A Review of Missing Data Treatment Methods

Liu Peng, Lei Lei
Department of Information Systems, Shanghai University of Finance and Economics,
Shanghai, 200433, P.R. China

ABSTRACT

Missing data is a common problem for data quality. Most real datasets have missing data. This paper analyzes the missing data mechanisms and treatment rules. Popular and conventional missing data treatment methods are introduced and compared. Suitable environments for method are analyzed in experiments. Methods are classified into certain categories according to different characters.

Key words: Missing data, Data mining, Knowledge Discovery in Databases

1. Introduction

Data Mining (DM) is the process of discovering interesting knowledge from large amounts of data stored either in databases; data warehouse; or other information repositories [1]. According to the study of Cabena, about 20% of the effort is spent on the problem and data understanding, about 60% on data preparation and about 20% on data mining and analysis of knowledge [2]. Why do people spend so much time on data preparation? Actually, there are a lot of serious data quality problems in real datasets: incomplete, redundant, inconsistent and noisy. These serious quality problems reduce the performance of data mining algorithms. Missing data is a common issue in almost every real dataset. If the rate of missing is less than 1%, missing data won't make any trouble for the Knowledge Discovery in Databases (KDD) process, 1-5% manageable, 5-15% requires sophisticated methods to handle and more than 15% may severely impact any kind of interpretation [3].

The paper emphasizes on the treatment methods of missing data. Missing mechanism and the guidelines for treatment are presented in section two. Section three introduces some popular treatment methods of missing data. Section four is experimental analysis and comparison. Characters and suitable environments for each method are analyzed and classified in section five. The last section is the conclusion of our job.

2. Missing Mechanism and Treatment Rules

The effect of the missing data treatment methods mainly depends on missing mechanism. For instance, missing data can be analyzed or surmised by the source information if we know. But if we don't know, the missing data treatment methods are supposed to be independent of how the missing data come into being. Statistician divides missing data into the following three categories [4]: (1) Missing completely at random (MCAR). It is the highest level of randomness. The probability of missing data on any attribute does not depend on any value of attribute. (2) Missing at random (MAR). The probability of missing data on any attribute does not depend on its own particular value, but on the values of other attributes. (3) Not missing at random (NMAR). Missing data depends on the values that are missing.

The treatment of missing values is an important task in KDD process. Especially, while the dataset contains a large amount of missing data, the treatment of missing data can improve the quality of KDD dramatically. Some data mining approaches treat missing data with internal algorithms, say decision tree C4.5. But it is still significant to construct complete datasets with treatment methods for missing data: (1) Data collectors with the knowledge about missing mechanism are able to construct a complete dataset which is very close to the real one. (2) All the data mining approaches can be used, if the dataset is a complete one. (3) It can prove a basic point for the comparison of the data mining approaches.

However, the data distribution should not be changed while handling missing data. Any missing data treatment method should satisfy the following rules: (1) Estimation without bias. Any missing data treatment method should not change the data distribution. (2) Relationship. The relationship among the attributes should be retained. (3) Cost. It is difficult for the method which is too complex and time cost to practice in real life.

3. Missing Data Treatment Methods

In this section we introduce some popular missing data treatment methods and our proposed models which are based on the concept of information gain and Naive Bayesian Classifier.

3.1. Case Deletion (CD)

This method omits those cases (instances) with missing data and does analysis on the remains. Though it is the most common method, it has two obvious disadvantages: a) A substantial decrease in the size of dataset available for the analysis. b) Data are not always missing completely at random. This method will bias the data distribution and statistical analysis. A variation of this method is to delete the cases (or attributes) with high missing rate. But before deleting any attribute, it is necessary to run relevance analysis, especially on the attributes with high levels of missing data.

3.2. Maximization Likelihood Methods (ML)

ML use all data observed in a database to construct the best possible first and second order moment estimates. It does not impute any data, but rather a vector of means and a covariance matrix among the variables in a database. This method is a development of expectation maximization (EM) approach. One advantage is that it has well-know statistical foundation. Disadvantages include the assumption of original data distribution and the assumption of incomplete missing at random.

3.3. Mean/Mode Imputation (MMI)

Replace a missing data with the mean (numeric attribute) or mode (nominal attribute) of all cases observed. To reduce the influence of exceptional data, median can also be used. This is one of the most common used methods. But there are some problems. Using constants to replace missing data will change the characteristic of the original dataset; ignoring the relationship among attributes will bias the following data mining algorithms. A variation of this method is to replace the missing data for a given attribute by the mean or mode of all known values of that attribute in the class where the instance with missing data belongs [5].

3.4. All Possible Values Imputation (APV)

It consists of replacing the missing data for a given attribute by all possible values of that attribute. In this method, an instance with a missing data will be replaced by a set of new instances. If there are more than one attribute with missing data, the substitution for one attribute will be done first, then the nest attribute be done, etc., until all attributes with missing data are replaced. This method also has a variation. All possible values of the attribute in the class are used to replace the missing data of the given attribute. That is restricting the method to the class [6].

3.5. Regression Methods (RM)

Regression imputation assumes that the value of one variable changes in some linear way with other variables. The missing data are replaced by a linear regression function instead of replacing all missing data with a statistics. This method depends on the assumption of linear relationship between attributes. But in the most case, the relationship is not linear. Predict the missing data in a linear way will bias the model.

3.6. Hot (cold) deck imputation (HDI)

In this method, a missing attribute value is filled in with a value from an estimated distribution for the missing value from the current data [3]. It is typically implemented in two stages: a) Data are partitioned into clusters. b) Missing data are replaced within a cluster. This can be done

by calculating the mean or mode of the attribute within a cluster. In Random Hot deck, a missing value of attribute is replaced by an observed value of the attribute chosen randomly. Cold deck imputation is similar to hot deck but the data source must be other than the current data source [3].

3.7. K-Nearest Neighbor Imputation (KNN)

This method uses k-nearest neighbor algorithms to estimate and replace missing data. The main advantages of this method are that: a) it can estimate both qualitative attributes (the most frequent value among the k nearest neighbors) and quantitative attributes (the mean of the k nearest neighbors); b) It is not necessary to build a predictive model for each attribute with missing data, even does not build visible models. Efficiency is the biggest trouble for this method. While the k-nearest neighbor algorithms look for the most similar instances, the whole dataset should be searched. However, the dataset is usually very huge for searching. On the other hand, how to select the value “k” and the measure of similar will impact the result greatly.

3.8. Multiple Imputation (MI)

The basic idea of MI is that: a) a model which incorporates random variation is used to impute missing data; b) do this M times, producing M complete datasets; c) run the analysis on each complete dataset and average the results of M cases to produce a single one. For MI, the data must be missing at random. In general, multiple methods outperform deterministic imputation methods. They can introduce random error in the imputation process and get approximately unbiased estimates of all parameters [5]. But the cost of calculating is too high for this method to implement in practice.

3.9. Internal treatment method for C4.5 (C4.5)

Decision tree C4.5 is a widely accepted classifier. One of the improvements for C4.5 is the development of the internal algorithms for missing data treatment. It uses probability approaches to handle missing data [4]: a) Select attribute according to the correctional information gain ratio and the correctional gene depends on the proportion of missing data on the attribute. b) All the instances with missing data are distributed into all the subsets according to the probability and the probability depends on the size of the subset they belonged to. c) While the decision tree is used to classify the new instance, all the possible paths are searched and then give a classification result in the form of probability, if the instance have missing data on the training attribute.

3.10. Bayesian Iteration Imputation (BII)

Naive Bayesian Classifier is a popular classifier, not only for its good performance, but also for its simple form. It is not sensitive to missing data and the efficiency of calculation is very high.

Bayesian Iteration Imputation uses Naive Bayesian Classifier to impute the missing data. It is consisted of two phases: a) Decide the order of the attribute to be treated according to some measurements such as information gain, missing rate, weighted index, etc.; b) Using the Naive Bayesian Classifier to estimate missing data. It is an iterative and repeating process. The algorithms replace missing data in the first attribute defined in phase one, and then turn to the next attribute on the base of those attributes which have be filled in. Generally, it is not necessary to replace all the missing data (usually 3~4 attributes) and the times for iterative can be reduced [7].

4. Experimental Analysis

There are a lot of algorithms dealing with missing data developed in recent years. The basic approaches about these popular algorithms have been introduced in Section 3. In this section, four experiments about missing data will be introduced. All the datasets used in experiments come from the Machine Learning Database Repository at the University of California, Irvine.

Experiment 1 comes from literature [6]. This experiment was carried out with dataset Breast cancer, Echocardiogram, Hdynet, Hepatitis, House, Im85, New-o, Primary tumor, Soybean and Tokt to evaluate the effect on the misclassification error rate of 9 methods for dealing with missing data: MMI, Concept Mean/Mode Imputation (CMMI), C4.5, APV, Concept All Possible Values Imputation (CAPV), CD, Event-Covering Method (EC), A Special LEM2 Algorithm (LEM2), New Values (NV). The experiments were conducted as follows. All of the nine were applied to ten datasets, which had been sampled into ten pairs of training and testing subsets. Then, new LERS was used to generate classification rules and classify the samples in testing subsets. The performance of different methods was compared by calculating the average error rate. For classifier new LERS, C4.5 is better than MMI, CD is better than MMI, LEM2 and NV. These experiments conclude that C4.5 approach is the best, CD is next to and MMI is the worst method among all nine approaches. APV and CAPV are excellent approaches based on their limited experimental results. However evidence is not enough to support this claim and further study is need.

Experiment 2 comes from literature [4].In this experiment, 4 datasets: Bupa, Cmc, Pima, Breast, were used to investigate the performance of the 4 methods to deal with missing data: MMI, KNN, C4.5 and CN2. Firstly, every original dataset are partitioned into 10 pairs of training and test subsets. A given percentage, varying from 10% to 60%, of missing data is artificially inserted into one, two or three attributes of the training subsets. All the missing data were treated by the four methods and then, two classifiers, C4.5 and CN2, were used to classify the test subsets. The average error rate of 10 iterations are estimated. The analysis indicates that KNN can outperform the internal methods used by C4.5 and CN2 to treat missing data, and can also outperform the MMI. Furthermore, the C4.5 is competitive to KNN when the number of attributes with missing data increasing. C4.5 tends to discard the attributes with missing data when those attributes were treated with mean or mode imputation or as the amount of missing data increased.

Experiment 3 comes from literature [3]. The experiments were carried using twelve datasets:

Hepatitis, Bupa, Breast, Iris, Sonar, Heartc, Ionosphere, Crx, Diabetes, Vehicle, German, Segment, to evaluate the performance of four missing data treatment methods: CD, MMI, MDI and KNN. Initially, missing data is inserted into each dataset completely at random in the percentages from 1% to 20%. Then, four methods are applied to treat the missing data and 10-fold cross-validation estimates of the misclassification error rate for both the LDA and KNN are calculated. The analysis indicates that there is not much difference between the MMI and MDI. Overall, KNN seems to perform better than CD, MMI and MDI because it is most robust to bias when the percentage of missing data increases.

Experiment 4 was carried out in this paper. Three datasets, dataset Crx, German and Nursery, were used to investigate the performance of three methods to deal with missing data: MMI, C4.5 and BII. Initially, the original dataset was partitioned into 10 pairs of training and testing subsets. Then, missing data were artificially inserted into one, two or three attributes of the training subset randomly in the percentages from 10% to 60%. Three methods were applied to treat the missing data and average misclassification error rate for C4.5 of the ten iterations are calculated. The results of dataset Nursery are displayed in TABLE 1 and FIGURE 1. On the whole, the performance of BII is superior to the performances of C4.5 and MMI. The type of the attributes with missing data affects the results of the methods. While the important attribute for classifying contains fewer missing data or none, C4.5 internal model performs very well. Comparatively, in the case of larger missing proportion and more attributes with missing data, BII will perform more satisfactorily.

TABLE 1. Comparative Results for Dataset Nursery

%?	Attr.	C4.5	MMI	BII	Attr.	C4.5	MMI	BII
0%		3.84±0.33%	-	-		3.84±0.33%	-	-
10%		4.23±0.17%	4.24±0.26%	3.91±0.11%		4.48±0.05%	4.54±0.31%	4.18±0.59%
20%	8 heal	4.71±0.30%	5.30±0.55%	5.02±0.62%	8 heal	5.07±0.44%	5.50±0.39%	5.31±0.22%
30%	1 par	5.74±1.02%	7.01±1.90%	6.41±0.48%	1 par	6.94±0.76%	6.78±0.27%	6.60±0.02%
40%		7.37±1.00%	7.04±0.53%	7.06±0.42%	6 fina	8.13±0.57%	7.41±0.17%	7.51±0.75%
50%		10.75±0.10%	10.45±0.98%	8.79±1.02%		10.98±0.04%	10.12±0.31%	10.09±0.42%
60%		11.04±0.34%	13.60±6.27%	9.77±1.18%		11.82±0.38%	15.39±6.91%	11.30±0.87%

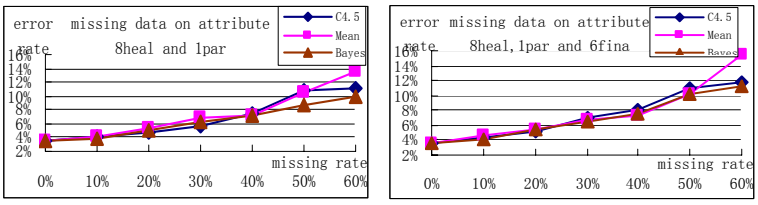

FIGURE 1. Comparative Results for Dataset Nursery

Summarily, MMI will be a better method for nominal data and KNN will be a better method for numeric. KNN, C4.5 and MMI are most common used methods for dealing with missing data these days. Though, they can perform very well, there are still drawbacks left to improve. Meanwhile, new developed methods, say Multiple Imputation, Hot deck imputation, become more competitive.

5. Classification of Missing Data Treatment Methods

In general, treatment methods for missing data can be divided into three kinds [4]: 1) Case deletion. It is the simplest one. 2) Parameter estimation. In this way, variants of Expectation-Maximization algorithm are used in the maximum likelihood procedures to estimate the parameters for missing data. Thanks to the full use of all the observed data, especially while the probability leading to missing data is known in the model, this kind of methods usually superior to case deletion. However, there are still some restrictions on the use of these methods. For example, the assumption of variable distributions, the high degree of complexity for calculation. 3) Imputation techniques, which uses the present information of the dataset to estimate and replace missing data correspondingly. It aims to recognize the relationships among the values in dataset and to estimate missing data under the help of these relationships. One advantage of this approach is that the treatment of missing data is independent of the learning algorithms used. This allows the user to select the most appropriate imputation method for each situation. But it depends on the assumption of the relationships existing among attributes. Values estimated in this way are usually more well-behaved than the true values would be, say, the predicted values are likely to be more consistent with this set of attributes than the true values would be [4]. Imputation methods can also be divided into three groups [5]: 1) Global imputation based on missing attribute. These methods use known values to impute the missing data of the attribute. 2) Global imputation based on non-missing attribute. These methods use the relationships among missing and non-missing attributes to predict missing data. 3) Local imputation. These methods subdivided samples into clusters and estimated missing data in cluster.

According to the phase of handling missing data in KDD process, methods can be classified into two groups: pre-replacing methods and embedded methods [8]. Pre-replacing methods are special methods which deal with missing data in data preparation phase of KDD process. Embedded methods deal with missing data in data mining phase of KDD process. The former one can be applied more flexibly and the later one can save more time and cost [4]. C4.5 is a typical method which contains embedded methods for dealing with missing data. According to the basic approach, methods can be classified into two groups: statistical methods and machine learning methods. In general, statistical methods are much simpler and machine learning methods have a higher accuracy, but more time costing. According to the available kinds of attributes, methods can be classified into three groups: numerical, nominal and both. Appropriate method should be selected according to different kind of attributes. According to the times of treatment, methods can be classified into two groups: deterministic methods and multiple methods. Deterministic methods only impute one value for replacing, which does not represent the uncertainty of imputed values. Multiple methods solve this problem by imputing several values for every missing data. MI has several desirable features and developed to be one of the most popular methods. There are many methods for dealing missing data, but no one is absolutely better than the others. Different situations, different classifiers require different methods [5]. Methods introduced in Section 3 are summarized in TABLE 2 from three aspects: basic approach, computing cost and available kinds of attributes.

TABLE 2. Summary of methods for dealing with missing data

No.	Method	Approach	Cost	Attr.	No.	Method	Approach	Cost	Attr.
1	CD	---	Low	Num & Nom	6	HDI	ML	Low	Num & Nom
2	ML	Statistic	Low	Num	7	KNN	ML	High	Num
3	MMI	Statistic	Low	Num & Nom	8	MI	ML	High	Num & Nom
4	APV	Statistic	High	Nom	9	C4.5	ML	Middle	Num & Nom
5	RM	Statistic	Low	Num	10	BII	ML	Middle	Nom

6. Conclusions

The topic of missing data has received considerable attention in the last decade. More and more missing data treatment methods have sprouted-up. Mainly methods for dealing with missing data are compared in this paper. Initially, missing data mechanism and treatment rules are presented. Popular methods for dealing with missing data and four comparative experiments about the effect of the methods are introduced. Characters and suitable environments for each method are analyzed and compared in experiments. KNN, C4.5 and MMI are most common used methods for dealing with missing data these days. Then, methods are classified in different aspect, such as available kinds of attributes, treatment phase, treatment times and basic approach. There exist many methods for dealing missing data, but no one is absolutely better than the others. Different situations require different solutions.

Reference

1. Han J. and Kamber M., *DataMining Concepts and Techniques*, Morgan Kaufmann Publishers, 2000
2. Cios K.J. and Kurgan L., *Trends in Data Mining and Knowledge Discovery*. In N.R. Pal, L.C. Jain, and Teoderesku N., editors, *Knowledge Discovery in Advanced Information Systems*. Springer, 2002.
3. Acuna E. and Rodriguez C., *The treatment of missing values and its effect in the classifier accuracy*. In D. Banks, L. House, F.R. McMorris,P. Arabie, W. Gaul (Eds).Classification, Clustering and Data Mining Applications. Springer-Verlag Berlin-Heidelberg, 639-648. 2004.
4. Gustavo E. A. P. A. Batista and Maria Carolina Monard, *An Analysis of Four Missing Data Treatment Methods for Supervised Learning*, Applied Artificial Intelligence 17(5-6): 519-533 , 2003
5. Magnani M., *Techniques for Dealing with Missing Data in Knowledge Discovery Tasks*, department of computer Science, University of Bologna, 2004
6. 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.
7. Liu P., Lei L. and Zhang X.F., *A Comparison Study of Missing Value Processing Methods*, Computer Science, 31(10):155-156, 2004.
8. Yoshikazu Fujikawa, *Efficient Algorithms for Dealing with Missing values in Knowledge Discovery*, Master Degree Thesis, Japan Advanced Institute of Science and Technology, 2001.