A Study of Impact on Missing Categorical Data - A Qualitative Review

Vijayakumar Kuppusamy* and Ilango Paramasivam
School of Computer Science and Engineering, VIT University, Vellore - 632014, Tamil Nadu, India; kvijayakumar@vit.ac.in, pilango@vit.ac.in

Abstract

Objectives: In this study, MissForest algorithm is used to analyze the impact of missing data. Statistical Analysis: For categorical dataset, MissForest package is used to impute various missing values and tested with missing data incrementally, first with 5% missing attributes of the records from the original dataset then with 10% and so on. Findings: The Proportion of Falsely Classified (PFC) is computed for categorical dataset. Since the good performance of MissForest leads to a value close to zero and bad performance to a value around one, the performance measured for the Nursery dataset is quite good. Application: This approach has good effect when the ratio of missing data is low. Missforest algorithm can be used for numerical or categorical data. The results are improved for continuous data.

Keywords: Categorical Data, Data Mining, Imputation, Missing Data, MissForest

1. Introduction

Data Mining is an analytical process of exploring large quantities of data and discovering the existing patterns and trends between them. Currently, missing data are found by pre-processing the datasets which is required for decision making. This is one of biggest research issues in imputing the missing values and using the complete datasets for analysis. As we have loads of data through various web services such as e-commerce, bank or credit card transactions, online purchases etc., and there is a lot of difficulty in solving these issues. Various techniques and methods are used to address the issue of missing data.

Data mining involves searching the series of multiple databases containing some amount of missing data along with a inconsistent percentage of inaccurate data, pollution, outliers, and noise. Ignoring the issue of missing data can cause prejudice in model evaluation and cause inaccurate mining results and thus should be tackled.

It is very common situation in the data set to have missing information. In the real world applications there are many possibilities of missing information, like automated sensor errors or failures, optional data fields in medical files, or simply refusal by respondents to answer questions compromising their privacy in surveys. Assuming that the data is complete and accurate by the data mining and statistical method has led to very little attention being to address this issue.

A common practice is to place an approximate value from the missing variable or an imputed value attained from the non-missing values of other variables in the same unit. Different methods are available depending on the various types of values with missing data for imputation.

The objective of this paper is to address the different methods available to impute the missing data and its effect on the data mining process in various domains. For example, to reduce the mailing cost in direct marketing, customers are likely to buy a new smartphone. To reach these consumers various techniques are followed that use similar data of the products that are launched before. With this, we can also predict the products the consumer decides to purchase and those not bought. Hence, {decisions like buy or don't buy} these decisions form the attributes shown in Figure 1.

The various information such as customer's lifestyle, type of business, place of residence, the monthly income are all collected in order to predict the sale of a newly
launched smartphone. Similarly, various problems are solved with the help of data mining. Missing data is one of the most common problems in the research community. This may be occurring due to various reasons such as non-responses to questionnaire, item in laboratory or field settings. Improper handling of missing data, like deletion, can lead to biased inference in complete case analysis and statistical techniques.

The most prevalent problem with data quality is the presence of missing data. There are different causes of missing data such as sudden death of patient, malfunction of equipment, or refusal to answer certain questions and so on. In addition to this, some fraction of data may be erroneous and the alternative is to discard this erroneous data.

In most of the cases, the data set attributes are dependent on each other. Thus with the identification of relationship among attributes, missing values can be determined. The term imputation means replacing the missing value in a data set with some plausible values. The benefit from this approach is in the treatment of missing data which is independent of the learning algorithm used. This also helps the user to select the most suitable imputation method for each situation.

Missing data problem affects the real world, and in most of the cases it produces or results
- Statistical power information loss.
- Making it inappropriate or difficult to apply common data analysis methods.
- Bias statistical model estimates derivation.

2. Imputing Missing Data Methods

This section describes different methods available for imputing missing data.

2.1 Missingness Completely At Random: (MCAR)

The highest level of randomness, occurring when the probability of any case with missing value of an attribute does not dependent on known values or missing data.

With complete random missing values the probability of missing data for all the unit is same. For example a survey respondent deciding to answer the “earning” question by tossing a coin and refusing it to answer if “heads” shows up. Hence, if the data is missing completely at random then we have to remove the missing data which does not go with our inferences.

If for example, if all the survey respondents answer the earnings question by rolling a die and do not answer if “6” show up. Discarding the missing data does not bias your inference if the missing data is totally random.

2.2 Missing At Random: (MAR)

Summarised as a situation between MCAR and NIM, the probability of a missing value is dependent on the values of some other variable rather than the variable itself.

2.3 Non-Ignorable Missingness: (NIM)

This is complete opposite of MCAR as the chances of having missing values in a variable is dependent on the variable itself (for example a question regarding skill will not be answered if the skill is low).

2.4 Deductive Imputation: (DI)

The technique uses the redundancy in the data to interpret the missing response from the auxiliary information. For instance, a missing value can be deduced by subtraction, if the record has a series of values and their total. This method can be used on data materials with very high
likelihood of correct value or very close to correct value. For example in a survey panel with a constant variable for most of the time, a missing value in one wave of the panel can be attributed to the value recorded in the preceding or succeeding waves.

2.5 Mean imputation Overall (MO)
All missing responses in this method are assigned an overall respondent mean, a deterministic degenerate form of the linear function with no auxiliary variables.

2.6 Random imputation Overall (RO)
From the total data sample of respondents, a non-respondent value is assigned a randomly selected respondent. It is a stochastic degenerate of linear function with no auxiliary variables.

\[ y_{mi} = y_r + e_{mi} \]

with

\[ e_{mi} = y_{rk} - y_r \]

Reducing to \( y_{mi} = y_{rk} \)

From the Epsom sampling scheme, a subsample of respondents can be chosen to act as donors initially from a given Epsom sample (e.g., unrestricted sampling, SRS, proportionate stratified sampling, or systematic sampling).

2.7 Mean imputation within Classes (MC)
Dividing the total sample into imputation classes on the basis of the auxiliary variable values. In each class, the variable respondent mean gets assigned to all the non-respondents of that class:

\[ y_{mhi} = y_{rh} \]

for the \( i^{th} \) non-respondent in the class \( h \)

\[ h = (1, 2, 3, \ldots H) \]

By definition all cells are the classes in the cross-tabulated categorized auxiliary variables, but this symmetry is not important; instead, some auxiliary variables can be used for one part and others used for another part of the sample. Alternatively, cells groups can be combined. In the case of all cells are used in the cross-tabulation, the linear function is indicated as a model with main effects and all levels of interaction for the auxiliary variables.

2.8 Random Imputation within Classes (RC)
This algorithm is very similar to the random overall method except for its application within imputation.

2.9 Hot-deck imputation
Hot-Deck Imputation as a term has broad meaning, but here refers to the procedural sequence used by the Census Bureau with the labour force here specifically.

Current Population Survey (CPS) components alternatively referred as the traditional hotdeck procedure, starts with specification of imputation classes, and subsequently assigning \( y \)-variable of each class a single value, providing a point of origin for the process.

Initial values are arrived, for instance, by taking each class’s respondent value or a representative value, typically the class mean, from the previous batch of survey. The current survey records are treated sequentially. A record responsive to \( y \)-variable replaces the previously stored value by that value for imputation class. A missing response in the record is assigned currently stored value for its imputation class. A major advantage of this procedure is the economy in computing, since all imputations are made from a single pass through the data file.

2.10 Flexible Matching imputation (FM)
Modified hot-deck procedure also termed as flexible matching imputation is in use since 1976 for the CPS March Income Supplement. In this method respondents and non-respondents are sorted into a large number of imputation classes, constructed from a detailed categorization of a sizeable set of auxiliary variables. Hierarchical basis is used to match non-respondents with the respondents. If a nonrespondent cannot be matched with a respondent in the initial imputation class, classes collapse and the match is made at a lower level.

3. Imputation Analysis
Nonparametric imputation method for any kind of data, MissForest can be used for mixed-dataset, linear, or non-linear pattern data matrix or of complex interrelation. It only requires observation (the data frame rows supplied to the function) to be independent pairs. The algorithm is based on random forest but is dependent on its R implementation RandomForest. For every variable MissForest...
fits observed criteria with a random forest and then predicts the missing criteria, continuing to repeat the two steps until a criteria is met or until the user specified iterations maxed out\textsuperscript{13}.

MissForest runs iteratively, with constantly updating the imputed matrix variable-wise, as well as performance assessment between the iterations. This assessment is carried by taking into account the difference(s) between the previous and the new imputation result. The algorithm stops as soon as the difference (in case of one type of variable) or differences (in case of mixed-type of variables) increases. The user is provided with an estimate of imputation error by MissForest, based on the out-of-bag (OOB) error estimate of random forest. This estimate error shows an appropriate representation of the true imputation error\textsuperscript{14}.

For categorical dataset, we have used MissForest package to impute various missing values tested with missing data incrementally (ie. Missing 5, 10 attributes of the records in the original dataset). The dataset used is nursery dataset from the standard UCI Machine Learning Repository for testing purpose. To know the imputation success rate of this package, we have checked some of the missing data from the original dataset and analyzed the details. R Studio package is an open source tool for statistical analysis. Here, R Studio was used for missing data imputation in Data Mining.

**Table 1.** Characteristics of the Nursery Dataset

| Data Set Characteristics | Multivariate  |
|--------------------------|---------------|
| Attribute Characteristics | Categorical    |
| Number of Records        | 12960         |
| Number of Attributes     | 8             |
| Area                     | Social        |
| Missing Values           | No            |

The Table 1 represents the characteristics of the Nursery dataset which has the information like No. of records, attributes, whether it consists of any missing values in the dataset and area in which it collected

**Table 2.** Description of the Nursery Dataset

| Attribute Name | Parameters                                      |
|----------------|-------------------------------------------------|
| Parents        | usual, pretentious, great_pret                 |
| Has nursery    | proper, less_proper, improper, critical, very_critical |

The missing data is imputed by the package Missforest of R. The difference between the previous and the new imputed continuous and categorical parts of the data set in the following equation.

\[
\sum_{j=1}^{n} \frac{(x_{\text{imp}}^j - x_{\text{old}}^j)^2}{\sum_{j=1}^{n} (x_{\text{imp}}^j)^2}
\]

The continuous variable with 5 missing variable from the first 15 attributes and its corresponding graph with their difference for the set is also attached within Figure 2 and Figure 3.

**Figure 2.** Graphical representation of proposition of missing.
The imputed results is analysed as follows:

```plaintext
> TEST1.imp$OOBerror
NRMSE       PFC
NaN             0.3979592
```

### Table 3. Overall Imputed Results with PFC Error rate

| Test No. | Number of Records used for every test | Number of Attributes Missing | Number of attributes Imputed | Error Rate PFC (Approx.) |
|----------|--------------------------------------|------------------------------|-----------------------------|--------------------------|
| Test 1   | 5                                    | 5                            | 5                           | 0.4                      |
| Test 2   | 8                                    | 8                            | 8                           | 0.45                     |
| Test 3   | 12                                   | 12                           | 12                          | 0.54                     |
| Test 4   | 15                                   | 15                           | 15                          | 0.44                     |

It is observed from Table 3 that only the categorical variable is missed in data set. The error of Normalized Root Mean Squared Error (NRMSE) is not applicable here and so it is denoted as NaN whereas the Proportion of Falsely Classified (PFC) is computed for categorical dataset and it is 0.3979592. Here, the value of PFC is close to 0. Since, the good performance of MissForest has yielded a value close to 0 and bad performance a value around 1, the performance for this dataset is quite good.

### 4. Limitations in Most Commonly used Imputation Algorithms

The primary imputation approaches include following groups: 1) Constant Completer and Mean Completer value imputation which do not consider the correlation structure among cases or attributes; 2) imputational approaches based on statistics, such as Regression Imputation\(^\text{15}\), Support Vector Regression (SVR)\(^\text{16}\) and Expectation Maximization (EM)\(^\text{17}\), developed normally for only numeric attributes; 3) Approaches taking advantages of some related classification like imputation approaches, and clustering method of data mining or machine learning techniques. Approaches based on classification include KNNimp\(^\text{18}\) (K-nearest Neighbour Imputation), SKNNimp\(^\text{19}\) (sequential K-nearest neighbour method based imputation), and MKNNimp\(^\text{20}\) (K-nearest neighbour imputation method based on Mahalanobis distance). Cluster-based imputation approaches include KMI\(^\text{21}\) (K-means based Imputation) and FCM impute (Fuzzy C-means Clustering Imputation).

### 5. Conclusion

As only the categorical variable is missing in data set, the error of NRMSE i.e. Normalized Root Mean Squared Error is not applicable here and so it is denoted as NaN whereas the PFC is computed for categorical dataset and it is 0.3979592. Here, the value of PFC is close to 0. A good performance of MissForest give a value close to 0 and bad performance a value around 1, the measurement of performance for this dataset is quite good.

Many of the imputational approaches mentioned here use only information pertaining to complete objects for estimating the missing data of the incomplete objects. This approach is good when the missing data ratio is low, however requires extension when the problem has high ratio of missing data. Unlike the discrete categorical data, also the commonest data type in data mining, many of these imputational approaches can be applied to missing data with only continuous attribute such as DNA data, spatial data, economic data, and so on. In the future work, rough set theory may be incorporated in the incomplete data analysis method to achieve better results with missing data of categorical attribute with discrete values.

### 6. References

1. Brown ML, Kros JF. Data mining and the impact of missing data. Ind Manag Data Syst [Internet]. 2003; 103(8):611–21. Available from: http://dx.doi.org/10.1108/02635570310497657
2. Paramasivam I, Thiagarajan H, Savarimuthu N. Genetic Chromo Dynamic Framework for Imputation of Missing Data in Type II Diabetes Databases. Special Issue Int Journal of Computer, Internet Manag. 2009; 17(sp1):51.1–51.11.
3. Aggarwal CC, Philip SY. A condensation approach to privacy preserving data mining. In: Advances in Database Technology-EDBT 2004. Springer; 2004. p. 183–99.
4. Ahmad A, Dey L. A k-mean clustering algorithm for mixed numeric and categorical data. Data Knowl Eng [Internet]. 2007 Nov [cited 2013 Oct 1]; 63(2):503–27. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0169023X0700050X
5. Ahmad A, Dey L. A method to compute distance between two categorical values of same attribute in unsupervised learning for categorical data set. Pattern Recognit Lett. Elsevier. 2007; 28(1):110–8.
6. Alexandre Evfimievski Cornell University, Ithaca N. Randomization in Privacy Preserving Data Mining. ACM SIGKDD Explor Newsl. 2002; 4(2):43–8.
7. Batista GE, Monard MC. An analysis of four missing data treatment methods for supervised learning. Appl Artif Intell. Taylor & Francis. 2003; 17(5-6):519–33.
8. Van Ginkel JR, der Ark LA, Sijtsma K. Multiple imputation of item scores in test and questionnaire data, and influence on psychometric results. Multivariate Behav Res. Taylor & Francis. 2007; 42(2):387–414.
9. Song Q, Shepperd M. Missing Data Imputation Techniques. Int J Bus Intell data Min. Inderscience Publishers. 2007; 2(3):261–91.
10. Sentas P, Angelis L. Categorical missing data imputation for software cost estimation by multinomial logistic regression. J Syst Softw [Internet]. 2006; 79(3):404–14. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0164121205000129
11. Kalton G, Kaspryzk D. Imputing for missing survey responses. In: Proceedings of the Section on Survey Research Methods, American Statistical Association. 1982.
12. Breiman L. Random forests. Mach Learn. Springer. 2001; 45(1):5–32.