JCSE International Journal of Computer Sciences and Engineering Open Access
Survey Paper Volume-5, Issue-2 E-ISSN: 2347-2693

A Study on Missing Data Management

Malay Mitra[1*] and R. K. Samanta[2]

[1*]Department of Computer Science and Application, University of North Bengal, Raja Rammuhunpur, India
[2]Department of Computer Science and Application, University of North Bengal, Raja Rammuhunpur, India

[*]*Corresponding Author: malay.mitra68@gmail.com*

Available online at: www.ijcseonline.org

Received: 15/Jan/2017 Revised: 20/Jan/2017 Accepted: 15/Feb/2017 Published: 28/Feb/2017

***Abstract*— Missing data, a persistent problem in most scientific research, should be handled very carefully, as role of data are vital in every analysis. Mishandling missing values may cause distorted analysis or may generate biased results. Valid and reliable models require good data preparation. Dozens of techniques have been proposed by methodologists to address the problem. Appropriate method should be taken into consideration for a particular study in order to achieve efficient and valid analysis. In this study we discuss different methods to handle missing data and compare three imputation methods: Arithmetic Mean Imputation, Regression Imputation and Multiple Imputation using EMB algorithm, performed on three data sets from UCI repository under the assumption of MAR based on Root Mean Square Error (RMSE) as an evaluation criteria.**

***Keywords*- UCI database, Missing At Random (MAR), Missing Completely At Random (MCAR), Missing Not At Random (MNAR), Multiple Imputation, Expectation Maximization with Bootstrap approach (EMB), Root Mean Square Error (RMSE).**

I. INTRODUCTION

In most scientific research domain like Biology [1], Medicine [2] missing data are common problems. One of the most challenging decision confronting researcher is to choose the most appropriate method to handle missing data. Numerous methods are used in literature to handle missing data. Moreover handling missing data are not typically addressed in most literature. Unfortunately most of the statistical packages implement old standby techniques which are prone to statistical bias. There are different methods which are being used by people:

- Delete the records containing missing data;
- Use attribute mean;
- Use attribute median;
- Use a global constant to fill in for missing values which seem not relevant to the decision attribute;
- Use a data mining method.

In this study we compare different imputation methods. We use three datasets – UCI Breast Cancer Dataset, UCI Chronic Kidney Disease Dataset and UCI Hepatitis Disease Dataset without missing value, based on evaluation criteria Root Mean Square Error (RMSE).

The paper is organized as follows. In section II, missing data mechanisms are discussed. Section III explains the techniques of handling missing data. Section IV describes data sets used in this study. Section V explains the principle of analysis. Section VI represents the evaluation criteria. Section VII presents the results. Lastly, our conclusions are summarized.

II. MISSING DATA MECHANISMS

Rubin [3] defined missing data based on three missingness mechanisms [4] – Missing at Random (MAR), Missing Completely at Random (MCAR) and Missing Not at Random (MNAR).

Data are missing at random when there is a relation between the probability of missing data for a variable to some other measured variable or variables, but not to the values of itself. MAR as its name does not imply missing in haphazard fashion, but it actually means that the probability of missing data is systematically related to other variable.

Data are missing completely at random when the probability of missing data for a variable is unrelated to any other measured variable and to the values of itself. MCAR implies missing completely in haphazard fashion. MCAR is a more restrictive condition than MAR as it assumes that missingness is completely unrelated to the data [5].

Data are missing not at random when the probability of missing data for a variable is related to the values of itself, even after controlling for other variable.

III. MISSING DATA HANDLING

Dealing with missing data includes – removing the cases with missing values or imputing the missing values. Dozens of techniques have been found in literature to handle missing data problem. Some of these techniques are – List-wise deletion, Pair-wise deletion, Arithmetic Mean Imputation, Regression Imputation, Multiple Imputation with EMB approach.

A. LIST-WISE DELTION

In list-wise deletion method data for any case which has one or more missing values are deleted. This is why the method is also known as complete-case analysis [6]. The main advantage of this method is that it is easy to implement and also available as standard option for statistical packages. In most situations the resulting reduced dataset as obtained by applying list-wise deletion may lead to decreased statistical analysis power and also important knowledge may be missed. Another disadvantage is that this method assumes MCAR. If data are not in MCAR, list-wise deletion produces distorted result. In particular for large dataset where missing values are very minimal, this method may be appropriate.

B. PAIR-WISE DELETION

To mitigate the loss of data that occurs in list-wise deletion, pair-wise deletion method eliminates cases on an analysis by analysis basis only on available cases. Pair-wise deletion uses the subset of cases with complete data for each pair of variables to compute correlation or covariance matrix. The strength of associationship between a pair of variables is measured by correlation. The correlation coefficients for each pair of variables for which data are available will take the data into account. Thus pair-wise deletion maximizes the use of data as much as possible, which increases the power of analysis. Pair-wise deletion method tends to be more powerful than list-wise deletion, particularly when the variables in a dataset have low to moderate correlations. The main advantage of pair-wise deletion is that it is easy to implement and also available in standard statistical packages.

The disadvantage of pair-wise deletion is that if the assumption of MCAR does not hold, it produces distorted result as it requires data in MCAR. In pair-wise deletion it is difficult to compute standard errors as average sample size is used to the entire correlation matrix. Thus it produces standard errors either underestimated or overestimated. Another disadvantage is that this technique may yield correlation outside [-1,1] which causes estimation problems for multivariate analyses that use correlation matrix as input.

C. SINGLE IMPUTATION

Single imputation methods impute data for unobserved values in the dataset prior to analysis. It replaces a single value for each missing value in the dataset. Out of many single imputation methods available we discussed two of them – Arithmetic Mean Imputation and Regression Imputation.

1) ARITHMETIC MEAN IMPUTATION: In this method the arithmetic mean of observed values for an attribute replaces all the missing values for that attribute. This is the simplest imputation method, but produces biased result. It increases the size of sample as well as the power of analysis. According to Rubin [4] mean substitution decreases the variability in the dataset, as mean that is the same value is used as a substitute for all the missing values.

2) REGRESSION IMPUTATION: It uses regression to predict missing values from other variables of known values. Variables containing missing data is assumed to be dependent while the other variables are considered as independent. If we consider bivariate dataset with attribute X and Y, missing values are computed from the regression equation :

$$Y = b*X + a \qquad \text{(i)}$$

Here we assume that value of dependent variable Y is to be predicted from independent variable X by estimating the regression with the available data of X and Y. The values of a and b are computed from the following formulae

$$a = \frac{\sum y \sum x^2 - \sum x \sum (x*y)}{n \sum x^2 - (\sum x)^2} \qquad \text{(ii)}$$

$$b = \frac{n \sum (x*y) - \sum x \sum y}{n \sum x^2 - (\sum x)^2} \qquad \text{(iii)}$$

Regression imputation is better than mean imputation, but it also has predictable biases.

D. MULTIPLE IMPUTATIONS

A bootstrap-based EMB algorithm [7] performs multiple imputation for missing values. In multiple imputation, values are imputed for each missing value of the data set and completed *m* data sets are generated. In these imputed data sets with complete data, the known values remain same for each set but the imputed missing values may be different for each set. After imputation, analysis is done with each imputed data set and the results are combined. There are different combination techniques one can adopt [7, 8].

Fig. 1 shows the schematic view of Multiple Imputation using EMB approach. Multiple imputations are found to produce more accurate results compared to list-wise deletion, arithmetic mean imputation. This technique reduces bias and increases efficiency. In this multiple imputation technique, MAR (missing at random) is assumed. It considers MAR, likelihood, law of iterated expectations, and a flat prior to compute posterior. From the posterior, it has to take draws. The EM [9] algorithm is to find the mode of the posterior. This EMB algorithm uses the EM algorithm with bootstrap approach to take draws from this posterior. For each draw, the data is bootstrapped to simulate estimation uncertainty and then run EM algorithm to find the mode of the posterior for the bootstrapped data, which also gives fundamental uncertainty [10]. After having draws imputations are done using observed part D(observed) and unobserved part D(missing) as well as mean vector μ and covariance matrix Σ with linear regression.

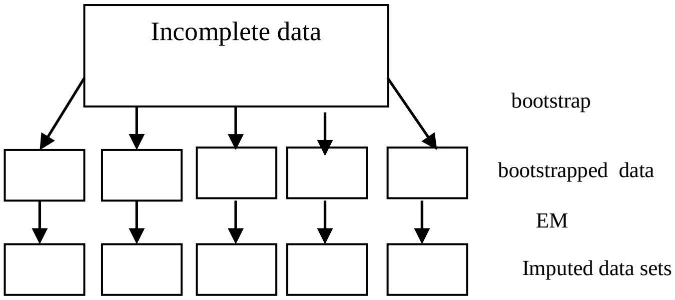

Figure 1. A schematic view of Multiple Imputation

IV. DATA SET

Considering the variability of relative performance of different methods across datasets, results were generated based on three reference datasets : Breast Cancer dataset, Chronic Kidney Disease Dataset and Hepatitis Disease Dataset.

The UCI Breast Cancer dataset is a very popular dataset contributed by Dr. William H. Wolberg (1989-91), University of Wisconsin Hospital, Madison, USA. The records came periodically as Dr. Woolberg reported his clinical cases. The data set contains 10 attributes plus one attribute for class (binary). The total number of instances are 699. In this data set there are 16 instances with missing values. After discarding these 16 instances we use 683 instances in this work.

The UCI Chronic Kidney Disease data set contains 24 attributes plus one attribute for class (binary). It contains 400 samples to two different classes (‘CKD’ – 250 cases and ‘NOTCKD’ – 150 cases). The dataset contains a number of missing values. After eliminating missing values 158 samples are used in this study.

Hepatitis data set from UCI Machine Learning Repository contains 19 attributes plus one attribute for class (binary). It contains 155 samples to two different classes (‘die’ – 32 cases; ‘live’ – 123 cases). There are a number of missing values in the data set. Number of samples used is 139 based on the attributes taken into consideration in this study.

V. PRINCIPLE OF ANALYSIS

Figure 2 shows the general principle of analysis. From the original data sets without missing values we produced bivariate data sets by selecting only two attributes from each data set and also introduced in the data a varying percentage of missing values (eg. 10%, 20% and 30%) in such a way that MAR is assumed. From Breast Cancer data set we selected the attributes – Clump Thickness and Uniformity of Cell Size. The two attributes which are selected from Chronic Kidney Disease data set are – Albumin and Serum Creatinine. Similarly, from Hepatitis Disease data set we chose attributes – Albumin and Billirubin. The values of dependent variables (for Breast Cancer data set – Uniformity of Cell Size, for CKD data set - Serum Creatinine, for Hepatitis data set – Billirubin) are missing at random (MAR) as they are systematically missing as a function of respective independent variables (for Breast Cancer data set – Clump Thickness, for CKD data set – Albumin, for Hepatitis data set – Albumin). These simulated missing values are imputed using 3 methods - Arithmetic Mean, Regression and Multiple Imputation using EMB approach. Performances are measured by evaluating Root Mean Square Error (RMSE).

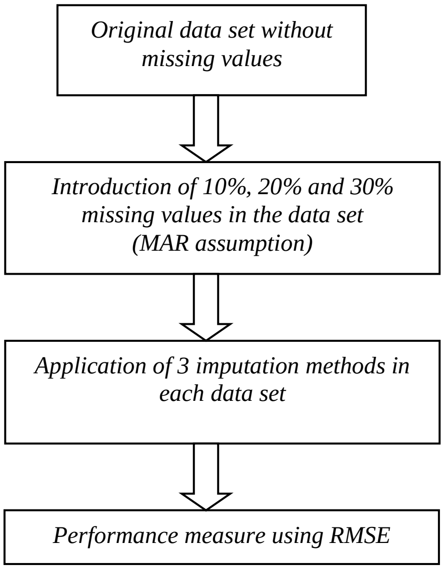
Original data set without missing values
Introduction of 10%, 20% and 30% missing values in the data set (MAR assumption)
Application of 3 imputation methods in each data set
Performance measure using RMSE

Figure 2. Block diagram of principle of analysis

VI. EVALUATION CRITERIA

We compare three imputation methods on the basis of Root Mean Square (RMSE) which measures the difference between imputed value and true value.

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(X_i^{obs} - X_i^{imputed})^2}{n}} \qquad (3)$$

VII. RESULTS

Results are summarized in Table I, Table II and Table III.

TABLE I. RESULTS FOR BREAST CANCER DATA SET OF UCI

Percentage of missing value	**Imputation Methods**	**Root Mean Square Error (RMSE)**
10	Arithmetic Mean	4.831
	Regression	**2.811**
	Multiple Imputation using EMB	2.867
20	Arithmetic Mean	5.226
	Regression	**2.86**
	Multiple Imputation using EMB	3.077
30	Arithmetic Mean	5.203
	Regression	**3.033**
	Multiple Imputation using EMB	3.494

TABLE II. RESULTS FOR CHRONIC KIDNEY DISEASE DATA SET OF *UCI*

Percentage of missing value	**Imputation Methods**	**Root Mean Square Error (RMSE)**
10	Arithmetic Mean	5.376
	Regression	**4.646**
	Multiple Imputation using EMB	4.785
20	Arithmetic Mean	6.52
	Regression	**5.591**
	Multiple Imputation using EMB	5.765
30	Arithmetic Mean	1.824
	Regression	**0.287**
	Multiple Imputation using EMB	1.193

TABLE III. RESULTS FOR HEPATITIS DISEASE DATA SET OF *UCI*

Percentage of missing value	**Imputation Methods**	**Root Mean Square Error (RMSE)**
10	Arithmetic Mean	**0.782**
	Regression	1.375
	Multiple Imputation using EMB	1.621
20	Arithmetic Mean	**0.836**
	Regression	1.74
	Multiple Imputation using EMB	1.573
30	Arithmetic Mean	0.948
	Regression	0.838
	Multiple Imputation using EMB	**0.795**

From the above tables it is observed that in almost all cases performance of Regression Imputation and Multiple Imputation using EMB are same, though in most of the cases regression imputation provides better result than the later. In case of Hepatitis Disease data set for 10% and 20 % missing values imputation using Arithmetic Mean leads to better result as compared to other two methods.

VIII. DISCUSSION AND CONCLUSION

Missing data, a part of many studies, are handled by several alternative ways to overcome the drawbacks. Comparative studies are needed to ensure which imputation method should be well suited for a particular study. Only a few literatures address an evaluation of existing imputation methods.

In this work, we performed a neutral comparative study of three imputation methods based on three UCI data sets of various sizes under the assumption of MAR. We did not consider elimination processes like List-Wise deletion and Pair-Wise deletion, as these methods are applicable only for large data set with minimal number of missing values, otherwise there may be a chance of losing important information. So, we concentrated only on imputation methods. Imputation accuracy is measured by Root Mean Square Error (RMSE).

The limitation of our study is that the results are limited to data matrices of numerical values. Careful attention should be taken into consideration for other type of variables also [11].

In conclusion, it can be suggested that there is no universal imputation method performing best in every situation, but for bi-variate data set if the data are missing at random, imputation using regression should be taken into consideration. For multivariate data set the regression imputation is somewhat complicated to implement. Regression imputation also requires data which are missing at random. So it is also suggested to consider multiple imputation approaches for multivariate data set which are in MAR or MCAR.

REFERENCES

[1] Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, “Missing value estimation methods for dna microarrays”, Bioinformatics Vol.17, pp.520-525, 2001.

[2] Lewis HD, “Missing data in clinical trials”, New England Journal of Medicine, Vol. 367, pp. 2557-2558, 2012.

[3] Rubin DB, “Inference and missing data”, Biometrica Vol. 63, pp. 581-592, 1976.

[4] Little RJA, Rubin DB, Statistical Analysis with Missing Data (2nd edn.), Wiley-Interscience, 2002.

[5] N.Durga, D.Ragupathi and V. Raj Kumar, "Uses of HDFS in Metadata Management System", International Journal of Computer Sciences and Engineering, Vol.2(9), pp.145-150, Sep 2014

[6] Schafer. J. L. & Graham, J.N., “Missing Data: Our view of the state of the art”, Psychological Methods, Vol. 7, pp. 147-177, 2002.

[7] Bhambri V., "Data Mining as a Solution for Data Management in Banking Sector", International Journal of Computer Sciences and Engineering, Vol.1(1), pp.20-25, Sep -2013.

[8] King G, Tomaz M, Wittenberg J, “Making the Most of Statistical Analyses: Improving and Presentation”, American Journal of Political Science, Vol. 44(2), pp. 341-355, 2000.

[9] Dempster A. P., Laird N. M., Rubin D. B., "Maximum Likelihood from Incomplete Data via the EM Algorithm", Journal of the Royal Statistical Society, Vol. 39(1) , pp. 1–38, 1977.

[10] Honaker J., King G., “What to do About Missing Values in Time Series Cross-Section Data”, *American J. of Political Science*, Vol. 54(2), pp.561-581, 2010.

[11] Horton NJ, Kleinman KP, “Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models”, The American Statistician Vol.61, pp. 79-90, 2007.