Imputation Methods for Missing Data for a Proposed VASA Dataset

S.Anitha, M.Vanitha

***Abstract:** Preprocessing is the presentation of raw data before apply the actual statistical method. Data preprocessing is one of the most vital steps in data mining process and it deals with the preparation and transformation of the initial dataset. It is prominent because the investigating data which is not properly preprocessed could lead to the result which is not accurate and meaningless. Almost every research have missing data and introduce an element into data analysis using some method. To consider the missing values that need to provide an efficient and valid analysis. Missing imputation is one of the process in data cleaning. Here, four different types of imputation methods are compared: Mean, Singular Value Decomposition (SVD), K-Nearest Neighbors (KNN), Bayesian Principal Component Analysis (BPCA). Comparison was performed in the real VASA dataset and based on performance evaluation criteria such as Mean Square Error (MSE) and Root Mean Square Error (RMSE). BPCA is the best imputation method of interest which deserve further consideration in practice.*

***Keywords:** Data Preprocessing, Missing Data, Imputation Methods, BPCA, RMSE.*

I. INTRODUCTION

In healthcare industry consists the huge amount of data, which is unfortunately not knowledgeable to discover the hidden information for effective decision making. Data cleaning, Data selection and transformation, Interpretation / Evaluation are the steps in Data Preprocessing. In real-world datasets may contain missing values for different reasons. Missing data is a vast issue that can affect the ability to produce proper result. Few data mining techniques can cope with such situations. Dataset may have NaNs or blanks. If missing data ignoring techniques is used to delete the cases that contain missing values. And also this may not lead to the most efficient utilization of the data. There are so many ways to handle this problem is to get rid of the observations that have missing values. Normally, four qualitatively various types of missing data, they are Structurally Missing, Missing Completely at Random (MCAR), Missing at Random, Missing Not at Random.

Revised Manuscript Received on November 30, 2019.
***** Correspondence Author

Anitha.S*, Research Scholar, Dept of Computer Applications, Alagappa University, Karaikudi, India. Email: nathan.anitha@gmail.com

Dr.Vanitha.M, Asst. Professor, Dept of Computer Applications, Alagappa University, Karaikudi, India. Email: mvanitharavi@gmail.com

II. VASA DATASET DESCRIPTION

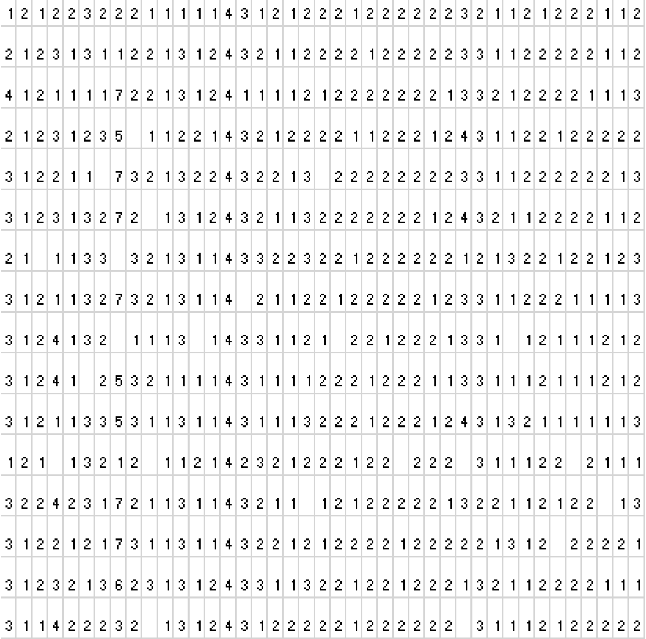
Preparation of questionnaire is the initial step of the process. The various factors in the questionnaire are : age, gender, marital status, Highest Qualification, Working Environment, Nature of works, Working Sector, Occupation, Working hours, Overtime hours, Job Satisfaction, Working Experience, Sufficient Income, Work pressure and various disease such as Diabetes, Blood Pressure, Headache, Mental Illness(depression), Heart Disease, Gastritis(Ulcer), Stroke and so on. The questionnaries are distributed to employees in different sectors and the data are collected. Each employee from separate concerns like schools, banks, private companies were asked to fill up the questionnaire and based on that data warehouse is organized.

Table 1. Attribute Description

Attribute	**Values**
Age	1.20-30,2.31-40,3.41-50,4.Above 50
Sex	1. Male, 2. Female
Marital status	1.Married,2.Single
Highest Qualification	1.Below SSLC, 2.Graduate, 3.Post Graduate, 4.Technical Qualification 5.Others
Working Environment	1. City, 2. Village
Nature of works	1.Physical,2.Mental,3.Both
Working Sector	1.Govt.2.Private,3.Business
Occupation	1.Teacher/Professor, 2.Doctor, 3.Engineer, 4.Bank Employees, 5.Business, 6.Labours, 7.others
Working hours	1.6hrs, 2.8hrs, 3.>8hrs
Overtime hours	1.2hrs, 2.5hrs, 3.>5hrs
Job Satisfaction	1. Yes, 2. No
Working Experience	1.2-5 years,2.6-10 years,3.>10 years
Sufficient Income	1. Yes, 2. No
Workpressure	1. Yes, 2. No

Addiction	1. Smoking, 2. Alcohol, 3,Tobacco
Food	1. Veg., 2.NV, 3. Both
Sleeping hours	1.<6hrs, 2.6-8hrs, 3.>8hrs
Physical Activities	1. Yes, 2. No
Type of roles	1. Major, 2. Minor
Leisure Time	1.Reading books, 2.Listening Music, 3.Watching TV, 4.Playing chess
Diabetes	1. Yes, 2. No
Blood Pressure	1. Yes, 2. No
Headache	1. Yes, 2. No
Mental Illness(depression)	1. Yes, 2. No
Heart Disease	1. Yes, 2. No
Gastritis(Ulcer)	1. Yes, 2. No
Stroke	1. Yes, 2. No
Exercise regularly	1. Yes, 2. No
Any continuous medication	1. Yes, 2. No
how long suffer	1.6 months,2.1 year,3.>1year
Body Weight	1.<40kg, 2.41-60kg, 3.>60kg
Feel tired or depressed	1. Yes, 2. No
Working conditions	1.satisfactory, 2.dissatisfactory, 3.Can't say
better on the job if conditions changed	1. Yes, 2. No
job affect your family	1. Yes, 2. No
Control over life	1. Yes, 2. No
Any argument	1. Yes, 2. No
Activity for stress relief in organization	1. Yes, 2. No
Are underpaid	1. Yes, 2. No
Are undervalued	1. Yes, 2. No
Appreciation for good work	1. Yes, 2. No

Table 2. Sample VASA Dataset with missing values



1	2	1	2	2	3	2	2	2	1	1	1	1	1	4	3	1	2	1	2	2	2	1	2	2	2	2	2	2	3	2	1	1	2	1	2	2	2	1	1	2
2	1	2	3	1	3	1	1	2	2	1	3	1	2	4	3	2	1	1	2	2	2	2	1	2	2	2	2	2	3	3	1	1	2	2	2	2	2	1	1	2
4	1	2	1	1	1	1	7	2	2	1	3	1	2	4	1	1	1	1	2	1	2	2	2	2	2	2	2	1	3	3	2	1	2	2	2	2	1	1	1	3
2	1	2	3	1	2	3	5		1	1	2	2	1	4	3	2	1	2	2	2	2	1	1	2	2	2	1	2	4	3	1	1	2	2	1	2	2	2	2	2
3	1	2	2	1	1		7	3	2	1	3	2	2	4	3	2	2	1	3		2	2	2	2	2	2	2	2	3	3	1	1	2	2	2	2	2	2	1	3
3	1	2	3	1	3	2	7	2		1	3	1	2	4	3	2	1	1	3	2	2	2	2	2	2	2	1	2	4	3	2	1	1	2	2	2	2	1	1	2
2	1		1	1	3	3		3	2	1	3	1	1	4	3	3	2	2	3	2	2	1	2	2	2	2	2	2	1	2	1	3	2	2	1	2	2	1	2	3
3	1	2	1	1	3	2	7	3	2	1	3	1	1	4		2	1	1	2	2	1	2	2	2	2	2	1	2	3	3	1	1	2	2	2	1	1	1	1	3
3	1	2	4	1	3	2		1	1	1	3		1	4	3	3	1	1	2	1		2	2	1	2	2	2	1	3	3	1		1	2	1	1	1	2	1	2
3	1	2	4	1		2	5	3	2	1	1	1	1	4	3	1	1	1	1	2	2	2	1	2	2	2	1	1	3	3	1	1	1	2	1	1	1	2	1	2
3	1	2	1	1	3	3	5	3	1	1	3	1	1	4	3	1	1	1	3	2	2	2	1	2	2	2	1	2	4	3	1	3	2	1	1	1	1	1	1	3
1	2	1		1	3	2	1	2		1	1	2	1	4	2	3	2	1	2	2	2	1	2	2		2	2	2		3	1	1	1	2	2		2	1	1	1
3	2	2	4	2	3	1	7	2	1	1	3	1	1	4	3	2	1	1		1	2	1	2	2	2	2	2	1	3	2	2	1	1	2	1	2	2		1	3
3	1	2	2	1	2	1	7	3	1	1	3	1	1	4	3	2	2	1	2	1	2	2	2	2	1	2	2	2	2	2	1	3	1	2		2	2	2	2	1
3	1	2	3	2	1	3	6	2	3	1	3	1	2	4	3	3	1	1	3	2	2	1	2	2	1	2	2	2	1	3	2	1	1	2	2	2	2	1	1	1
3	1	1	4	2	2	2	3	2		1	3	1	2	4	3	1	2	2	2	2	2	1	2	2	2	2	2	2		3	1	1	1	2	1	2	2	2	2	2

III. IMPUTATION USING MEAN VALUES

First calculating mean/median of the non-missing values in a column and then replacing the missing values with each column separately and independently from the others. This method is easy and fast but working well in numeric datasets only. It works on the column level only and not very accurate.

$$\begin{matrix} 1 & 2 & 2 & 1 \\ 2 & NaN & 1 & 1 \\ 4 & 1 & NaN & 2 \\ 2 & 1 & 5 & 1 \\ 3 & 1 & 7 & 2 \end{matrix} \quad \xrightarrow{\text{Mean}} \quad \begin{matrix} 1.0 & 2.0 & 2.0 & 1.0 \\ 2.0 & 1.2 & 1.0 & 1.0 \\ 4.0 & 1.0 & 3.7 & 2.0 \\ 2.0 & 1.0 & 5.0 & 1.0 \\ 3.0 & 1.0 & 7.0 & 2.0 \end{matrix}$$

IV. IMPUTATION USING SINGULAR VALUE DECOMPOSITION

Singular value decomposition (SVD) is a simple way for missing values imputation. In this method, computation is so easy and steps of a description for $x_{ij,}$ one missing value in X followed:

Step 1 : $\bar{U}$ and $\bar{V}$ are orthonormal matrices (i.e., $\bar{U}'\bar{U}$ = $\bar{U}\bar{U}'$=1). The i th case (row) is ignored from X and calculation the SVD of the remaining (n-1). Data matrix p, specified by X^{-i} = $\bar{U}\bar{D}\bar{V}'$ with $\bar{U}=\{\bar{u}_{st}\}$, $\bar{V}=\{\bar{v}_{st}\}$ and $\bar{D}=diag\{\bar{d}_1,\dots\dots,\bar{d}_p\}$.

Step 2: The j th variable (column) is ignored from X and calculation the SVD of the remaining n. Data matrix (p-1), specified by $X_{-j}=\tilde{U}\tilde{D}\tilde{V}$ with $\tilde{U}=\{\tilde{u}_{st}\}$, $\tilde{V}=\{\tilde{v}_{st}\}$ and $\tilde{D}=diag\{\tilde{d}_1,\dots,\tilde{d}_{p-1}\}$.

Step 3: For (i,j) th missing case imputation with

$$X_{ij}^{*} = \sum_{t=1}^{p-1}(\tilde{u}_{it}\,\tilde{d}_t^{1/2})(\bar{u}_{jt}\bar{d}_t^{-1/2})$$

Retrieval Number: A5204119119/2019©BEIESP
DOI: 10.35940/ijitee.A5204119119
Journal Website: www.ijitee.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

V. IMPUTATION USING KNN

The k-nearest neighbours algorithm is one of the simplest algorithm for classification. The main advantage of this approach is to predict both qualitative and quantitative attributes. Qualitative and quantitative attribute means for finding most frequent value and calculate mean among the k-nearest neighbours respectively. And also no need to create predictive model for each attribute of missing data. This approach can be easily adapted to work with any attribute.
K-nearest neighbour algorithm classifies a data point based on how its neighbours are classified. Here, all available cases are stored and also classified new cases as per similarity measure. This algorithm is based on **feature similarity:** choosing the right value of k is a process called parameter tuning and important for better accuracy.

$$\begin{matrix} 1 & 2 & 2 & 1 \\ 2 & \text{NaN} & 1 & 1 \\ 4 & 1 & \text{NaN} & 2 \\ 2 & 1 & 5 & 1 \\ 3 & 1 & 7 & 2 \end{matrix} \quad \xrightarrow{\text{Knnimpute()}} \quad \begin{matrix} 1 & 2 & 2 & 1 \\ 2 & 1 & 1 & 1 \\ 4 & 1 & 4 & 2 \\ 2 & 1 & 5 & 1 \\ 3 & 1 & 7 & 2 \end{matrix}$$

VI. IMPUTATION USING BAYESIAN PRINCIPAL COMPONENT ANALYSIS (BPCA)

This is an missing values estimation method which is based on Bayesian principal component analysis. This is the probabilistic model methodology and calculated the latent variables simultaneously within the framework of Bayesian inference. In terms of statistical methodology, the implementation that makes it possible to estimate arbitrary missing variables is new method.
This is a global based imputation method based on Eigen values. Here, some continuous hyper parameters are introduced to determine the latent space. BPCA required parameter optimization. A different Bayes algorithm is used to iteratively estimate the posterior distribution of the model parameters. And the missing values until convergence is reached.

VII. EVALUATION CRITERIA

Based on measures of performance, compared the above imputation methods. Generally Mean Square Error and Root Mean Square Error are used to evaluate the errors. MSE measures the average squared difference between imputed value and original values. RMSE represents the square root of mean square error.

$$MSE = \frac{\sum_{i=1}^{n} \left(X_i^{obs} - X_i^{imputed}\right)^2}{n}$$

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

where X^{obs} is the true value of predictors assigned as missing and $X^{imputed}$ is its imputed value; n represents the number of values in one subsample.

Table.3.Comparison of Evaluation Criteria

Methods	Mean	SVD	KNN	BPCA
MSE	3.8032	4.5976	4.7502	0.0244
RMSE	1.9502	2.1442	2.1795	0.1562

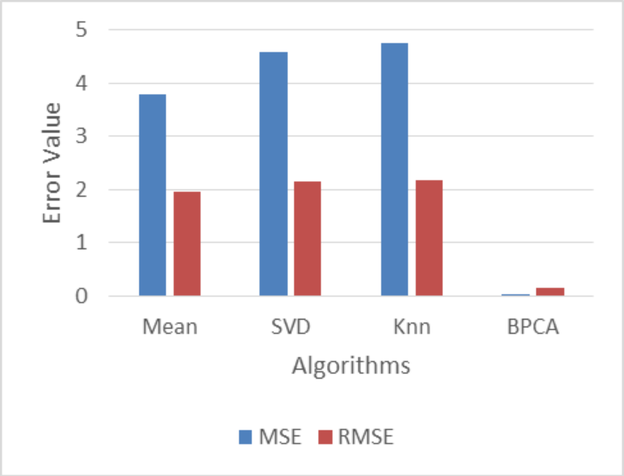


Fig.1.Comparison of Missing value imputation algorithms

VIII. CONCLUSION AND FUTURE WORK

Missing value is one of the challenge in the fields of data analysis. In this paper, discussed various missing data imputation techniques and also evaluated the performance using MSE & RMSE. While comparing the algorithms using the evaluation methods based on the real dataset, BPCA produced lower error rate than other methods. And also discovered that the BPCA algorithm outperforms the other methods, which is easy and common approach for missing data imputation. The graph is showed the comparative analysis clearly. In future, reduction technique will be applied for reducing the number of attributes in the real time datasets.

ACKNOWLEDGMENT

This research work has been supported by RUSA PHASE 2.0, Alagappa University, Karaikudi.

REFERENCES

1. K. Manimekalai and A. Kavitha, "Missing Value Imputation And Normalization Techniques in Myocardial Infarction", ICTACT Journal on Soft Computing,Vol.08 Issue:03,April 2018.
2. Peter Schmitt, J Biomet Biostat ,Jonas Mandel et al,"A Comparison of Six Methods for Missing Data Imputation", in Journal of Biometrics & Biostatistics, 2015.
3. V.B. Kambie and S.N. Deshmukh ,"Comparision between Accuracy and MSE, RMSE by using Proposed Method with Imputation Technique", Oriental Journal of Computer Science and Technology, ISSN: 0974-6471, Vol. 10, No. (4) 2017

4. Dr.Durairaj.M, Sivagowry.S , “A Pragmatic Approach of Preprocessing the Data Set for Heart Disease Prediction”, International Journal of Innovative Research in Computer and Communication Engineering, ISSN(Online): 2320-9801, Vol. 2, Issue 11, November 2014
5. Qinbao Song, Martin Shepperd et al, “Can k-NN Imputation Improve the Performance of C4.5 With Small Software Project Data Sets? A Comparative”, JSS ,December 2008.
6. Shigeyuki Oba, Masa-aki Sato, Ichiro Takemasa, MoritoMonden, Ken-ichi Matsubara and Shin Ishi, “A Bayesian missing value estimation method for gene expression profile data”, Bioinformatics, Vol. 19 no. 16, pages 2088–2096,2003.
7. Gustavo E. A. P. A. Batista and Maria Carolina ,“An Analysis of Four Missing Data Treatment Methods for Supervised Learning”, 2014.
8. Bjorn Grung , Rolf Manne ,“Missing values in principal component analysis”, Chemometrics and Intelligent Laboratory Systems,1998.
9. Julian Luengo, Salvador Garcia,Francisco ,“On the choice of the best imputation methods for missing values considering three groups of classification methods”, KnowlInfSyst, 2012.
10. Folch-Fortuny, F.Arteaga et al, “PCA model building with missing data: new proposals and a comparative study”, Chemometrics and Intelligent Laboratory Systems,2015.
11. Kristen A. Severson, Mark C. Molaro † and Richard D. Braatz,“Principal Component Analysis of Process Datasets with Missing Values”,MDPI Processes,2017.
12. Arjun Puri, Dr. Manoj Gupta, “Review on Missing Value Imputation Techniques in Data Mining”, International Conference on Machine Learning and Computational Intelligence-2017.
13. Dan Li, JitenderDeogun et al,“Towards Missing Data Imputation: A Study of Fuzzy K-means Clustering Method”, Springer-Verlag Berlin Heidelberg 2004.
14. SowmyaChandrasekaran et al, ”Data Preprocessing: A New Algorithm for Univariate Imputation Designed Specifically for Industrial Needs”, Technology Arts Science,Workshop Computational Intelligence, 2016.
15. Z. Mahesh Kumar , R. Manjula ,”Regression model approach to predict missing values in the Excel sheet databases”, International Journal of Computer Science & Engineering Technology (IJCSET), ISSN : 2229-3345, 2012.
16. Cosmin Lazar, Laurent Gatto, Myriam Ferro, Christophe Bruley and Thomas Burger, “Accounting for the Multiple Natures of Missing Values in Label-Free Quantitative Proteomics Data Sets to Compare Imputation Strategies”,American Chemical Soceity 2016.
17. Min-Wei Huang,Wei-Chao Lin,Chih-Wen Chen,Shih-Wen Ke,Chih-Fong Tsai,WilliamEberie,“Data preprocessing issues for incomplete medical datasets”,Wiley Online Library,2016.
18. R. Manimaran, Dr.M.Vanitha,”Novel Approach to Prediction of Diabetes using Classification Mining Algorithm “, International Journal of Innovative Research in Science, Engineering and Technology,July 2017.

AUTHORS PROFILE

Anitha.S.,obtained Master of Computer Science in 2004 from Alagappa University, Karaikudi, Tamilnadu.She obtained Master of philosophy in computer science in 2006 from Bharathidasan University,Tiruchirapalli,Tamilnadu.At present, she is pursuing her Ph.D in computer science ,Alagappa University Karaikudi.

Dr.Vanitha. M, M.Sc(OR & CA).,M.Sc.,M.Phil.,Ph.D (CS).Presently working as a Assistant professor in the Department of Computer Application, Alagappa University, Karaikudi. She has more than 10 years of experience in Research and nearly 10 years of experience in teaching. She has published more than 50 papers in international journals and acted as session Chairperson and reviewer.