An Analysis of Four Missing Data Treatment Methods for Supervised Learning

Gustavo E. A. P. A. Batista and Maria Carolina Monard
University of São Paulo - USP
Institute of Mathematics and Computer Science - ICMC
Department of Computer Science and Statistics - SCE
Laboratory of Computational Intelligence - LABIC
P. O. Box 668, 13560-970 - São Carlos, SP, Brazil
{gbatista, mcmonard}@icmc.usp.br

Abstract. One relevant problem in data quality is the presence of missing data. Despite the frequent occurrence and the relevance of missing data problem, many Machine Learning algorithms handle missing data in a rather naive way. However, missing data treatment should be carefully thought, otherwise bias might be introduced into the knowledge induced. In this work we analyse the use of the k-nearest neighbour as an imputation method. Imputation is a term that denotes a procedure that replaces the missing values in a data set by some plausible values. One advantage of this approach is that the missing data treatment is independent of the learning algorithm used. This allows the user to select the most suitable imputation method for each situation. Our analysis indicates that missing data imputation based on the k-nearest neighbour algorithm can outperform the internal methods used by C4.5 and CN2 to treat missing data, and can also outperform the mean or mode imputation method, which is a method broadly used to treat missing values.

1 Introduction

One relevant problem in data quality is the presence of missing data. Missing data may have different sources such as death of patients, equipment malfunctions, refusal of respondents to answer certain questions, and so on. In addition, a significant fraction of data can be erroneous, and the only alternative may be discarding the erroneous data.

Data quality is a major concern in Machine Learning — ML — and other correlated areas, such as Data Mining — DM — and Knowledge Discovery from Databases — KDD. Despite the frequent occurrence of missing data in real world data sets, ML algorithms handle missing data in a rather naive way. Missing data treatment should be carefully thought, otherwise bias might be introduced into the knowledge induced.

In most cases, data sets attributes are not independent from each other. Thus, through the identification of relationships among attributes, missing values can be determined. *Imputation* is a term that denotes a procedure that replaces the missing values in a data set by some plausible values. One advantage of this approach is that

the missing data treatment is independent of the learning algorithm used. This allows the user to select the most suitable imputation method for each situation.

The objective of this work is to analyse the performance of the k-nearest neighbour as an imputation method, comparing its performance with other three missing data treatment methods. The first method is the mean or mode imputation. This method is very simple and broadly used. It consists of replacing every missing value of an attribute by the mean (if the attribute is quantitative) or mode (if the attribute is qualitative) of its known values. The other two methods are the internal missing data treatment strategies used by two well known ML algorithms: CN2 [3] and C4.5 [12].

This work is organized as follows: Section 2 describes the taxonomy proposed in [10] to classify the degree of randomness of missing data in a data set, and surveys the most widely used methods for missing data treatment; Section 3 describes the imputation method; Section 4 presents the k-nearest neighbour as an imputation method for treating missing values; Section 5 describes how the ML algorithms C4.5 and CN2 treat missing data internally; Section 6 performs a comparative study of the k-nearest neighbour algorithm as an imputation method with the internal methods used by C4.5 and CN2 to treat missing data, as well as the mean or mode imputation; finally, Section 7 presents the conclusions of this work.

2 Randomness of Missing Data and Methods for Treating Missing Data

Missing data randomness can be divided into three classes, as proposed by [10]:

1. *Missing completely at random (MCAR).* This is the highest level of randomness. It occurs when the probability of an instance (case) having a missing value for an attribute does not depend on either the known values or the missing data. In this level of randomness, any missing data treatment method can be applied without risk of introducing bias on the data;

2. *Missing at random (MAR).* When the probability of an instance having a missing value for an attribute may depend on the known values, but not on the value of the missing data itself;

3. *Not missing at random (NMAR).* When the probability of an instance having a missing value for an attribute could depend on the value of that attribute.

Several methods have been proposed in the literature to treat missing data. Many of these methods, such as case substitution, were developed for dealing with missing data in sample surveys, and have some drawbacks when applied to the Data Mining context. Other methods, such as replacement of missing values by the attribute mean or mode, are very naive and should be carefully used to avoid insertion of bias.

In a general way, missing data treatment methods can be divided into the following three categories [10]:

1. *Ignoring and discarding data.* There are two main ways to discard data with missing values. The first one is known as *complete case analysis*. It is available in all statistical packages and is the default method in many programs. This method consists of discarding all instances (cases) with missing data. The second method is known

as *discarding instances and/or attributes*. This method consists of determining the extent of missing data on each instance and attribute, and delete the instances and/or attributes with high levels of missing data. Before deleting any attribute, it is necessary to evaluate its relevance to the analysis. Unfortunately, relevant attributes should be kept even with a high degree of missing values. Both methods, complete case analysis and discarding instances and/or attributes, should be applied only if missing data are MCAR, because missing data that are not MCAR have non-random elements that can bias the results;

2. *Parameter estimation.* Maximum likelihood procedures are used to estimate the parameters of a model defined for the complete data. Maximum likelihood procedures that use variants of the Expectation-Maximization algorithm [4] can handle parameter estimation in the presence of missing data;

3. *Imputation.* Imputation is a class of procedures that aims to fill in the missing values with estimated ones. The objective is to employ known relationships that can be identified in the valid values of the data set to assist in estimating the missing values. This papers focus on imputation of missing data. More details about this class of methods are described next.

3 Imputation Methods

Imputation methods involve replacing missing values with estimated ones based on information available in the data set. There are many options varying from naive methods, like mean imputation, to some more robust methods based on relationships among attributes. A description of some widely used imputation methods follows:

1. *Case substitution.* This method is typically used in sample surveys. One instance with missing data (for example, a person that cannot be contacted) is replaced by another nonsampled instance;

2. *Mean or mode imputation.* This is one of the most frequently used methods. It consists of replacing the missing data for a given attribute by the mean (quantitative attribute) or mode (qualitative attribute) of all known values of that attribute;

3. *Hot deck and cold deck.* In the hot deck method, a missing attribute value is filled in with a value from an estimated distribution for the missing value from the current data. Hot deck is typically implemented in two stages. In the first stage, the data are partitioned into clusters. In the second stage, each instance with missing data is associated with one cluster. The complete cases in a cluster are used to fill in the missing values. This can be done by calculating the mean or mode of the attribute within a cluster. Cold deck imputation is similar to hot deck but the data source must be other than the current data source;

4. *Prediction model.* Prediction models are sophisticated procedures for handling missing data. These methods consist of creating a predictive model to estimate values that will substitute the missing data. The attribute with missing data is used as class-attribute, and the remaining attributes are used as input for the predictive

model. An important argument in favour of this approach is that, frequently, attributes have relationships (correlations) among themselves. In this way, those correlations could be used to create a predictive model for classification or regression for, respectively, qualitative and quantitative attributes with missing data. Some of these relationships among the attributes may be maintained if they are captured by the predictive model. An important drawback of this approach is that the model estimated values are usually more well-behaved than the true values would be, *i.e.*, since the missing values are predicted from a set of attributes, the predicted values are likely to be more consistent with this set of attributes than the true (not known) values would be. A second drawback is the requirement for correlation among the attributes. If there are no relationships among attributes in the data set and the attribute with missing data, then the model will not be precise for estimating missing values.

4 Imputation with k-Nearest Neighbour

This work proposes the use of the k-nearest neighbour algorithm to estimate and substitute missing data. The main benefits of this approach are: *(i)* k-nearest neighbour can predict both qualitative attributes (the most frequent value among the k nearest neighbours) and quantitative attributes (the mean among the k nearest neighbours); *(ii)* There is no necessity for creating a predictive model for each attribute with missing data. Actually, the k-nearest neighbour algorithm does not create explicit models (like a decision tree or a set of rules), since the data set is used as a "lazy" model. Thus, the k-nearest neighbour algorithm can be easily adapted to work with any attribute as class, by just modifying the attributes to be considered in the distance metric. Also, this approach can easily treat examples with multiple missing values.

The main drawback of the k-nearest neighbour approach is that, whenever the k-nearest neighbour looks for the most similar instances, the algorithm searches through all the data set. This limitation can be very critical for KDD, since this research area has, as one of its main objectives, the analysis of large databases. Several works that aim to solve this limitation can be found in the literature. One method is the creation of a reduced training set for the k-nearest neighbour composed only by prototypical examples [13]. This work uses an access method called M-tree [2, 6], that was implemented in the k-nearest neighbour algorithm employed. Furthermore, M-trees can organize and search data sets based on a generic metric space. M-trees can drastically reduce the number of distance computations in similarity queries.

5 Missing Data Treatment by C4.5 and CN2

Both algorithm, C4.5 and CN2, were selected because they are well considered by the ML community. They induce propositional concepts: decision trees and rules, respectively. Furthermore, C4.5 seems to have a good internal algorithm to treat missing values, as shown in [5]. On the other hand, CN2 seems to use a rather simple method to treat missing data.

C4.5 and CN2 can handle missing values in any attribute, except the class attribute, for both training and test sets.

C4.5 uses a probabilistic approach to handle missing data. Given a training set, T, C4.5 finds a suitable test, based on a single attribute, that has one or more mutually exclusive outcomes $O_1, O_2, \ldots, O_n$. T is partitioned into subsets T_1, T_2, ..., T_n, where T_i contains all the instances in T that satisfy the test with outcome O_i. The same algorithm is applied to each subset T_i until a stop criteria is obeyed. C4.5 uses the *information gain ratio* measure to choose a good test to partition the instances. If there exist missing values in an attribute X, C4.5 uses the subset with all known values of X to calculate the information gain.

Once a test based on an attribute X is chosen, C4.5 uses a probabilistic approach to partition the instances with missing values in X. When an instance in T with known value is assigned to a subset T_i, this indicates that the probability of that instance belonging to subset T_i is 1 and to all other subsets is 0. When the value is not known, only a weaker probabilistic statement can be made. C4.5 associates to each instance in T_i a *weight* representing the probability of that instance belonging to T_i. If the instance has a known value, and satisfies the test with outcome O_i, then this instance is assigned to T_i with weight 1; if the instance has an unknown value, this instance is assigned to all partitions with different weights for each one. The weight for the partition T_i is the probability that instance belongs to T_i. This probability is estimated as the sum of the weights of instances in T known to satisfy the test with outcome O_i, divided by the sum of weights of the cases in T with known values on the attribute X.

The CN2 algorithm uses a rather simple imputation method to treat missing data. Every missing value is filled in with its attribute most common known value, before calculating the entropy measure [3].

6 Experimental Analysis

The main objective of the experiments conducted in this work is to evaluate the efficiency of the k-nearest neighbour algorithm as an imputation method to treat missing data, comparing its performance with the performance obtained by the internal algorithms used by C4.5 and CN2 to learn with missing data, and by the mean or mode imputation method.

In these experiments, missing values were artificially implanted, in different rates and attributes, into the data sets. The performance of all four missing data treatments were compared using cross-validation estimated error rates. In particular, we are interested in analysing the behaviour of these treatments when the amount of missing data is high since some researchers have reported finding databases where more than 50% of the data were missing [8].

The experiments were carried using four data sets from UCI [11]: Bupa, Cmc, Pima and Breast. The first three data sets have no missing values. Breast has very few cases with missing values (in total 16 cases or 2.28%) which were removed before starting the experiments. The main reason for not using data with missing values is the wish to have total control over the missing data in the data set. For instance, we would like the test sets to have no missing data. If some test set has missing data, then the inducer's ability to classify missing data properly may influence the result. This influence is undesirable since the objective of this work is to analyse the viability of the k-nearest neighbour as an imputation method for missing data and the inducer learning ability when missing

values are present.

Table 1 summarizes the data sets employed in this study. It shows, for each data set, the number of instances (#Instances), number and percentage of duplicate (appearing more than once) or conflicting (same attribute-value but different class attribute) instances, number of attributes (#Attributes), number of quantitative and qualitative attributes, class attribute distribution and the majority class error. This information was obtained using the MLC++ *info* utility [7].

Data set	# Instances	#Duplicate or conflicting (%)	#Attributes (quanti., quali.)	Class	Class %	Majority Error
bupa	345	4 (1.16%)	6 (6,0)	1	42.03%	42.03%
				2	57.97%	on value 2
cmc	1473	115 (7.81%)	9 (2,7)	1	42.70%	57.30%
				2	22.61%	on value 1
				3	34.69%	
pima	769	1 (0.13%)	8 (8,0)	0	65.02%	34.98%
				1	34.98%	on value 0
breast	699	8 (1.15%)	9 (9,0)	2	65.52%	34.48%
				4	34.48%	on value 2

Table 1: Data sets summary descriptions.

Initially, the original data set was partitioned into 10 pairs of training and test sets through the application of 10-fold cross validation resampling method. Then, missing values were inserted into the training set. Six copies of this training set were used, two were given directly to C4.5 and CN2 without any missing data treatment. Other two copies had their missing values treated by the mean or mode imputation method and, the last two copies were given to the k-nearest neighbour to estimate and substitute the missing values. After the missing data treatment, the training sets were given to C4.5 and CN2. All classifiers, *i.e.* the two induced with untreated data and the other four induced with treated data, were used to classify the test set. At the end of 10 iterations, the true error rate was estimated by calculating the mean of the error rates of each iteration. Finally, the performances of C4.5 and CN2 allied to the k-nearest neighbour missing data treatment method were analysed and compared to the performances of the methods used internally by C4.5 and CN2 to learn when missing values are present, and to the performances of C4.5 and CN2 allied to the mean or mode imputation.

In order to insert missing data into the training sets, some attributes have to be chosen, and some of their values modified to unknown. Which attributes will be chosen and how many of their values will be modified to unknown is an important decision. It is straightforward to see that the most representative attributes of the data set are a sensible choice for the attributes that should have their values modified to unknown. Otherwise, the analysis may be compromised by treating non-representative attributes that will not be incorporated into the classifier by the learning system. Since finding the most representative attributes of a data set is not a trivial task, we used the results of [9] to select the three most relevant attributes according to several feature subset selection methods such as wrapper and filter.

Related to the amount of missing data to be inserted into the training sets, we want to analyse the behaviour of the methods with different amounts of missing data. In this way, missing data was inserted completely at random (MCAR) in the following percentages: 10%, 20%, 30%, 40%, 50% and 60% of the total of instances. The experiments were performed with missing data inserted into one, two and three of the

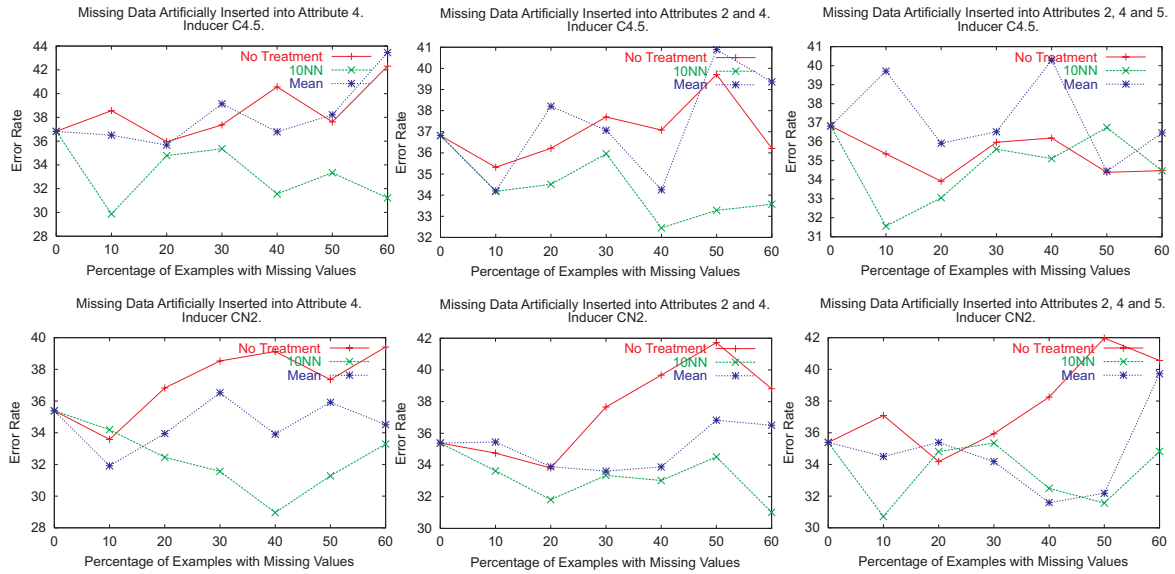

Figure 1: Comparative results for Bupa data set.

attributes selected as the most representatives. The missing values were replaced by estimated values using 1, 3, 5, 10, 20, 30, 50 and 100 nearest neighbours. Unfortunately, due to lack of space, only results with 10-nearest neighbour, identified as 10-NNI, will be showed in this work.

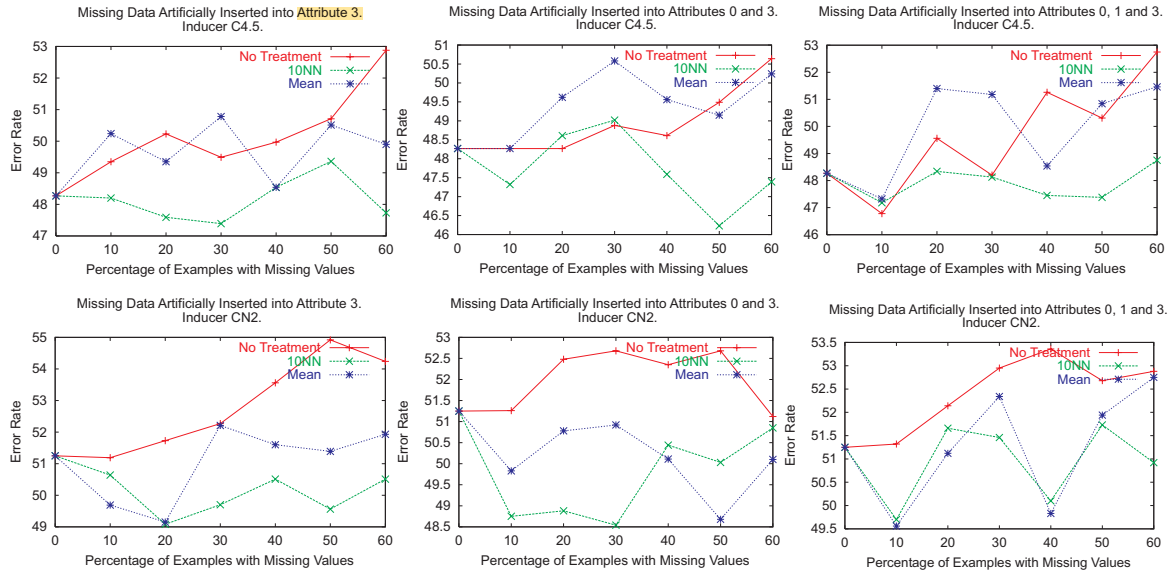

Figure 2: Comparative results for Cmc data set.

Considering the results shown in Figure 1, it can be observed that the performance of 10-NNI is superior to the performances of C4.5 and CN2 internal algorithms, and the mean imputation for Bupa data set. Furthermore, the C4.5 internal algorithm is competitive to 10-NNI only when missing values were inserted into the attributes 2, 4 and 5. The mean or mode imputation obtained good results when missing values are inserted into the attributes 2, 4 and 5, for the CN2 inducer.

Similar results are shown in Figure 2. The performance of 10-NNI is in most cases

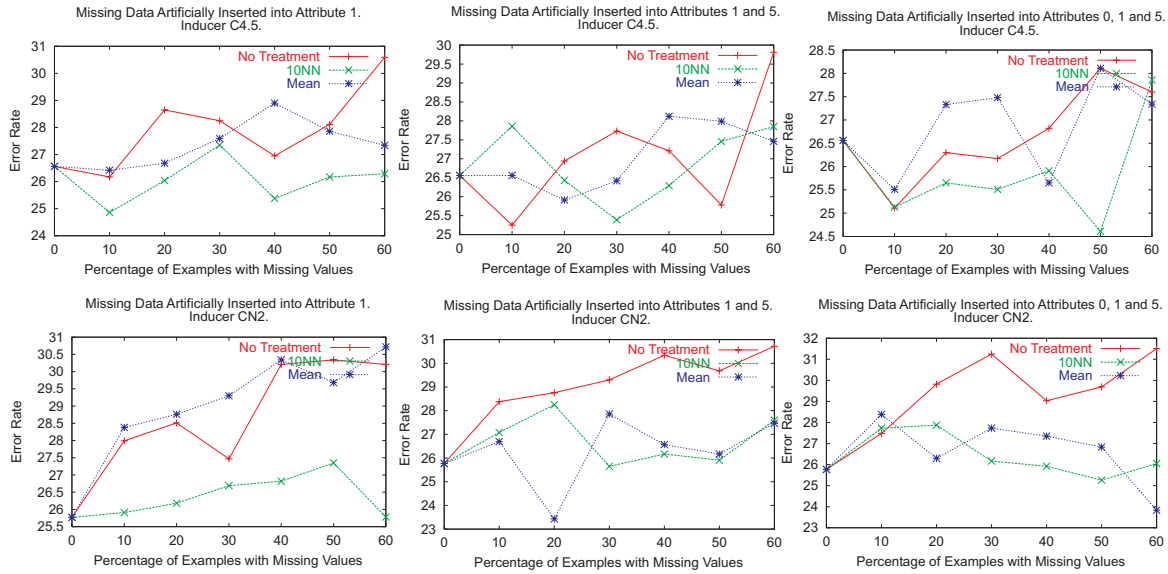

Figure 3: Comparative results for the Pima data set.

superior to the performance obtained without missing data treatment, for both C4.5 and CN2. The performance of 10-NNI is also superior or, in some few cases, competitive to the performance of the mean or mode imputation method. In fact, the mean or mode imputation method is competitive to 10-NNI only when missing values were inserted into the attributes 0 and 3 and 0, 1, and 3, using CN2 as inducer.

Figure 3 shows the comparative results for Pima data set. In this data set, the 10-NNI method shows a slightly superior performance compared with C4.5 without missing data treatment, and a superior performance compared with CN2 without missing data treatment. Besides, 10-NNI is superior to the mean or mode imputation when missing data were inserted into attribute 1 for both inducers. With missing data inserted into more than one attribute, 10-NNI and mean or mode imputation show similar results.

Table 2 shows some numerical results related to the graphs presented in Figures 1, 2 and 3. This table shows the error rates and standard deviations. More detailed results can be found in [1].

			C4.5			CN2		
Data set	Attr.	%?	No Imputation	Mean/Mode	10-NNI	No Imputation	Mean/Mode	10-NNI
Bupa	4	0%	36.82 ± 2.69	-	-	35.39 ± 2.47	-	-
		10%	38.56 ± 1.74	36.50 ± 1.76	29.87 ± 1.76	33.58 ± 1.94	31.91 ± 1.88	34.19 ± 1.45
		20%	35.95 ± 1.24	35.66 ± 1.61	34.78 ± 2.43	36.82 ± 0.96	33.95 ± 1.70	32.45 ± 0.95
		30%	37.36 ± 1.89	39.14 ± 2.41	35.36 ± 2.71	38.53 ± 2.16	36.52 ± 1.74	31.56 ± 2.71
		40%	40.56 ± 2.05	36.78 ± 1.72	31.55 ± 1.86	39.13 ± 1.09	33.91 ± 1.36	28.96 ± 2.24
		50%	37.62 ± 2.35	38.22 ± 3.03	33.34 ± 2.54	37.35 ± 2.74	35.92 ± 2.09	31.28 ± 1.91
		60%	42.31 ± 2.11	43.45 ± 2.08	31.22 ± 3.33	39.41 ± 1.20	34.51 ± 2.78	33.29 ± 2.64
	4, 2	0%	36.82 ± 2.69	-	-	35.39 ± 2.47	-	-
		10%	35.32 ± 2.36	34.20 ± 2.23	34.18 ± 1.72	34.75 ± 2.01	35.45 ± 2.21	33.63 ± 1.77
		20%	36.22 ± 2.18	38.21 ± 2.54	34.51 ± 2.16	33.81 ± 3.23	33.89 ± 1.49	31.81 ± 2.65
		30%	37.70 ± 2.40	37.07 ± 2.44	35.96 ± 2.05	37.66 ± 1.48	33.61 ± 1.96	33.34 ± 1.88
		40%	37.08 ± 1.42	34.25 ± 1.76	32.45 ± 1.09	39.67 ± 1.98	33.88 ± 1.27	33.02 ± 2.44
		50%	39.71 ± 2.76	40.89 ± 2.31	33.28 ± 3.07	41.72 ± 1.38	36.83 ± 1.88	34.51 ± 2.40
		60%	36.21 ± 1.84	39.36 ± 2.30	33.57 ± 2.38	38.81 ± 1.58	36.51 ± 2.33	31.01 ± 1.48
	4, 2, 5	0%	36.82 ± 2.69	-	-	35.39 ± 2.47	-	-
		10%	35.36 ± 1.76	39.71 ± 1.91	31.56 ± 2.44	37.09 ± 2.55	34.50 ± 1.81	30.71 ± 2.47
		20%	33.92 ± 2.07	35.92 ± 1.17	33.05 ± 2.09	34.18 ± 2.03	35.39 ± 1.75	34.81 ± 1.49
		30%	35.97 ± 2.90	36.52 ± 1.68	35.61 ± 3.00	35.94 ± 2.14	34.18 ± 1.92	35.35 ± 1.39
		40%	36.19 ± 2.39	40.29 ± 2.47	35.11 ± 2.14	38.25 ± 1.49	31.59 ± 2.51	32.49 ± 1.20

		50%	34.39 ± 2.84	34.45 ± 1.75	36.75 ± 2.12	41.97 ± 1.58	32.18 ± 2.24	31.56 ± 1.58
		60%	34.48 ± 1.77	36.46 ± 1.71	34.47 ± 3.02	40.56 ± 1.88	39.72 ± 1.63	34.82 ± 2.04
Cmc	3	0%	48.27 ± 0.83	-	-	51.25 ± 0.80	-	-
		10%	49.35 ± 1.14	50.24 ± 1.15	48.20 ± 1.16	51.19 ± 1.51	49.69 ± 1.34	50.64 ± 1.22
		20%	50.23 ± 1.12	49.35 ± 0.85	47.59 ± 0.98	51.73 ± 1.17	49.15 ± 1.42	49.08 ± 0.95
		30%	49.49 ± 0.95	50.78 ± 1.45	47.39 ± 1.48	52.27 ± 0.94	52.21 ± 1.13	49.70 ± 1.71
		40%	49.97 ± 0.87	48.54 ± 1.46	48.54 ± 1.12	53.56 ± 1.47	51.60 ± 0.73	50.51 ± 1.11
		50%	50.71 ± 1.11	50.51 ± 1.15	49.36 ± 0.91	54.92 ± 0.95	51.39 ± 1.39	49.56 ± 1.74
		60%	52.88 ± 1.25	49.90 ± 1.07	47.73 ± 0.95	54.24 ± 1.31	51.93 ± 1.50	50.51 ± 1.12
	3, 0	0%	48.27 ± 0.83	-	-	51.25 ± 0.80	-	-
		10%	48.27 ± 0.67	48.27 ± 1.37	47.32 ± 1.30	51.26 ± 0.80	49.83 ± 0.77	48.75 ± 1.42
		20%	48.27 ± 0.99	49.62 ± 1.42	48.61 ± 1.30	52.48 ± 1.51	50.78 ± 1.20	48.88 ± 1.46
		30%	48.88 ± 1.40	50.58 ± 0.98	49.02 ± 1.36	52.68 ± 0.91	50.92 ± 0.95	48.54 ± 1.34
		40%	48.61 ± 1.20	49.56 ± 1.33	47.59 ± 1.53	52.35 ± 1.10	50.11 ± 1.43	50.44 ± 1.09
		50%	49.49 ± 0.84	49.15 ± 1.38	46.23 ± 1.06	52.68 ± 0.81	48.68 ± 1.18	50.03 ± 1.76
		60%	50.64 ± 1.16	50.24 ± 0.91	47.39 ± 1.87	51.12 ± 1.53	50.10 ± 1.38	50.85 ± 1.49
	3, 0, 1	0%	48.27 ± 0.83	-	-	51.25 ± 0.80	-	-
		10%	46.78 ± 1.46	47.32 ± 0.78	47.18 ± 1.19	51.32 ± 1.19	49.56 ± 1.46	49.70 ± 1.61
		20%	49.56 ± 1.34	51.40 ± 1.49	48.34 ± 1.29	52.14 ± 1.04	51.12 ± 1.07	51.66 ± 1.06
		30%	48.20 ± 1.19	51.18 ± 0.89	48.13 ± 1.51	52.95 ± 1.25	52.34 ± 1.45	51.46 ± 1.15
		40%	51.26 ± 1.33	48.54 ± 1.12	47.45 ± 1.46	53.36 ± 1.23	49.83 ± 0.85	50.10 ± 1.49
		50%	50.31 ± 1.23	50.84 ± 1.61	47.38 ± 1.74	52.68 ± 1.02	51.94 ± 1.29	51.73 ± 1.82
		60%	52.75 ± 1.16	51.46 ± 1.06	48.75 ± 1.86	52.88 ± 0.76	52.75 ± 1.05	50.92 ± 1.35
Pima	1	0%	26.56 ± 1.16	-	-	25.77 ± 1.12	-	-
		10%	26.17 ± 1.03	26.42 ± 1.48	24.86 ± 0.88	27.99 ± 0.98	28.38 ± 0.87	25.91 ± 0.86
		20%	28.65 ± 1.15	26.68 ± 1.18	26.04 ± 1.68	28.51 ± 1.06	28.76 ± 1.51	26.18 ± 0.78
		30%	28.25 ± 1.85	27.59 ± 1.38	27.35 ± 1.03	27.47 ± 1.11	29.30 ± 1.23	26.69 ± 1.61
		40%	26.95 ± 1.67	28.90 ± 1.23	25.38 ± 1.15	30.21 ± 1.08	30.34 ± 1.59	26.82 ± 0.98
		50%	28.11 ± 1.14	27.86 ± 0.84	26.17 ± 1.11	30.34 ± 1.21	29.68 ± 1.58	27.35 ± 1.47
		60%	30.59 ± 1.13	27.34 ± 1.05	26.29 ± 1.90	30.21 ± 1.28	30.72 ± 1.47	25.78 ± 1.33
	1, 5	0%	26.56 ± 1.16	-	-	25.77 ± 1.12	-	-
		10%	25.25 ± 1.10	26.56 ± 1.08	27.86 ± 1.15	28.38 ± 0.87	26.69 ± 1.31	27.08 ± 0.98
		20%	26.94 ± 1.22	25.91 ± 1.34	26.43 ± 1.08	28.76 ± 1.51	23.43 ± 0.68	28.25 ± 1.09
		30%	27.73 ± 1.60	26.42 ± 1.27	25.39 ± 0.81	29.30 ± 1.23	27.86 ± 1.16	25.65 ± 1.13
		40%	27.21 ± 1.45	28.12 ± 1.11	26.29 ± 1.69	30.34 ± 1.59	26.57 ± 1.73	26.17 ± 1.07
		50%	25.78 ± 1.13	27.99 ± 1.37	27.46 ± 1.16	29.68 ± 1.58	26.17 ± 0.82	25.91 ± 1.08
		60%	29.81 ± 1.43	27.46 ± 1.67	27.85 ± 1.51	30.72 ± 1.47	27.47 ± 0.75	27.60 ± 1.47
	1, 5, 0	0%	26.56 ± 1.16	-	-	25.77 ± 1.12	-	-
		10%	25.11 ± 1.70	25.51 ± 1.90	25.13 ± 0.90	27.48 ± 1.00	28.38 ± 0.99	27.73 ± 0.68
		20%	26.30 ± 1.01	27.33 ± 1.42	25.65 ± 1.35	29.82 ± 0.82	26.30 ± 1.13	27.87 ± 1.26
		30%	26.17 ± 1.35	27.48 ± 1.19	25.51 ± 1.75	31.25 ± 0.89	27.73 ± 0.91	26.17 ± 1.32
		40%	26.82 ± 1.28	25.65 ± 0.84	25.91 ± 1.44	29.03 ± 0.90	27.35 ± 0.92	25.92 ± 1.32
		50%	28.11 ± 1.32	28.11 ± 1.65	24.61 ± 1.16	29.69 ± 0.41	26.83 ± 1.29	25.26 ± 0.68
		60%	27.60 ± 1.05	27.34 ± 1.53	27.86 ± 1.55	31.51 ± 1.17	23.83 ± 0.95	26.05 ± 0.86
Breast	1	0%	4.24 ± 0.67	-	-	4.68 ± 0.60	-	-
		10%	3.80 ± 0.93	3.66 ± 0.82	4.25 ± 0.67	4.39 ± 0.44	4.24 ± 0.46	5.12 ± 0.84
		20%	3.95 ± 0.90	3.51 ± 0.88	5.11 ± 0.99	4.68 ± 0.75	4.83 ± 0.69	4.39 ± 0.57
		30%	3.95 ± 0.90	3.80 ± 0.93	4.09 ± 0.91	4.97 ± 0.82	4.67 ± 1.03	4.97 ± 0.62
		40%	3.95 ± 0.90	3.95 ± 0.90	4.53 ± 0.82	4.53 ± 0.73	5.12 ± 0.90	4.53 ± 0.70
		50%	3.95 ± 0.90	3.95 ± 0.90	5.41 ± 1.00	4.53 ± 0.91	4.82 ± 0.87	4.53 ± 0.63
		60%	3.95 ± 0.90	3.95 ± 0.90	6.00 ± 0.88	4.83 ± 0.84	4.83 ± 1.07	5.12 ± 0.69
	1, 5	0%	4.24 ± 0.67	-	-	4.68 ± 0.60	-	-
		10%	4.83 ± 0.61	3.80 ± 0.85	4.10 ± 0.61	4.38 ± 0.65	3.80 ± 0.66	4.38 ± 0.75
		20%	4.97 ± 0.65	4.68 ± 0.64	3.80 ± 0.88	3.65 ± 0.84	4.53 ± 0.67	5.56 ± 0.77
		30%	4.68 ± 0.61	4.39 ± 0.65	4.83 ± 0.69	3.95 ± 0.54	4.09 ± 0.97	4.69 ± 0.57
		40%	4.39 ± 0.65	4.97 ± 0.44	4.98 ± 0.54	3.95 ± 0.87	3.51 ± 0.66	4.96 ± 0.99
		50%	4.98 ± 0.73	4.69 ± 0.37	3.81 ± 0.63	4.53 ± 0.63	4.68 ± 0.75	4.98 ± 0.76
		60%	4.54 ± 0.71	4.68 ± 0.65	5.85 ± 0.53	4.39 ± 0.95	3.66 ± 0.66	4.25 ± 0.64
	1, 5, 0	0%	4.24 ± 0.67	-	-	4.68 ± 0.60	-	-
		10%	4.68 ± 0.75	4.10 ± 0.61	4.83 ± 0.81	4.25 ± 0.71	4.10 ± 0.52	4.83 ± 0.76
		20%	5.12 ± 0.73	4.83 ± 0.69	4.69 ± 0.68	4.97 ± 0.79	4.39 ± 0.66	3.80 ± 0.62
		30%	5.42 ± 0.69	4.98 ± 0.50	4.69 ± 1.02	5.12 ± 0.54	5.41 ± 0.78	4.24 ± 0.80
		40%	4.97 ± 0.62	4.09 ± 0.68	5.27 ± 0.85	5.13 ± 0.55	3.65 ± 0.82	4.83 ± 0.62
		50%	5.41 ± 0.57	4.83 ± 0.61	4.10 ± 0.84	5.85 ± 0.76	3.07 ± 0.82	4.24 ± 0.91
		60%	4.97 ± 0.73	4.68 ± 0.78	4.68 ± 0.80	5.85 ± 0.61	3.80 ± 0.73	5.11 ± 0.97

Table 2: Comparative results for Bupa, Cmc, Pima and Breast data sets.

It is important to say that, for Bupa, Cmc and Pima data sets, the internal meth-

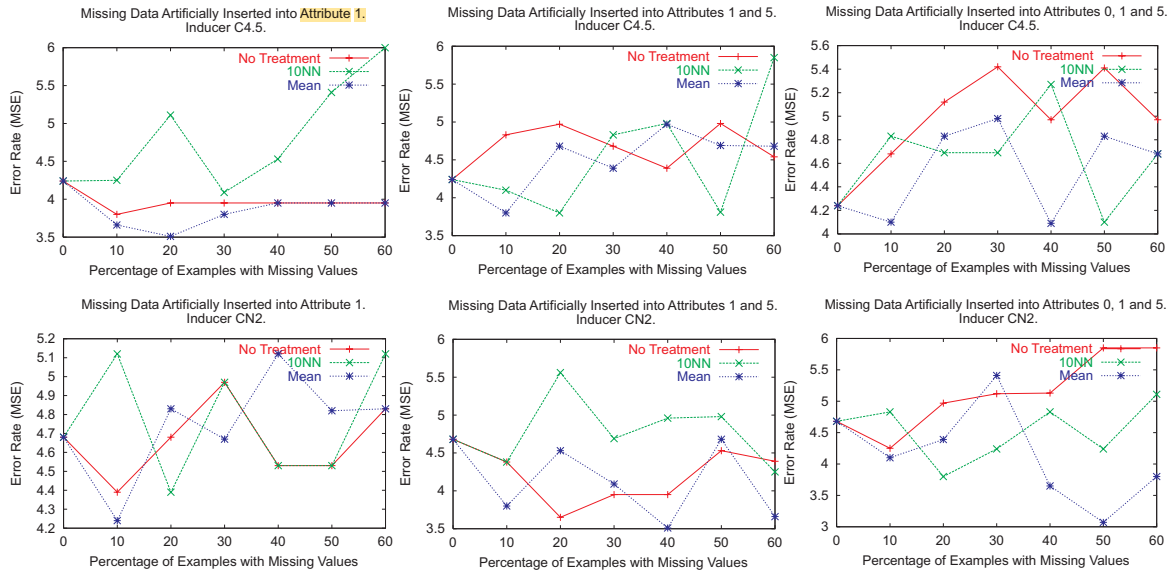

Figure 4: Comparative results for the Breast data set.

ods used by C4.5 and CN2 to treat missing data show a lower error rates compared to 10-NNI in only 11 of 108 measurements (8 for C4.5 and 3 for CN2). In none of these 11 measurements the internal methods show a statistically significant difference. On the other hand, 10-NNI shows statistically significant difference in 35 measurements (13 of them are highly significant). Comparing 10-NNI to the mean or mode imputation method, the mean or mode imputation shows a lower error rate in 20 of 108 measurements (5 for C4.5 and 15 for CN2), 1 of them is a highly significant difference. 10-NNI shows statistically significant differences in 11 measurements (3 of them are highly significant).

Although missing data imputation with k-nearest neighbour can provide good results, there are occasions that its use should be avoided. This is illustrated by the Breast data set. Breast was chosen because its attributes have strong correlations among each other. These correlations cause an interesting situation: in one hand, the k-nearest neighbour can predict the missing values with precision; on the other hand, the inducer can decide not to use the treated attribute, replacing it by another attribute with high correlation. The results for Breast data set are shown in Figure 4, where it can be seen that 10-NNI does not outperform the others missing data treatment methods. This scenario is interesting because 10-NNI was able to predict the missing data with higher precision than the mean or mode imputation. As missing values were artificially implanted into the data, the mean square error (*MSE*) between the predicted values and the actual ones can be measured. These errors are presented in Table 3.

Attribute	MSE 10-NNI	MSE Mean/Mode
0 (Clump Thickness)	4.02 ± 0.14	7.70 ± 0.28
1 (Uniformity of Cell Size)	1.72 ± 0.11	8.96 ± 0.36
5 (Bare Nuclei)	4.23 ± 0.30	13.29 ± 0.46

Table 3: Mean square error (MSE) between predicted and actual values for 10-NNI and mean or mode imputation — Breast data set.

If 10-NNI method was more accurate in predicting the missing values, why this

higher accuracy is not translated into a more precise classifier? The answer may be in the high correlation among the data set attributes and, because (or consequently) Breast data set has several attributes with similar predicting power.

In order to perform a deeper analysis, we need to verify how each attribute is used into the induced classifier. For instance, it is interesting to understand how C4.5 was able to obtain a constant error rate even with high levels of missing values inserted into attribute 1 (Figure 4). Analysing the decision trees generated by C4.5, it is possible to verify that C4.5 was able to substitute attribute 1 — Uniformity of Cell Size — by attribute 2 — Uniformity of Cell Shape. This substitution was possible because these two attributes have a high level of correlation (linear correlation coefficient $r = 0.9072$). In a general way, for Breast data set, C4.5 was able to replace every attribute with missing values by others attributes, and still be competitive with 10-NNI. Using the highest level of the decision tree in which the attribute was incorporated as a heuristical measure of attribute importance in the model, Table 4 shows that C4.5 was able to gradually discard the attributes with missing values as the amount of missing data increased. In a similar way, C4.5 shows a tendency to discard the attributes with missing values when those attributes were treated with mean or mode imputation. This result is expected since in mean or mode imputation all missing values are replaced by the same value (the attribute mean or mode). Consequently, the attribute discriminatory power (measured by the C4.5 decision tree algorithm through entropy) tends to decrease. The same did not occur when the missing data were treated by 10-NNI. In this scenario, C4.5 kept the attributes with missing values as the upmost attributes into the decision tree. This situation would had been an advantage if Breast data set do not have other attributes with similar predicting power.

% of Missing	No Imputation			Mean/Mode			10-NNI		
	Attr. 1	Attr. 5	Attr. 0	Attr. 1	Attr. 5	Attr. 0	Attr. 1	Attr. 5	Attr. 0
0%	1	2	3	1	2	3	1	2	3
10%	2	2	3	2	2	3	1	2	3
20%	-	2	3	-	3	3	1	2	3
30%	-	5	-	-	3	-	1	2	3
40%	5	4	-	3	-	-	1	2	3
50%	-	-	-	6	7	3	1	3	2
60%	-	5	-	-	3	-	1	2	3

Table 4: Level in which the attributes 1, 5 and 0 of Breast data set were incorporated into the decision tree induced by C4.5. "-" means that the attribute was not selected to be part of the tree. Level 1 is the root of the decision tree.

7 Conclusions and Limitations

This work analyses the behaviour of four methods for missing data treatment: the 10-NNI method using a k-nearest neighbour algorithm for missing data imputation; the mean or mode imputation; and the internal algorithms used by C4.5 and CN2 to treat missing data. These methods were analysed inserting different percentages of missing data into different attributes of four data sets, showing promising results. The 10-NNI method provides very good results, even for training sets having a large amount of missing data.

The Breast data set provided a valuable insight into the limitations of the missing data treatment methods. The first decision to be taken is if the attribute should be

treated. The existence of others attributes with similar information (high correlation), or similar predicting power can make the missing data imputation useless, or even harmful. Missing data imputation can be harmful because even the most advanced imputation method is only able to approximate the actual (missing) value. The predicted values are usually more well-behaved, since they conform with other attributes values. In the experiments carried out, as more attributes with missing values were inserted and as the amount of missing data increased, more simple were the induced models. In this way, missing data imputation should be carefully applied, under the risk of oversimplifying the problem under study.

In future works, the missing data treatment methods will be analyzed in other data sets. Furthermore, in this work missing values were inserted completely at random (MCAR). In a future work, we will analyze the behaviour of these methods when missing values are not randomly distributed. In this case, there is a possibility of creating invalid knowledge. For an effective analysis, we will have to inspect not only the error rate, but also the quality of the knowledge induced by the learning system.

Acknowledgements. This research is partially supported by Brazilian Research Councils CAPES and FINEP. The authors would like to thank André C.P.L.F. de Carvalho for his suggestions.

References

[1] G. E. A. P. A. Batista and M. C. Monard. K-Nearest Neighbour as Imputation Method: Experimental Results (in print). Technical report, ICMC-USP, 2002. ISSN-0103-2569.

[2] P. Ciaccia, M. Patella, and P. Zezula. M-tree: An Efficient Access Method for Similarity Search in Metric Spaces. In *VLDB'97*, pages 426–435, 1997.

[3] P. Clark and T. Niblett. The CN2 Induction Algorithm. *Machine Learning*, 3(4):261–283, 1989.

[4] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm (with Discussion). *Journal of Royal Statistical Society*, B39:1–38, 1977.

[5] J. W. Grzymala-Busse and M. Hu. A Comparison of Several Approaches to Missing Attribute Values in Data Mining. In *RSCTC'2000*, pages 340–347, 2000.

[6] C. Traina Jr., A. Traina, B. Seeger, and C. Faloutsos. Slim-trees: High Performance Metric Trees Minimizing Overlap Between Nodes. In *EDBT'2000*, pages 51–65, 2000.

[7] R. Kohavi, D. Sommerfield, and J. Dougherty. Data Mining using MLC++: A Machine Learning Library in C++. *Tools with Artificial Intelligence*, pages 234–245, 1996.

[8] K. Lakshminarayan, S. A. Harp, and T. Samad. Imputation of Missing Data in Industrial Databases. *Applied Intelligence*, 11:259–275, 1999.

[9] H. D. Lee, M. C. Monard, and J. A. Baranauskas. Empirical Comparison of Wrapper and Filter Approaches for Feature Subset Selection. Technical Report 94, ICMC-USP, 1999. ISSN-0103-2569.

[10] R. J. Little and D. B. Rubin. *Statistical Analysis with Missing Data*. John Wiley and Sons, New York, 1987.

[11] C. J. Merz and P. M. Murphy. UCI Repository of Machine Learning Datasets, 1998. http://www.ics.uci.edu/ mlearn/MLRepository.html.

[12] J. R. Quinlan. *C4.5 Programs for Machine Learning*. Morgan Kaufmann, CA, 1988.

[13] D. R. Wilson and T. R. Martinez. Reduction Techniques for Exemplar-Based Learning Algorithms. *Machine Learning*, 38(3):257–286, March 2000.