International Journal on Recent Researches In Science, Engineering & Technology
(Division of Computer Science and Engineering)
A Journal Established in early 2000 as National journal and upgraded to International journal in 2013 and is in existence for the last 10 years. It is run by Retired Professors from NIT, Trichy. It is an absolutely free (No processing charges, No publishing charges etc) Journal Indexed in JIR, DIIF and SJIF.
Research Paper
Available online at: www.jrrset.com

ISSN (Print) : 2347-6729
ISSN (Online) : 2348-3105
Volume 4, Issue 12, December 2016.
JIR IF : 2.54
DIIF IF : 1.46
SJIF IF : 1.329

A Study on Missing Data Management

Dr.C.N.Ravi
Department of Computer Science and Engineering,Shadan College of Engineering and Technology HYD,T.S,INDIA

Abstract
Missing data, a power hassle in most scientific research, need to be handled very carefully, as positions of data are fundamental in each analysis. Mishandling missing values may additionally cause distorted evaluation or may also generate biased results. Valid and reliable fashions require accurate statistics preparation. Dozens of methods have been proposed by means of methodologists to tackle the problem. Appropriate method be taken into consideration for a precise study in order to acquire environment friendly and valid analysis. In this find out about we talk about special methods to cope with missing records and examine three imputation methods: Arithmetic Mean Imputation, Regression Imputation and Multiple Imputation using EMB algorithm, performed on three facts sets from UCI repository under the assumption of MAR primarily based on Root Mean Square Error (RMSE) as evaluation criteria.
Keywords - Multiple Imputation, Expectation Maximization with Bootstrap approach (EMB), Root Mean Square Error (RMSE)., UCI database, Missing At Random (MAR), Missing Completely At Random (MCAR), Missing Not At Random (MNAR)

I. INTRODUCTION

In most scientific research domain like Biology [1], Medicine [2] missing statistics are common problems. One of the most difficult decision confronting researcher is to select the most terrific approach to take care of lacking data. Numerous techniques are used in literature to manage lacking data. Moreover dealing with missing statistics are no longer typically addressed in most literature. Unfortunately most of the statistical packages implement historic standby methods which are prone to statistical bias. There are exceptional techniques which are being used by people: Delete the data containing lacking data;

- Use attribute mean;
- Use attribute median;
- Use a global consistent to fill in for lacking values two which appear no longer applicable to the decision attribute;
- Use a data mining method.

In this learn about we examine one-of-a-kind 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 standards Root Mean Square Error (RMSE). The paper is organized as follows. In area II, missing data mechanisms are discussed. Section III explains the methods of handling missing data. Section IV describes statistics units used in this study. Section V

explains the precept of analysis. Section VI represents the contrast criteria. Section VII conclusions are summarized.

II. MISSING DATA MECHANISMS

Rubin [3] described lacking records primarily based on three missingness mechanisms [4] – Missing at Random (MAR), Missing Completely at Random (MCAR) and Missing Not at Random (MNAR). Data are lacking at random when there is a relation between the chance of lacking data for a variable to some different measured variable or variables, but not to the values of itself. MAR as its identify does not imply missing in haphazard fashion, however it certainly potential that the likelihood of missing records is systematically associated to different variable. Data are missing completely at random when the probability of missing information for a variable is unrelated to any other measured variable and to the values of itself. MCAR implies lacking totally in haphazard fashion. MCAR is a greater restrictive condition than MAR as it assumes that missingness is absolutely unrelated to the statistics [5]. Data are missing now not at random when the probability of missing information for a variable is associated to the values of itself, even after controlling for other variable.

III. MISSING DATA HANDLING

Dealing with lacking records includes – disposing of the cases with missing values or imputing the lacking values. Dozens of techniques have been observed in literature to handle lacking facts problem. Some of these strategies 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 records for any case which has one or extra lacking values are deleted. This is why the method is also regarded as complete-case analysis [6]. The most important advantage of this method is that it is effortless to enforce and also accessible as standard alternative for statistical packages. In most conditions the ensuing decreased dataset as got by using making use of list-wise deletion can also lead to reduced statistical evaluation strength and also necessary understanding may be missed. Another downside is that this technique assumes MCAR. If facts are no longer in MCAR, list-wise deletion produces distorted result. In particular for giant dataset the place missing values are very minimal, this technique may also be appropriate.

B. PAIR-WISE DELETION

To mitigate the loss of records that takes place in list-wise deletion, pair-wise deletion technique eliminates instances on an analysis by evaluation foundation only on on hand cases. Pair-wise deletion uses the subset of cases with entire data for each pair of variables to compute correlation or covariance matrix. The electricity of associationship between a pair of variables is measured by correlation. The correlation coefficients for every pair of variables for which records are accessible will take the records into account. Thus pair-wise deletion maximizes the use of data as tons as possible, which will increase the energy of analysis. Pair-wise deletion technique tends to be greater powerful than list-wise deletion, specifically when the variables in a dataset have low to average correlations. The essential gain of pair-wise deletion is that it is convenient to put into effect and additionally on hand in general statistical packages. The downside of pair-wise deletion is that if the assumption of MCAR does now not hold, it produces distorted end result as it requires facts in MCAR. In pair-wise deletion it is challenging to compute fashionable mistakes as average sample measurement is used to the entire correlation matrix. Thus it produces wellknown mistakes both underestimated or overestimated. Another drawback is that this approach may yield correlation outside [-1,1] which causes estimation issues for multivariate analyses that use correlation matrix as input.

C. SINGLE IMPUTATION Single imputation techniques impute statistics for unobserved values in the dataset prior to analysis. It replaces a single price for each missing fee in the dataset. Out of many single imputation strategies handy we discussed two of them – Arithmetic Mean Imputation and Regression Imputation.

1) ARITHMETIC MEAN IMPUTATION:

In this technique the arithmetic imply of discovered values for an attribute replaces all the missing values for that attribute. This is the easiest imputation method, however produces biased result. It increases the size of sample as properly as the strength of analysis. According to Rubin [4] suggest 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 makes use of regression to predict missing values from different variables of regarded values. Variables containing missing statistics is assumed to be based whilst the other variables are viewed as independent. If we think about bivariate dataset with attribute X and Y, missing values are computed from the regression equation :

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

Here we count on that fee of based variable Y is to be anticipated from unbiased variable X by estimating the regression with the available data of X and Y. The values of a and b are computed

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

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

from the following formulae Regression imputation is better than mean imputation, but it also has predictable biases.

D. MULTIPLE IMPUTATIONS

A bootstrap-based EMB algorithm [7] performs more than one imputations for lacking values. In more than one imputation, values are imputed for each lacking cost of the information set and executed m information units are generated. In these imputed statistics units with whole data, the recognized values stay identical for every set but the imputed missing values may be exceptional for each set. After imputation, evaluation is achieved with every imputed statistics set and the consequences are combined. There are one of a kind mixture techniques one can undertake [7, 8]. Fig. 1 shows the schematic view of Multiple Imputation using EMB approach. Multiple imputations are observed to produce greater accurate consequences compared to list-wise deletion, arithmetic mean imputation. This technique reduces bias and increases efficiency. In this a couple of imputation technique, MAR (missing at random) is assumed. It considers MAR, likelihood, regulation of iterated expectations, and a flat prior to compute posterior. From the posterior, it has to take draws. The EM [9] algorithm is to discover the mode of the posterior. This EMB algorithm uses the EM algorithm with bootstrap method to take attracts 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 offers critical uncertainty [10]. After having draws imputations are done using observed phase D(observed) and unobserved part D(□ and covariance matrix □missing) as

properly as imply vector with linear regression.

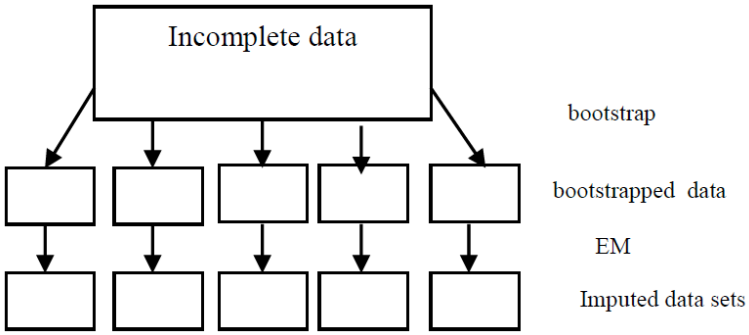

Fig.1. Schematic View of Multiple Imputation

IV. DATA SET

Considering the variability of relative overall performance of different strategies across datasets, consequences have been 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 using Dr. William H. Wolberg (1989-91), University of Wisconsin Hospital, Madison, USA. The documents came periodically as Dr. Woolberg mentioned his clinical cases. The facts set includes 10 attributes plus one attribute for class (binary). The total range of situations are 699. In this facts set there are sixteen situations with lacking values. After discarding these 16 situations we use 683 instances in this work. The UCI Chronic Kidney Disease statistics set contains 24 attributes plus one attribute for type (binary). It carries four hundred samples to two exceptional lessons ('CKD' – 250 cases and 'NOTCKD' – a hundred and fifty cases). The dataset carries a variety of missing values. After removing lacking values 158 samples are used in this study. Hepatitis statistics set from UCI Machine Learning Repository carries 19 attributes plus one attribute for class (binary). It includes a hundred and fifty five samples to two one of a kind training ('die' – 32 cases; 'live' – 123 cases). There are a variety of lacking values in the facts set. Number of samples used is 139 primarily based on the attributes taken into consideration in this study.

VI. EVALUATION CRITERIA

We evaluate three imputation methods on the source of Root Mean Square (RMSE) which measures the dissimilarity between imputed value and true value.

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

From the above it is observed that in nearly all cases performance of Regression Imputation and Multiple Imputation the usage of EMB are same, though in most of the cases regression imputation offers higher result than the later. In case of Hepatitis Disease information set for 10% and 20 percent lacking values imputation the use of Arithmetic Mean leads to higher end result as compared to other two methods.

VII. DISCUSSION AND CONCLUSION

Missing data, a phase of many studies, are dealt with via quite a few choice methods to overcome the drawbacks. Comparative studies are wished to ensure which imputation method be nicely applicable for a specific study. Only a few literatures tackle an evaluation of existing imputation methods. In this work, we performed a impartial comparative study of three imputation methods based totally on three UCI records units of a number sizes below the assumption of MAR. We did now not consider elimination methods like List-Wise deletion and Pair-Wise deletion, as these methods are applicable

solely for giant data set with minimal range of lacking values, in any other case there may also be a hazard of losing vital information. So, we concentrated solely on imputation methods. Imputation accuracy is measured with the aid of Root Mean Square Error (RMSE). The problem of our find out about is that the results are restricted to statistics matrices of numerical values. Careful attention should be taken into consideration for different kind of variables additionally [11]. In conclusion, it can be cautioned that there is no everyday imputation method performing pleasant in each situation, but for bi-variate facts set if the records are missing at random, imputation the usage of regression should be taken into consideration. For multivariate information set the regression imputation is somewhat problematic to implement. Regression imputation additionally requires information which are lacking at random. So it is also cautioned to reflect onconsideration on a couple of imputation procedures for multivariate facts 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.