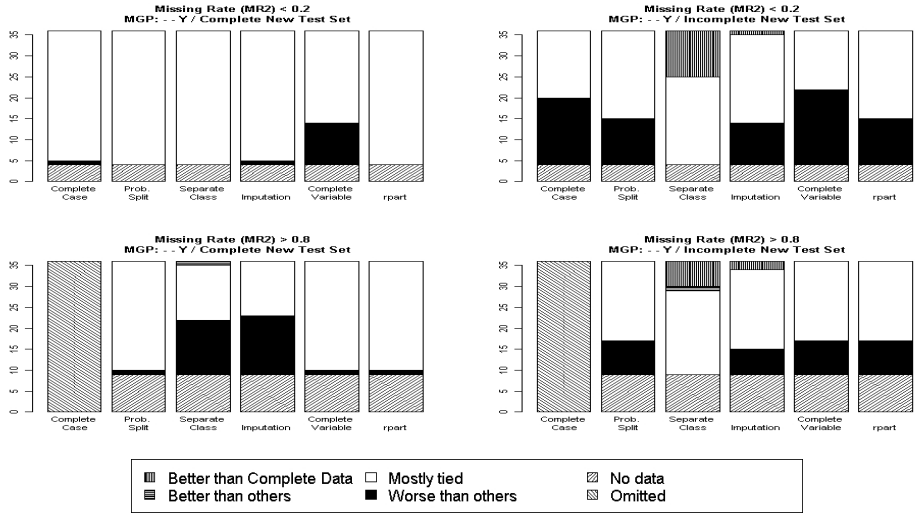

Figure 11: A comparison of the relative out-of-sample performance with low and high missing rates. Shown here, as an example, is the relative performance when the missingness is dependent upon the response variable. The left column is for the cases where the test set is fully observed and the right column for the cases where the test set has missing values. Top row shows the cases with low missing rate ($MR_2 < 0.2$) and bottom row shows the cases with high missing rate ($MR_2 > 0.8$)

.

information gained by separate class may not be enough to compensate for the lost information in the missing values, making all of the methods more comparable. This observation is consistent with Figure 2, where it is very clear that separate class gains the most when the missing rate is around 50%, as well as Figure 7, where the bottom row shows that separate class has better performance when the missing rate is neither too high or too low.

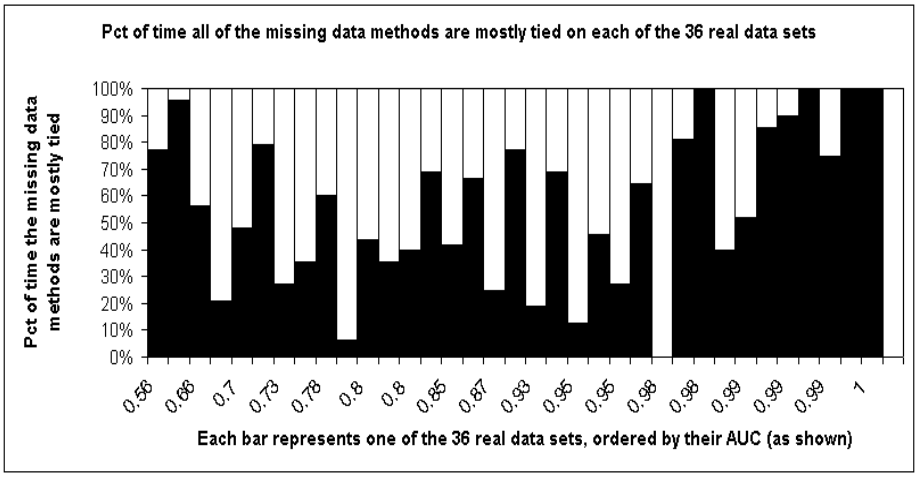

Figure 12: A tally of the missing data methods performance differentiation by data separability (measured by AUC).

4.1.3 IMPACT OF THE DATA SEPARABILITY, MEASURED BY ORIGINAL FULL DATA AUC

The experiment with these 36 data sets also shows that data separability (measured by AUC) is informative about the performance differentiation between different missing data methods (see Figure 12). In the graph, each vertical bar represents one of the 36 data sets, which are ordered from left to right according to their maximum full data AUC (as calculated by Perlich et al. 2003) from smallest to the greatest. The X-axis label shows the AUCs of the data sets. The height of each black bar shows the percentage of time when all of the missing data methods have mostly tied performance on the data set. The percentage is calculated as follows. There are three simulated missingness patterns (MCAR, NMAR and missingness depending on Y), four different testing sets (complete training set, complete new test set, incomplete training set and incomplete new test set) and four performance measures (accuracy, AUC and their bagged versions). This yields 48 measurement blocks for each data set. The performances of all of the missing data methods are compared within each block. If within a block, all of the missing data methods have very similar performance, the block is marked as mostly tied. Otherwise, the block is marked as having at least one method performing differently. The percentage is the proportion of the 48 blocks that are marked as mostly tied.

Figure 12 shows that when data separability is very high, as indicated by an AUC very close to 1 (the right end of the graph), the performances of different missing data methods are more likely to be tied. This is presumably due to the fact that strong signals in the data are less likely to be affected by missing data. The last data set (Nurse) is an exception because there is only one useful predictor. Since we picked the most significant predictor to create missing values in, when the complete variable method is used, the only useful predictor was always deleted and thus the complete variable method always had worse performance than others. As a result, on this data set, none of the measurement blocks is marked as mostly tied. This is consistent with the observations made in Figure 8.

4.2 A Real Data Set With Missing Values

We now present a real data example with naturally occurred missing values. In this example, we try to model a company's bankruptcy status given its key financial statement items. The data are annual financial statement data and the predictions are sequential. That is, we build the tree on one year's data and then test its performance on the following year's data. For example, we build a tree on 1987's data and test its performance on 1988's data, then build a tree on 1988's data and test it on 1989 data, and so on.

The data are retrieved from Compustat North America (a database of U.S. and Canadian fundamental and market information on more than 24,000 active and inactive publicly held companies). Following Altman and Sabato (2005), twelve variables from the data base are used as potential predictors: Current Assets, Current Liabilities, Assets, Sales, Operating Income Before Depreciation, Retained Earnings, Net Income, Operating Income After Depreciation, Working Capital, Liabilities, Stockholder's Equity and year. The response variable, bankruptcy status, is determined using two footnote variables, the footnote for Sales and the footnote for Assets. Companies with remarks corresponding to "Reflects the adoption of fresh-start accounting upon emerging from Chapter 11 bankruptcy" or "Company in bankruptcy or liquidation" are marked as bankruptcy. The data include all active companies, and span 19 years from 1987 to 2005. There are 177560 observations in the original retrieved data, but 76504 of the observations have no data except for the company identifications, and are removed from the data set, resulting in 99056 observations. There are 19238 (19.4%) observations containing missing values and there are 56820 (4.8%) missing data values.

According to the results in Sections 3 and 4.1, there are two criteria that differentiate the performance of different missing data methods, that is, whether or not there are missing values in the testing set and whether or not the missingness depends on the response variable. In the bankruptcy data, there are missing values in every year's data, and thus missing values in each testing data set. To assess the dependence of the missingness on the response variable, the following test is carried out. First, we define twelve new binary missingness indicators corresponding to the original twelve predictors. Each indicator takes on value 1 if the original value for the associated variable is missing and 0 if the original value is observed for that observation. We then build a tree for each year's data using the indicators as the predictors and the original response variable, the bankruptcy status, as the response variable. From 1987 to 2000, the tree makes no split, indicating the tree algorithm is not able to establish a relationship between the missingness and the response variable. From 2001 to 2005, the classification tree consistently splits on the missingness indicators of Sales and Retained Earnings. This indicates that the missingness of these predictors has information about the response variable in these years, and the MGP across the years is fairly consistent in missingness in sales and retained earnings being related to bankruptcy status. However, the AUC values calculated from the trees built with the missingness indicators are not very high, all being between 0.5 and 0.6. Therefore, the relationship is not a very strong one.

Given these observations and the fact that the sample sizes are fairly large, we would make the following propositions based on our earlier conclusions. First, from 1988 to 2001 (since the tree tested on 2001 data is built on 2000 data), different missing data methods should have similar performance, with no clear winners. However, from year 2002 to year 2005, the separate class method should have better performance than the others (but perhaps not much better since the relationship between missingness and the response is not very strong). The actual relative performance of different missing data methods is shown in Figure 13. Since surrogate split is realized using

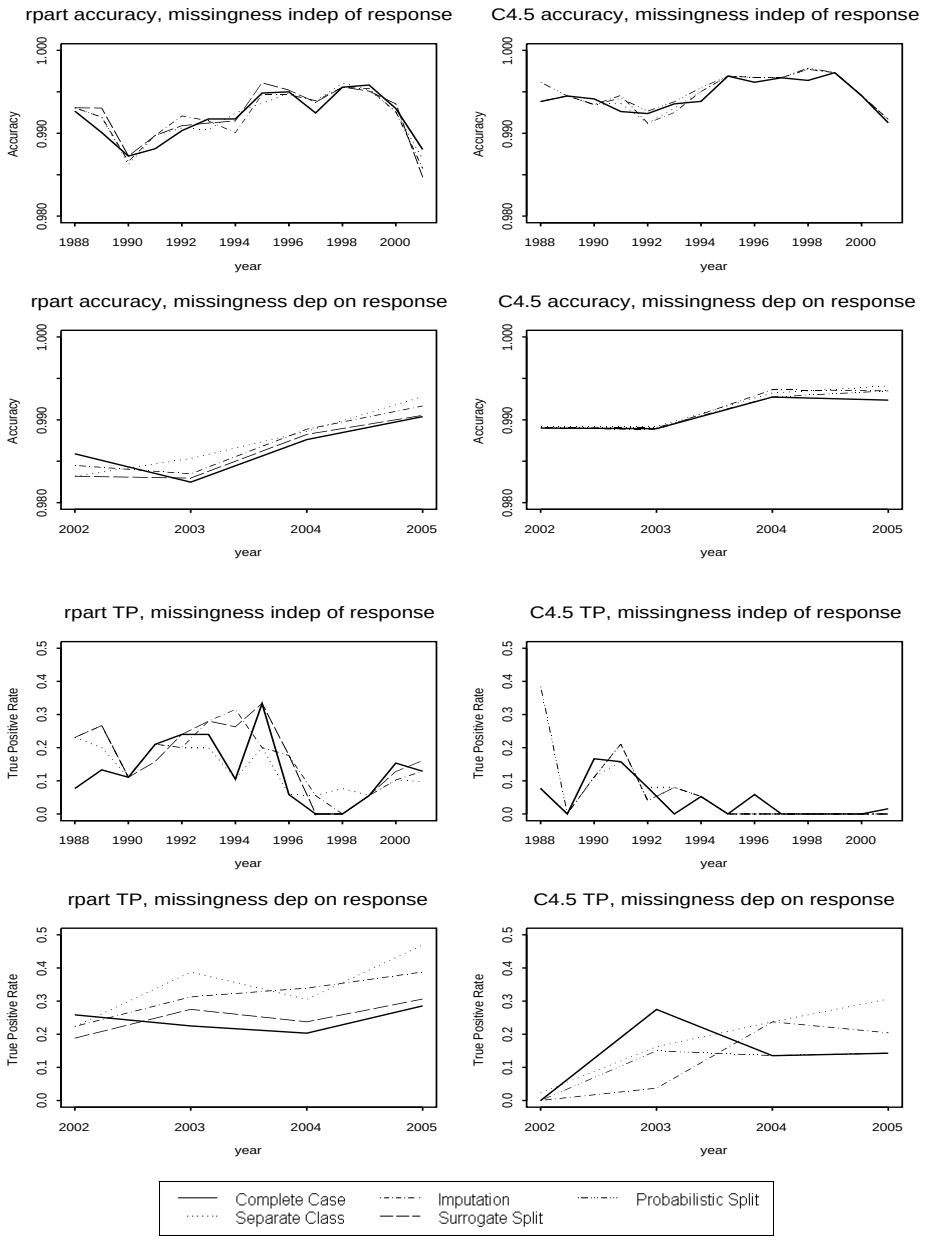

Figure 13: The relative performance of all of the missing data methods on the bankruptcy data. The left column gives methods using RPART (and includes all of the methods except for probabilistic split) and the right column gives methods using C4.5 (and includes all of the methods except for surrogate split). The top rows are performance in terms of accuracy while the bottom rows are in terms of true positive rate.

RPART while probabilistic split is realized using C4.5, we run all of the other methods using both RPART and C4.5 so that we can compare both surrogate split and probabilistic split with all of the other methods. In Figure 13, the plots on the left are the results from RPART, which include all of

the missing data methods except for probabilistic split. The plots on the right are the results from C4.5, which include all of the missing data methods except for surrogate split. The performances of methods common to both plots are slightly different because of differences between C4.5 and RPART in splitting and pruning rules. Both the accuracy and the true positive rates are shown. Since the number of actual bankruptcy cases in the data is small, the accuracy is always very high. The true positive rate is defined as

$$TP = \frac{\text{Number of correctly predicted bankruptcy cases}}{\text{Actual number of bankruptcy cases}}.$$

The graphs in the first and the second rows are for accuracies, with the first row for the first time period from 1988 to 2001 and the second row for the second time period from 2002 to 2005. The graphs in the third and the fourth rows are for true positive rates, with the third row for the first time period from 1988 to 2001 and the fourth row for the second time period from 2002 to 2005. It is apparent that in the first time period, there are no clear winners. However, in the second time period, separate class is a little better than the others, in line with expectations.

5. Extension To Logistic Regression

One obvious observation from this study is that when missing values occur in both the model building and model application stages, it should be considered as part of the data generating process rather than a separate mechanism. That is, taking the missingness into consideration can improve predictive performance, sometimes significantly. This should also apply to other supervised learning methodologies, non-parametric or parametric, when predictive performance is concerned. We present here the results from a real data analysis study involving logistic regression, similar to the one presented in Section 4.1. Missing values are generated the same way as in Section 4.1 and then logistic regression models (without variable selection) with different ways of handling missing data are applied to those data sets. Finally a tally is made on the relative performances of different missing data methods. Results measured in accuracy, bagged accuracy, AUC and bagged AUC are almost identical to each other; results in terms of accuracy are shown in Figure 14.

Included in the study are five ways of handling missing data: using only complete cases (complete case method), including a missingness dummy variable in the explanatory variable (dummy method, sometimes called the missing-indicator method),[1] building separate models for data with values missing and data without missing values (by-group method),[2] imputing missing values with grand mean/mode (imputation method), and only using predictors without missing values (complete variable method). Note that the methods using a dummy variable and building separate models for

1. If explanatory variable X_1 has missing values, then we create a missingness dummy variable M_1 that has value 1 if X_1 is observed and 0 otherwise. Then M_1 and $X_1 * M_1$ are both used as explanatory variables. The result of this set-up is that the effect of X_1 is fit on the observations with X_1 observed but a single mean value is fit to the observations with X_1 missing. All of the observations, with or without X_1 values, have the same coefficients for all of the other explanatory variables. Jones (1996) showed that this method can result in biased coefficient estimates in regression modeling, but did not address the question of predictive accuracy that is the focus here.
2. The biggest difference between the by-group method and the dummy method is whether the explanatory variables, other than the one containing missing values, have different coefficients or not. The by-group method fits two separate models to observations with and without missing values. Therefore, even if an explanatory variable is fully observed, its coefficient would most likely be different for fully observed observations and for observations with missing values. The dummy method, on the other hand, fits a single model to the entire data set so that variables that are fully observed will have the same coefficients whether an observation has missing values or not.

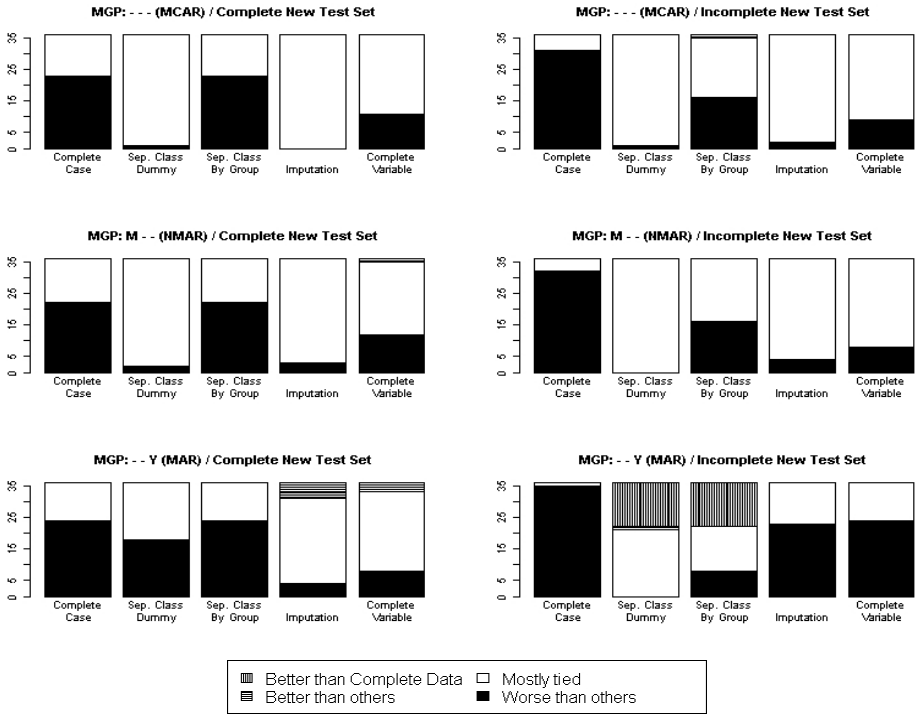

Figure 14: A tally of the relative out-of-sample performance with logistic regression measured in accuracy of all of the missing data methods on the 36 data sets.

observations with and without missing values each are analogous to the separate class method for trees. The most obvious observation is that when missingness is related to the response variable and missingness occurs in the test set, the dummy method and the by-group method dominate the other methods; in fact, more than a third of the time, they perform better than logistic regression on the original full data. Comparing Figure 14 with Figure 9, we see a clear similarity, in that the methods using a separate class model missingness directly, and thus use the information contained in the missingness about the response variable most efficiently. This suggests that the result that predictive performance of supervised learning methods is driven by the dependence (or lack of dependence) on the response variable is not limited to trees, but is rather a general phenomenon.

6. Conclusion And Future Study

The main conclusions from this study are as follows:

1. The two most important criteria that differentiate the performance of different missing data methods are whether or not the testing set is complete and whether or not the missingness depends on the response variable. There is strong evidence, both analytically and empirically, that separate class is the best missing data method to use when the testing data also contains missing values and the missingness is dependent upon the response variable.

   In practice, one way to detect the dependence of missingness on the response variable is to try building a model, with a classification tree being a natural choice, of the response variable on

the missingness indicators (which equals to 1 if the corresponding original value is missing and 0 otherwise). If such a model supports a relationship, then it is an indication that the missingness is related to the response variable.

2. The performance of classification trees is on average not too negatively affected by missing values, except for the complete case method and the mean imputation method, which are sensitive to different missing rates. Separate class tend to perform better when the missing rate is neither too high nor too low, trading off between information loss due to missing values and information gain from the informative MGP.

3. The original full data AUC has an impact on the performance of separate class method. The higher the original AUC, the more severe the information loss due to missing value, and thus relatively the worse the performance of the separate class method.

The consistency of these results across the theoretical analyses, simulations from the artificial data, and simulations based on real data provides strong support for their general validity.

The findings here also have implications beyond analysis of the data at hand. For example, since missingness that is dependent on the response variable can actually improve predictive performance, it is clear that expending time, effort, and money to recover the missing values is potentially a poor way to allocate resources. Another interesting implication of these results is related to data disclosure limitation. It is clear that any masking of values must be done in a way that is independent of the response variables of interest (or any predictors highly related to such variables), since otherwise data disclosure using regression-type methods (Palley and Simonoff, 1987) could actually increase.

Classification trees are designed for the situation where the response variable is categorical, not just binary; it would be interesting to see how these results carry over to that situation. Tree-based methodologies for the situation with a numerical response have also been developed (i.e., regression trees), and the problems of missing data occur in that context also. Investigating such trees would be a natural extension of this paper. In this paper, we focused on base form classification trees using C4.5 and RPART, although bagging was included in the study. It would be worthwhile to see how the performance of different missing data methods is affected by different tree features such as stopping and pruning or when techniques such as cross-validation, tree ensembles, etc. are used.

Moreover, as was shown in Section 5, the relationship between the missingness and the response variable can be helpful in prediction when missingness occurs in both the training data and the testing data in situations other than classification trees. This is very likely true for other supervised learning methods, and thus testing more learning methods would also be a natural extension to this study.

Acknowledgments

The authors would like to thank Foster Provost for his helpful comments, which have helped us to greatly improve the quality of this paper. The authors also thank several anonymous referees for helpful comments that greatly improved the presentation of results in the paper.

$P(Y=0\|X=0,$ with Missing Data)	$>T$	$>T$
$P(Y=0\|X=1,$ with Missing Data)	$>T$	$\leq T$
P(Y=0|X=0) P(Y=0|X=1)		
$>T$ $>T$	1	$\frac{P(X=0,Y=0)+P(X=1,Y=1)}{P(Y=0)}$
$>T$ $\leq T$	$\frac{P(Y=0)}{P(X=0,Y=0)+P(X=1,Y=1)}$	1
$\leq T$ $>T$	$\frac{P(Y=0)}{P(X=0,Y=1)+P(X=1,Y=0)}$	$\frac{P(X=0,Y=0)+P(X=1,Y=1)}{P(X=0,Y=1)+P(X=1,Y=0)}$
$\leq T$ $\leq T$	$\frac{P(Y=0)}{P(Y=1)}$	$\frac{P(X=0,Y=0)+P(X=1,Y=1)}{P(Y=1)}$
$P(Y=0\|X=0,$ with Missing Data)	$\leq T$	$\leq T$
$P(Y=0\|X=1,$ with Missing Data)	$>T$	$\leq T$
P(Y=0|X=0) P(Y=0|X=1)		
$>T$ $>T$	$\frac{P(X=0,Y=1)+P(X=1,Y=0)}{P(Y=0)}$	$\frac{P(Y=1)}{P(Y=0)}$
$>T$ $\leq T$	$\frac{P(X=0,Y=1)+P(X=1,Y=0)}{P(X=0,Y=0)+P(X=1,Y=1)}$	$\frac{P(Y=1)}{P(X=0,Y=0)+P(X=1,Y=1)}$
$\leq T$ $>T$	1	$\frac{P(Y=1)}{P(X=0,Y=1)+P(X=1,Y=0)}$
$\leq T$ $\leq T$	$\frac{P(X=0,Y=1)+P(X=1,Y=0)}{P(Y=1)}$	1

Table 3: *RelAcc* of tree built on data with missing values and tested on the original full data set when there is no variation from true DGP

Appendix A. Proofs of the Theorems

The relative accuracy (*RelAcc*) when there are missing values in the training set but not in the testing set can be summarized into Table 3, where T is the threshold value (an observation will be classified as class 0 if the predicted probability for it to be 0 is greater than T). The value of T reflects the misclassification cost. It is taken as 0.5 reflecting an equal misclassification cost. In Table 3, the columns show different rules given by the classification trees when there are missing values, and the rows show actual DGP's. The entries are the *RelAcc* values under different scenarios. For example, all of the entries on the diagonal are one's because the rules given by the classification trees when there are missing values are the same as the true DGP's and thus the accuracy achieved by the trees are the same with or without the missing values and thus $RelAcc = 1$. Cell $(1,2)$, for example, shows that if the true DGP is $P(Y=0|X=0) > T$ and $P(Y=0|X=1) > T$ but the classification tree gives rule $P(Y=0|X=0) > T$ and $P(Y=0|X=1) \leq T$ when there are missing values, that is, $P(Y=0|X=0, \text{with missing value}) > T$ and $P(Y=0|X=1 \text{with missing value}) \leq T$, then the relative accuracy is determined to be

$$\frac{P(X=0,Y=0)+P(X=1,Y=1)}{P(Y=0)}.$$

Proof of Theorem 1 : The expected performance of the complete case method when the missingness does not depend on the response variable and the testing set is complete.

Proof

First, we define A as the case-wise missingness indicator which equals 1 if the observation contains missing values in one or more of the predictors or 0 if the observation does not

contain missing values in any of the predictors. Y is the response variable and $\underline{X}$ is the vector of the predictors.

If only the complete cases are used, if $P(Y|A=0,\underline{X}) = P(Y|\underline{X})$, then only the diagonal in Table 3 can be achieved, and thus there is no loss in accuracy.

This condition will be satisfied if and only if the MGP is conditionally independent of Y given $\underline{X}$, that is, $P(A=0|\underline{X},Y) = P(A=0|\underline{X})$.

1. "$P(Y|A=0,\underline{X}) = P(Y|\underline{X}) \Rightarrow P(A=0|\underline{X},Y) = P(A=0|\underline{X})$"

$$\begin{aligned} P(A=0|\underline{X},Y) &= \frac{P(A=0,\underline{X},Y)}{P(\underline{X},Y)} \\ &= \frac{P(Y|A=0,\underline{X})P(A=0,\underline{X})}{P(\underline{X},Y)} \\ &= \frac{P(Y|\underline{X})P(A=0,\underline{X})}{P(Y|\underline{X})P(\underline{X})} \\ &= P(A=0|\underline{X}) \end{aligned}$$

2. "$P(Y|A=0,\underline{X}) = P(Y|\underline{X}) \Leftarrow P(A=0|\underline{X},Y) = P(A=0|\underline{X})$"

$$\begin{aligned} P(Y|A=0,\underline{X}) &= \frac{P(A=0,\underline{X},Y)}{P(A=0,\underline{X})} \\ &= \frac{P(A=0|\underline{X},Y)P(\underline{X},Y)}{P(A=0,\underline{X})} \\ &= \frac{P(A=0|\underline{X})P(\underline{X},Y)}{P(A=0|\underline{X})P(\underline{X})} \\ &= \frac{P(\underline{X},Y)}{P(\underline{X})} \\ &= P(Y|\underline{X}) \end{aligned}$$

■

Proof of Theorems 2 and 3 : The expected performance of the complete case method when the missingness depends on the response variable and the testing set is complete.

We first observe the following lemmas.

Lemma 7 *For the partition defined by the tree built on the original full data (and not changed by missing values), let the k^{th} section contain P^k proportion of data and within the partition, the majority class have proportion P^k_{mj}. Note that $\sum_{k=1}^{K} P^k = 1$, while the full data set accuracy, that is, the accuracy achievable with the full data set, is $\sum_k P^k P^k_{mj}$.)*

The rule for the k^{th} section will be classifying it as the majority class of the section. The impact of missing data on its rule is to either leave it unchanged or make it classify the data as the minority class instead of the majority class.

The smallest missing rate needed in k^{th} section to change the rule is $P(A=1|k) = 2P^k_{mj} - 1$, where A is defined as in Theorem 1, that is, it is the case-wise indicator, which takes value 1 if the observation contains missing value or 0 otherwise. If the rule is changed the loss in accuracy within that section is $2P^k_{mj} - 1$.

Proof

We assume the partition of the data is not changed by the missing values. The structure of the trees need not to be the same because different trees may lead to the same partition of data.

For any k, to make the rule of the k^{th} section change, we need to observe more minority class cases than the majority ones within that section. To achieve this in the most efficient way, we only make the majority ones missing. Originally, there are P_{mj}^k majorities and $1-P_{mj}^k$ minorities. Only when there are $P_{mj}^k-(1-P_{mj}^k)=2P_{mj}^k-1$ majorities missing will it become less than the minorities, so this is the smallest missing rate we need to make the rule change.

After the rule is changed, only $1-P_{mj}^k$ of the data, that is, the minorities, will be correctly classified. Therefore, the loss in accuracy is $P_{mj}^k-(1-P_{mj}^k)=2P_{mj}^k-1$.

■

Lemma 8 *For a given data set and the partition defined by the tree built on the full data set (which is not changed by the missing values), the largest loss in accuracy is* $\sum_k 2P_{mj}^k-1$. *The smallest missing rate needed to achieve this is also* $\sum_k 2P_{mj}^k-1$.

Proof

The largest loss is achieved if and only if the rules are changed in every section of data in the partition. The result then follows from Lemma 7.

■

Lemma 9 *For a certain missing rate, say* P_m, *the largest effect it can have on the classification accuracy of any data that won't be split is* P_m *itself.*

In this case, the data set has its majority proportion $P_{mj}=\frac{1}{2}(1+P_m)$.

Proof

Similar to the proof of Lemma 7, for missing values to have an impact on the classification rule, it has to switch the order status of the majority and minority. To achieve this, it has to be that $P_{mj}-(1-P_{mj})\leq P_m$. We know that once the rule is changed, the loss in accuracy is $P_{mj}-(1-P_{mj})$. Therefore, the largest loss is P_m when the equality holds. In this case, we have $P_{mj}=\frac{1}{2}(1+P_m)$.

■

We now prove Theorem 2.

Proof

For any data set, once it is partitioned and the partition is not changed by missing values, the rules in different sections of data are independent of each other, so we can look at them separately.

Suppose the data are partitioned into K segments, in which some contain missing data and the others do not. Let K_0 be the set of sections whose rules are changed by missing data and K_1 be the set of all other sections. Also let the k^{th} segment ($k = 1 \ldots K$) contain proportion P^k of the data. We have $\sum_{k=1}^{K} P^k = 1$.

Assume that the k^{th} segment ($k \in K_0$) contains proportion P_m^k of missing data. Then we have

$$\sum_{k \in K_0} P_m^k P^k \leq P_m.$$

For the k^{th} segment ($k \in K_0$), by Lemma 9, the largest possible loss in accuracy is P_m^k and it occurs if and only if $P_{mj}^k = \frac{1}{2}(1 + P_m^k)$. Therefore, the possible loss for the entire data set is

$$\sum_{k \in K_0} P_m^k P^k \leq P_m,$$

the largest loss being achieved when the equality holds. In that case, the rules in all of the categories that contain missing values are changed and the maximum loss is P_m.

■

We now prove Theorem 3.

Proof

Assuming the partitions of data are not changed by the missing values, we have

$$\begin{aligned} RelAcc &= \frac{\sum_{k=1}^{K} P_{mj}^k P^k - \sum_{k \in K_0} (\text{loss in accuracy in } k^{th} \text{ segment})}{\sum_{k=1}^{K} P_{mj}^k P^k} \\ &= 1 - \frac{\sum_{k \in K_0} (\text{loss in accuracy in } k^{th} \text{ segment})}{\sum_{k=1}^{K} P_{mj}^k P^k} \\ &= 1 - \frac{\sum_{k \in K_0} (\text{loss in accuracy in } k^{th} \text{ segment})}{\sum_{k \in K_0} P_{mj}^k P^k + \sum_{k \in K_1} P_{mj}^k P^k} \end{aligned}$$

This is an increasing function of $\sum_{k \in K_1} P_{mj}^k P^k$ in the denominator, which is independent of other factors; setting it to zero minimize the *relative accuracy*, so

$$RelAcc \leq 1 - \frac{\sum_{k \in K_0} (\text{loss in accuracy in } k^{th} \text{ segment})}{\sum_{k \in K_0} P_{mj}^k P^k}$$

Denote the numerator $\sum_{k \in K_0} (\text{loss in accuracy in } k^{th} \text{ segment})$ as a. Now, from the proof of Theorem 2, the numerator $a \leq P_m$ and the denominator $\sum_{k \in K_0} P_{mj}^k P^k = \frac{1}{2}(1 + a)$. So,

$$RelAcc \leq 1 - \frac{a}{\frac{1}{2}(1 + a)}$$

This is a decreasing function of a and subject to $a \leq P_m$. Therefore, the minimum *RelAcc* is achieved when $a = P_m$. This gives

$$\begin{aligned} RelAcc &\leq 1 - \frac{P_m}{\frac{1}{2}(1+P_m)} \\ &= \frac{1-P_m}{1+P_m} \end{aligned}$$

■

Proof of Theorem 4 : Some properties of probabilistic split when the missingness does not depend on both the predictor and the response variable.

Proof

1. Part 1
   - If the MGP is independent of Y given X, that is, $P(M|X,Y) = P(M|X)$ then $P(Y|M,X) = P(Y|X)$ by the proof of Theorem 1.
     The rules given by probabilistic split when there are missing values are as follows:

$$\begin{aligned} & P(Y=0|X=0, Prob\_split) \\ = \; & P(Y=0|M=0,X=0)P(M=0) + P(Y=0|M=1)P(M=1) \\ = \; & P(Y=0|X=0)P(M=0) + P(Y=0|M=1)P(M=1) \\ = \; & P(Y=0|X=0)P(M=0) \\ & +[P(Y=0,X=0|M=1) + P(Y=0,X=1|M=1)]P(M=1) \\ = \; & P(Y=0|X=0)P(M=0) \\ & +[P(Y=0|M=1,X=0)P(X=0|M=1) \\ & +P(Y=0|M=1,X=1)P(X=1|M=1)]P(M=1) \\ = \; & P(Y=0|X=0)P(M=0) + [P(Y=0|X=0)P(X=0|M=1) \\ & +P(Y=0|X=1)P(X=1|M=1)]P(M=1) \\ = \; & P(Y=0|X=0)P(M=0) + P(Y=0|X=0)P(M=1,X=0) \\ & +P(Y=0|X=1)P(M=1,X=1) \\ = \; & P(Y=0|X=0)[P(M=0) + P(M=1,X=0)] \\ & +P(Y=0|X=1)P(M=1,X=1) \end{aligned}$$

     and following the similar route, we can get

$$\begin{aligned} & P(Y=0|X=1, Prob\_split) \\ = \; & P(Y=0|X=1)[P(M=0) + P(M=1,X=1)] \\ & +P(Y=0|X=0)P(M=1,X=0). \end{aligned}$$

     Note that

$$P(M=0) + P(M=1,X=1) + P(M=1,X=0) = 1.$$

Therefore, both $P(Y=0|X=0,Prob\_split)$ and$P(Y=0|X=1,Prob_split)$ are weighted averages of $P(Y=0|X=0)$ and $P(Y=0|X=1)$.
It follows that if both $P(Y=0|X=0)$ and $P(Y=0|X=1)$ are greater (less) than 0.5, then both $P(Y=0|X=0,Prob\_split)$ and $P(Y=0|X=1,Prob\_split)$ are also greater (less) than 0.5.

- If the MGP is independent of X given Y, without loss of generality, we prove the case when $P(Y=0|X=0)>T=0.5$ and $P(Y=0|X=1)>T=0.5$.

$$
\begin{aligned}
& P(Y=0|X=0,Prob\_split) \\
= \; & \frac{P(M=0,X=0,Y=0)}{P(M=0,X=0)}P(M=0)+P(Y=0|M=1)P(M=1) \\
= \; & \frac{P(M=0|X=0,Y=0)P(X=0,Y=0)P(M=0)}{P(M=0,X=0)} \\
& +P(M=1,Y=0) \\
= \; & \frac{P(M=0|Y=0)P(X=0,Y=0)P(M=0)}{P(M=0,X=0)} \\
& +P(M=1,X=0,Y=0)+P(M=1,X=1,Y=0) \\
= \; & \frac{P(M=0|Y=0)P(Y=0|X=0)P(M=0)}{P(M=0,X=0)} \\
& +P(M=1,X=0|Y=0)P(Y=0) \\
& +P(M=1,X=1|Y=0)P(Y=0) \\
= \; & \frac{P(M=0|Y=0)P(Y=0|X=0)P(M=0)}{P(M=0,X=0)} \\
& +P(M=1|Y=0)P(X=0|Y=0)P(Y=0) \\
& +P(M=1|Y=0)P(X=1|Y=0)P(Y=0) \\
= \; & \frac{P(M=0|Y=0)P(Y=0|X=0)P(M=0)}{P(M=0,X=0)} \\
& +P(M=1|Y=0)P(Y=0|X=0)P(X=0) \\
& +P(M=1|Y=0)P(Y=0|X=1)P(X=1) \\
> \; & T[\frac{P(M=0|Y=0)P(M=0)}{P(M=0,X=0)} \\
& +P(M=1|Y=0)P(X=0)+P(M=1|Y=0)P(X=1)] \\
= \; & T[\frac{P(M=0|Y=0)P(M=0)}{P(M=0,X=0)}+P(M=1|Y=0)] \\
> \; & T(P(M=0|Y=0)+P(M=1|Y=0)) \\
= \; & T
\end{aligned}
$$

A similar argument gives $P(Y=0|X=1,Prob\_split)>T$.

2. Part 2

- If the MGP is independent of Y given X, then from the proof of part 1,

$$P(Y=0|X=0,Prob\_split)$$

$$\begin{aligned} &= P(Y=0|X=0)(P(M=0)+P(M=1,X=0)) \\ &\quad +P(Y=0|X=1)P(M=1,X=1) \end{aligned}$$

and

$$\begin{aligned} &P(Y=0|X=1,Prob\_split) \\ &= P(Y=0|X=1)(P(M=0)+P(M=1,X=1)) \\ &\quad +P(Y=0|X=0)P(M=1,X=0). \end{aligned}$$

Taking the difference, we get

$$\begin{aligned} &P(Y=0|X=0,Prob\_split)-P(Y=0|X=1,Prob\_split) \\ &= P(Y=0|X=0)(P(M=0)+P(M=1,X=0)) \\ &\quad +P(Y=0|X=1)P(M=1,X=1) \\ &\quad -[P(Y=0|X=1)(P(M=0)+P(M=1,X=1)) \\ &\quad +P(Y=0|X=0)P(M=1,X=0)] \\ &= P(Y=0|X=0)P(M=0)-P(Y=0|X=1)P(M=0) \\ &= (P(Y=0|X=0)-P(Y=0|X=1))P(M=0). \end{aligned}$$

Without loss of generality, assume $P(Y=0|X=0,Prob\_split) > T$ and $P(Y=0|X=1,Prob\_split) < T$. It then follows that $P(Y=0|X=0) > P(Y=0|X=1)$. There are three possibilities:

(a) $P(Y=0|X=0) > T > P(Y=0|X=1)$

(b) $T > P(Y=0|X=0) > P(Y=0|X=1)$

(c) $P(Y=0|X=0) > P(Y=0|X=1) > T$

Conditions (b) and (c) are not possible because in these two cases, X is actually not informative and by Part 1, probabilistic split will show they are not informative. Therefore, it holds that $P(Y=0|X=0) > T > P(Y=0|X=1)$.

- If the MGP is independent of X given Y, that is, $P(M|X,Y) = P(M|Y)$, we have

$$\begin{aligned} &P(Y=0|X=0,Prob\_split) \\ &= \frac{P(M=0,X=0,Y=0)}{P(M=0,X=0)}P(M=0)+P(Y=0|M=1)P(M=1) \\ &= \frac{P(M=0|X=0,Y=0)P(X=0,Y=0)P(M=0)}{P(M=0|X=0,Y=0)P(X=0,Y=0)+P(M=0|X=0,Y=1)P(X=0,Y=1)} \\ &\quad +P(Y=0|M=1)P(M=1) \\ &= \frac{P(M=0|Y=0)P(X=0,Y=0)P(M=0)}{P(M=0|Y=0)P(X=0,Y=0)+P(M=0|Y=1)P(X=0,Y=1)} \\ &\quad +P(Y=0|M=1)P(M=1) \\ &= \frac{P(M=0|Y=0)P(Y=0|X=0)P(M=0)}{P(M=0|Y=0)P(Y=0|X=0)+P(M=0|Y=1)P(Y=1|X=0)} \\ &\quad +P(Y=0|M=1)P(M=1), \end{aligned}$$

and following the same route, we have

$$\begin{aligned} &P(Y=0|X=1,Prob\_split) \\ &= \frac{P(M=0|Y=0)P(Y=0|X=1)P(M=0)}{P(M=0|Y=0)P(Y=0|X=1)+P(M=0|Y=1)P(Y=1|X=1)} \\ &\quad +P(Y=0|M=1)P(M=1). \end{aligned}$$

Therefore,

$$\begin{aligned}
& P(Y=0|X=0,Prob_split)-P(Y=0|X=1,Prob_split) \\
= & \frac{P(M=0|Y=0)P(Y=0|X=0)P(M=0)}{P(M=0|Y=0)P(Y=0|X=0)+P(M=0|Y=1)P(Y=1|X=0)} \\
& -\frac{P(M=0|Y=0)P(Y=0|X=1)P(M=0)}{P(M=0|Y=0)P(Y=0|X=1)+P(M=0|Y=1)P(Y=1|X=1)} \\
= & [P(Y=0|X=0)P(M=0|Y=0)P(Y=0|X=1) \\
& +P(Y=0|X=0)P(M=0|Y=1)P(Y=1|X=1) \\
& -P(Y=0|X=1)P(M=0|Y=0)P(Y=0|X=0) \\
& -P(Y=0|X=1)P(M=0|Y=1)P(Y=1|X=0)]\frac{P(M=0|Y=0)P(M=0)}{D_1D_2} \\
= & [P(Y=0|X=0)-P(Y=0|X=1)]\frac{P(M=0|Y=1)P(M=0|Y=0)P(M=0)}{D_1D_2} \\
= & [P(Y=0|X=0)-P(Y=0|X=1)]K
\end{aligned}$$

where

$$D_1 = P(M=0|Y=0)P(Y=0|X=0)+P(M=0|Y=1)P(Y=1|X=0),$$
$$D_2 = P(M=0|Y=0)P(Y=0|X=1)+P(M=0|Y=1)P(Y=1|X=1)$$

and

$$K = \frac{P(M=0|Y=1)P(M=0|Y=0)P(M=0)}{D_1D_2}.$$

Since K is always positive as long as there are different Y values observed, we can see that the probabilistic split preserves the order of the conditional probability of Y given X.

Now, without loss of generality, assume $P(Y=0|X=0,Prob\_split) > T$ and $P(Y=0|X=1,Prob\_split) < T$. It follows that $P(Y=0|X=0) > P(Y=0|X=1)$ because probabilistic split preserves the correct order. There are three possibilities:

(a) $P(Y=0|X=0) > T > P(Y=0|X=1)$

(b) $T > P(Y=0|X=0) > P(Y=0|X=1)$

(c) $P(Y=0|X=0) > P(Y=0|X=1) > T$

Conditions (b) and (c) are not possible because in these two cases, X is actually not informative and by the earlier result in Part 1, probabilistic split will show they are not informative. Therefore, it holds that $P(Y=0|X=0) > T > P(Y=0|X=1)$.

3. Part 3

The results of Part 1 and Part 2 lead to the simplification of Table 3 into Table 4.

Without loss of generality, we provide the proof only for the case when $P(Y=0|X=0) > T$ and $P(Y=0|X=1) \leq T$ but $P(Y=0|X=0,prob\_split) > T$ and $P(Y=0|X=1,prob\_split) > T$, where *RelAcc* is

$$RelAcc = \frac{P(Y=0)}{P(X=0,Y=0)+P(X=1,Y=1)}.$$

It suffices to show that $P(Y=0) > 0.5$

- If M is independent of Y given X,

$$\begin{aligned}
& P(Y=0) \\
= & P(X=0,Y=0)+P(X=1,Y=0)
\end{aligned}$$

Simplified possibilities	$>T$ $>T$	$>T$ $\leq T$	$\leq T$ $>T$	$\leq T$ $\leq T$
Full data				
$>T$ $>T$	1	—	—	—
$>T$ $\leq T$	$\frac{P(Y=0)}{P(X=0,Y=0)+P(X=1,Y=1)}$	1	—	$\frac{P(Y=1)}{P(X=0,Y=0)+P(X=1,Y=1)}$
$\leq T$ $>T$	$\frac{P(Y=0)}{P(X=0,Y=1)+P(X=1,Y=0)}$	—	1	$\frac{P(Y=1)}{P(X=0,Y=1)+P(X=1,Y=0)}$
$\leq T$ $\leq T$	—	—	—	1

Table 4: *RelAcc* with a 2×2 table of probabilistic split when the missingness is independent of either X or Y or both

$$\begin{aligned}
&= P(M=0,X=0,Y=0)+P(M=1,X=0,Y=0)\\
&\quad +P(Y=0|X=1)P(X=1)\\
&= P(Y=0|M=0,X=0)P(M=0,X=0)\\
&\quad +P(Y=0|M=1,X=0)P(M=1,X=0)+P(Y=0|X=1)P(X=1)\\
&=^1 P(Y=0|X=0)P(M=0,X=0)+P(Y=0|X=0)P(M=1,X=0)\\
&\quad +P(Y=0|X=1)P(X=1)\\
&>^2 P(Y=0|X=0)P(M=1,X=0)+P(Y=0|X=1)P(M=0,X=0)\\
&\quad +P(Y=0|X=1)P(X=1)\\
&= P(Y=0|X=1)(P(M=0)+P(M=1,X=1))\\
&\quad +P(Y=0|X=0)P(M=1,X=0)\\
&= P(Y=0|X=1,prob_split)\\
&> 0.5
\end{aligned}$$

where 1 follows because $P(Y|M,X)=P(Y|X)$ and 2 follows because $P(Y=0|X=0)>T\geq P(Y=0|X=1)$. Therefore,

$$\frac{P(Y=0)}{P(X=0,Y=0)+P(X=1,Y=1)}>P(Y=0)>0.5$$

- If M is independent of X given Y,

$$P(Y=0)=P(M=0,Y=0)+P(M=1,Y=0)$$

and by assumption,

$$\begin{aligned}
&P(Y=0|X=1,prob_split)\\
=\;&P(Y=0|M=0,X=1)P(M=0)+P(M=1,Y=0)\\
>\;&0.5
\end{aligned}$$

If $P(M=0,Y=0)>P(Y=0|M=0,X=1)P(M=0)$, then $P(Y=0)>P(Y=0|X=1,prob\_split)>0.5$, it suffices to show

$$P(M=0,Y=0)>P(Y=0|M=0,X=1)P(M=0).$$

By the earlier results in Part 2, probabilistic split preserves the order of conditional probabilities of Y given X when the missingness is conditionally independent of X given Y, that is, in this case, since

$$P(Y=0|X=0) > T \geq P(Y=0|X=1),$$

we have

$$\begin{aligned}
& P(Y=0|X=0,Prob_split) - P(Y=0|X=1,Prob_split) \\
= \;& P(Y=0|M=0,X=0)P(M=0) + P(Y=0|M=1)P(M=1) \\
& -(P(Y=0|M=0,X=1)P(M=0) + P(Y=0|M=1)P(M=1)) \\
= \;& (P(Y=0|M=0,X=0) - P(Y=0|M=0,X=1))P(M=0) \\
> \;& 0.
\end{aligned}$$

That is, $P(Y=0|M=0,X=0) > P(Y=0|M=0,X=1)$. We then have

$$\begin{aligned}
& P(Y=0|M=0,X=0) > P(Y=0|M=0,X=1) \\
\Rightarrow \;& P(Y=0|M=0,X=0)P(M=0,X=0) \\
& > P(Y=0|M=0,X=1)P(M=0,X=0) \\
\Rightarrow \;& P(M=0,X=0,Y=0) > P(Y=0|M=0,X=1)P(M=0,X=0) \\
\Rightarrow \;& P(M=0,X=0,Y=0) + P(M=0,X=1,Y=0) \\
& > P(Y=0|M=0,X=1)P(M=0,X=0) + P(M=0,X=1,Y=0) \\
\Rightarrow \;& P(M=0,Y=0) \\
& > P(Y=0|M=0,X=1)P(M=0,X=0) + P(Y=0|M=0,X=1)P(M=0,X=1) \\
\Rightarrow \;& P(M=0,Y=0) > P(Y=0|M=0,X=1)(P(M=0,X=0) + P(M=0,X=1) \\
\Rightarrow \;& P(M=0,Y=0) > P(Y=0|M=0,X=1)P(M=0)
\end{aligned}$$

■

Proof of Theorem 5 : Some properties of the mode imputation when the missingness does not depend on the response variable.

Proof

Without loss of generality, we assume that $P(X=0|M=0) > P(X=1|M=0)$, that is, there are more X=0 cases observed than X=1 ones. As a result, all of the missing X values will be labeled as X=0, the observed mode. Then the decision rules when the mode imputation is used can be written as

$$\begin{aligned}
& P(Y=0|X=0,Imp) \\
= \;& \frac{P(M=0,X=0,Y=0) + P(M=1,Y=0)}{P(M=0,X=0) + P(M=1)} \\
= \;& \frac{P(M=0,X=0,Y=0) + P(M=1,X=0,Y=0) + P(M=1,X=1,Y=0)}{P(M=0,X=0) + P(M=1)} \\
= \;& \frac{P(X=0,Y=0) + P(M=1,X=1,Y=0)}{P(X=0) + P(M=1,X=1)} \\
= \;& \frac{P(X=0,Y=0) + P(Y=0|M=1,X=1)P(M=1,X=1)}{P(X=0) + P(M=1,X=1)}
\end{aligned}$$

$$
\begin{aligned}
= & \frac{P(Y=0|X=0)P(X=0)+P(Y=0|X=1)P(M=1,X=1)}{P(X=0)+P(M=1,X=1)} \\
& P(Y=0|X=1,Imp) \\
= & P(Y=0|M=0,X=1) \\
= & P(Y=0|X=1)
\end{aligned}
$$

1. Note that $P(Y=0|X=0,Imp)$ is a weighted average of $P(Y=0|X=0)$ and $P(Y=0|X=1)$. Therefore, if they are both larger (or smaller) than 0.5, $P(Y=0|X=0,Imp)$ will also be, and thus it gives the same rule as $P(Y=0|X=0)$. Moreover, $P(Y=0|X=1,Imp)=P(Y=0|X=1)$, so it also gives the correct rule.
2. Suppose

$$
\begin{aligned}
P(Y=0|X=0,Imp) &> 0.5 \\
P(Y=0|X=1,Imp) &< 0.5,
\end{aligned}
$$

   then $P(Y=0|X=1)=P(Y=0|X=1,Imp)<0.5$, which is always correct. Moreover, note that $P(Y=0|X=0,Imp)$ is a weighted average of $P(Y=0|X=0)$ and $P(Y=0|X=1)$. Since $P(Y=0|X=0,Imp)>0.5$ and $P(Y=0|X=1)<0.5$, we must have $P(Y=0|X=0)>0.5$. Therefore, $P(Y=0|X=0,Imp)$ gives the correct rule.
3. Again the possibilities simplify to Table 4. Without loss of generality, we prove the situation when both $P(Y=0|X=0,Imp)$ and $P(Y=0|X=1,Imp)$ are greater than 0.5, that is

$$
\begin{aligned}
& P(Y=0|X=0,Imp) \\
= & \frac{P(Y=0|X=0)P(X=0)+P(Y=0|X=1)P(M=1,X=1)}{P(X=0)+P(M=1,X=1)} \\
> & 0.5 \\
& P(Y=0|X=1,Imp) \\
= & P(Y=0|X=1) \\
> & 0.5
\end{aligned}
$$

   Under the assumption that $P(X=0|M=0)>P(X=1|M=0)$, the missing values have an effect only if $P(Y=0|X=0)<0.5$ and $P(Y=0|X=1)>0.5$. In this case, the relative accuracy is $\frac{P(Y=0)}{P(X=0,Y=1)+P(X=1,Y=0)}$. This is the cell of the 3^{rd} row and the 1^{st} column in Table 4.
   But,

$$
\begin{aligned}
& \frac{P(Y=0)}{P(X=0,Y=1)+P(X=1,Y=0)} \\
> & P(Y=0) \\
= & P(X=0,Y=0)+P(X=1,Y=0) \\
>^{1} & 0.5(P(X=0)+P(M=1,X=1))-P(Y=0|X=1)P(M=1,X=1) \\
& +P(X=1,Y=0) \\
= & 0.5(P(X=0)+P(M=1,X=1))+P(Y=0|X=1)(P(X=1)-P(M=1,X=1)) \\
= & 0.5(1-P(M=0,X=1))+P(Y=0|X=1)P(M=0,X=1) \\
= & 0.5-0.5P(M=0,X=1))+P(Y=0|X=1)P(M=0,X=1) \\
>^{2} & 0.5-0.5P(M=0,X=1))+0.5P(M=0,X=1) \\
> & 0.5,
\end{aligned}
$$

where 1 follows because

$$\begin{aligned} & P(Y=0|X=0,Imp) \\ = & \frac{P(Y=0|X=0)P(X=0)+P(Y=0|X=1)P(M=1,X=1)}{P(X=0)+P(M=1,X=1)} \\ > & 0.5. \end{aligned}$$

By rearranging terms,

$$\begin{aligned} & P(Y=0|X=0)P(X=0) \\ = & P(X=0,Y=0) \\ > & 0.5(P(X=0)+P(M=1,X=1))-P(Y=0|X=1)P(M=1,X=1), \end{aligned}$$

where 2 follows because P(Y=0|X=1)=P(Y=0|X=1,Imp)>0.5.

■

Proof of Theorem 6 : The dominance of the separate class method when there are missing values in both the training set and the testing set and the missingness depends on the response variable.

Proof

When there are missing data in X in both the training set and the testing set, the finest partition of the data will be $X=0$, $X=1$ and X is missing. The best rule we can derive is to classify the majority class in each of these three partitions. This is achieved by using the separate class method.

■

References

P.D. Allison. Missing data. In *Sage University Papers Series on Quantitative Applications in the Social Sciences*, 07-136. Sage, Thousand Oaks, CA, 2001.

E.I. Altman and G. Sabato. Effects of the new basel capital accord on bank capital requirements for smes. *Journal of Financial Services Research*, 28(1):15–42, 2005. URL `http://econpapers.repec.org/RePEc:kap:jfsres:v:28:y:2005:i:1:p:15-42`.

G.E.A.P.A. Batista and M.C. Monard. An analysis of four missing data treatment methods for supervised learning. *Applied Artificial Intelligence*, 17:519–533, 2003.

L. Breiman, J. Friedman, R.A. Olshen, and C.J. Stone. *Classification and Regression Trees*. Chapman and Hall/CRC, Boca Raton, Fla, 1998.

Y. Ding and J.S. Simonoff. An investigation of missing data methods for classification trees applied to binary response data. Working paper 2008-SOR-1, Stern School of Business, New York University, 2008.

R.O. Duda, P.E. Hart, and D.G. Stork. *Pattern Classification*. Wiley-Interscience, 2001.

A. Feelders. Handling missing data in trees: Surrogate splits or statistical imputation? *Principles of Data Mining and Knowledge Discovery*, 1704:329–334, 1999.

Y. Fujikawa and T.B. Ho. Cluster-based algorithms for filling missing values. *Lecture Notes in Artificial Intelligence*, 2336:549–554, 2002. In *6th Pacific-Asia Conf. on Knowledge Discovery and Data Mining*, Taiwan, 6-9 May.

D.J. Hand. *Construction and Assessment of Classification Rules*. John Wiley and Sons, Chichester, 1997.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, pages 246–249. Springer Series in Statistics, 2001.

M.P. Jones. Indicator and stratification methods for missing explanatory variables in multiple linear regression. *Journal of the American Statistical Association*, 91:222–230, 1996.

A. Kalousis and M. Hilario. Supervised knowledge discovery from incomplete data. Cambridge, UK, 2000. Proceedings of the 2nd International Conference on Data Mining 2000, WIT Press.

H. Kim and S. Yates. Missing value algorithms in decision trees. In H. Bozdogan, editor, *Statistical Data Mining and Knowledge Discovery*, pages 155–172. Chapman & Hall/CRC, Boca Raton, Fla, 2003.

R.J.A. Little and D.B. Rubin. *Statistical Analysis with Missing Data*. Wiley, New York, second edition, 2002.

W.Z. Liu, A.P. White, S.G. Thompson, and M.A. Bramer. Techniques for dealing with missing values in classification. *Lecture Notes in Computer Science*, 1280:527–536, 1997.

M.A. Palley and J.S. Simonoff. The use of regression methodology for the compromise of confidential information in statistical databases. *ACM Transactions on Database Systems*, 12:593–608, 1987.

C. Perlich, F. Provost, and J.S. Simonoff. Tree induction vs. logistic regression: A learning curve analysis. *Journal of Machine Learning Research*, 4:211–255, 2003.

J.R. Quinlan. *C4.5 Programs for Machine Learning*. Morgan Kaufmann Publishers, San Francisco, CA, 1993.

D.B. Rubin. Inference and missing data. *Biometrika*, 63:581–592, 1976.

M. Saar-Tsechansky and F. Provost. Handling missing values when applying classification models. *Journal of Machine Learning Research*, 8:1625–1657, 2007.

T.M. Therneau and E.J. Atkinson. An introduction to recursive partitioning using the rpart routines. Technical report, Mayo Foundation, 1997.

B. Twala. An empirical comparison of techniques for handling incomplete data using decision trees. *Applied Artificial Intelligence*, 23:373–405, 2009.

B. Twala, M.C. Jones, and D.J. Hand. Good methods for coping with missing data in decision trees. *Pattern Recognition Letters*, 29:950–956, 2008.

J. Wang and X. Shen. Large margin semi-supervised learning. *Journal of Machine Learning Research*, 8:1867–1891, 2007.

S. Zhang, Z. Qin, C.X. Ling, and S. Sheng. Missing is useful: Missing values in cost-sensitive decision trees. *IEEE Transactions on Knowledge and Data Engineering*, 17:1689–1693, 2005.