International Conference on Machine Learning and Computational Intelligence-2017
International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 7 | ISSN : 2456-3307

Review on Missing Value Imputation Techniques in Data Mining

Arjun Puri, Dr. Manoj Gupta
Department of Computer Science and Engineering, Shri Mata Vaishno Devi University, Katra, Jammu and Kashmir, India

ABSTRACT

Now days, there are huge amount of data available for analysis, the main problem with the data is inconsistency. The inconsistent data (missing value) need to replace with most appropriate fit values. Some missing values are dependent on some known variable in the dataset need to be taken for further calculation. There are different methods to impute these missing values. In this paper, we discuss various technique based on their classification and also discuss their behavior in different datasets under different types of missing values.

Keywords : Missing value imputation, data mining, data preprocessing, Techniques for missing value imputation, MCAR, MAR, NMAR.

I. INTRODUCTION

In real world scenario, we are dealing with data and analysis of data. In order to deal with the extraction of knowledge from the given raw data, data mining is one of the important branch. There are many steps involved in the getting meaningful information from raw data. One of the important step is data preprocessing; the technique which helps in improving the quality of data and also improve mining results. One of the important issues in data preprocessing is missing value. It plays a vital role in deciding the computational results obtained by data preprocessing. Missing value can be caused from different sources like: sensor failure, corrupted datasets, incomplete survey etc. (Irfan Pratama, 2016). Inconsistence of data (missed values) is of different types, some of them are discussed below:

1. Missing completely at random (**MCAR**), if there is no dependency in the missing data is related to its known values. In this type of missing data we assume that a whole distribution of data is completely missed.
2. Missing at random (**MAR),** when the missing value depend on the already known value and does not depend upon missed value itself.
3. Not missing at random (**NMAR**), when missed value does not depend upon any given or missed value. [(Irfan Pratama, 2016), (Shichao Zhang, 2011), (Julián Luengo, 2012)]

These types of anomalies generally arise due to different sources like: MCAR can arise due to sensor recording failure because no data is dependence in between them whereas, MAR can arise during the survey question some question are not answered by the people but there are other questions related with them (Irfan Pratama, 2016).

In order to deal with the missing values there are many techniques developed so for, some of them usually ignore the missing values and some of them delete and some techniques use imputation. Broadly speaking, these techniques divided into two main types: conventional techniques (like mean, mode, median and deleting values) and modern techniques (hot deck, cold deck (Geeta Chhabra, 2017),

classification techniques like SVM)[(Alireza Farhangfara, 2008), (Julián Luengo, 2012)]. In this research paper, we survey some of the techniques to deal with missing value imputation and compare them in contrast in the following sections: literature survey, methods deal with missing value imputation, discussion of different techniques on different datasets and conclusion.

II. LITERATURE SURVEY

Various researcher compare different techniques on different datasets analyze their outputs and also suggest that which technique is suitable for which dataset [(Alireza Farhangfara, 2008) (Geeta Chhabra, 2017) (Julián Luengo, 2012) (Schmitt P, 2015) (Irfan Pratama, 2016). Some other researchers develop technique to improve accuracy in imputation of values.

In Alireza Farhangfara et al(2008), in which a compartative study was made which includes six single and multiple imputation methods on 15 discrete incomplete datasets. In this paper researcher find that imputation improves by using classification techniques, except for the mean imputation method which shows poor results with high rate of missing values (50%). In this paper,researcher conclude that Naive-Bayes based imputation shows better result by using RIPPER classification on datasets with high amount of missing values, i.e. 40% and 50%. Researcher also show that multiple imputation polytomous regression method shows best result with SVM on different datasets. Finally, shows that the mean imputation is least beneficial.

In Sasi et al.(2016), an intelligent approach was suggested by the authors to deal with data of different types. In this authors take 3 different datasets(name: iris, credit and adult) and preform missing value imputation by using different approaches(name: Mean/Mode, K-Nearest Neighbor, Hot-Deck, Expectation Maximization, and C5.0 imputation techniques) by using and IITMV menthod which decide which dataset missing values and operated by which technique. In this paper authors, concluded that IITMV technique shows better results compared with C5.0 algorithm.

In Esther-Lydia Silva-Ramírez et al(2011), A methodology for data imputation by means of artificial neural networks has been proposed and empirically compared with three classic methods: mean/mode imputation, regression models and hot-deck. Fifteen datasets are used for elvaluation and observed that multilayer perceptron provide better results.

In Irene Erlyn Wina Rachmawan et al(2015), An algorithm from machine learning for missing value called Reinforcement Programming was proposed. Reinforcement programming shows better result as compared with zero imputation, Mean imputation and Genetic Algorithm.During evaluation researcher find that Reinforcement Programming could run better in solving Missing imputation.

In Schmitt.P et al(2015), a comparison of six imputation methods(Mean, KNN, SVD, maximization expectation imputation, bPCA and MICE) based on four real datasets (iris,e.coli,breast cancer1,breast cancer 2) of various sizes, under an MCAR assumption with missing values ratio percentage of missing values (from 5% to 45%increased by 10%). By using different evaluation techniques authors identifed performance of different techniques : Root mean squared error (RMSE), unsupervised classifcation error (UCE), supervised classifcation error (SCE) and execution time. While much attention has been paid to the imputation accuracy measured by RMSE. Results shows that, MICE technique is complex in structure and show better results in case of small datasets. While bPCA and FKM shows better results with large datasets.

In Ibrahim Berkan Aydilek et al (2013), a hybrid method was proposed by the resaercher by using support vector regression and genetic algorithm with fuzzy clustering to estimate missing values. A Complete train data were clustered based on their similarity, and fuzzy principles were used during clustering. Therefore, each missing value becomes a member of more than one cluster centroids, which yields more sensible imputation results. Six datasets with different characteristics and also with differnet missing value ratios were used in this paper, and resulted showed that better result obtained as compared to other techniques.

Some of the techiques for missing values imputation discussed during literature survey are as under.

2.1 CLASSFICATION OF MISSING VALUE TECHNIQUES

There are so many approach and techniques developed so far to deal with missing values. Researchers develop many techniques ranges from simple to complex. Researcher generally makes division but these techniques some of them deal with low missing values and some of them deal with higher missing values. These techniques are discussed as under:-

1. **Mean imputation:** In this technique, mean of missing value is calculated by using the corresponding attribute value. This technique is faster over other techniques. It shows good result when data is small, but result is not good for big data. This model is helpful for only MAR but not useful for MCAR [(Irfan Pratama, 2016), (Jason Van Hulse, 2008), (Sasi, 2016)].
2. **Hot deck imputation:** this method is used for categorical data and it is beneficial for big data and not for small data. In this method, missed valued is replaced by the most similar values of that attribute, this method becomes problematic when there is no other similar data is available [(Irfan Pratama, 2016), (Alireza Farhangfara, 2008), (Sasi, 2016) (Esther-Lydia Silva-Ramírez, 2011)].
3. **K-nearest Neighbor imputation (KNN):** This technique used Euclidean distance to determine the similarity between two values and replace the missing one with similar one .The main benefits of this approach are as given as under:
   ·KNN is useful for datasets having both qualitative and quantitative attribute values.
   ·There is no need for creating a predictive model for each attribute of missing data and helpful for multiple missing values.
   The main drawback of the KNN approach is that, whenever the KNN looks for the most similar instances, the algorithm searches through all of the data set (Sasi, 2016).
4. **Regression Imputation:** This technique is applied by using known values for the construction of model and calculates the regression between variables and then applied that model to calculate the missing values. This technique gives more accurate results than mean imputation (Jason Van Hulse, 2008).
5. **REPTree imputation:** REPTree is a decision tree used for the analysis of independent variables in comparison with quantitative dependent variables. In this process recursive technique are applied to complete the incomplete dataset with least error by using reduced error pruning by using variance. (Jason Van Hulse, 2008).
6. **Support Vector Regression:** This method is extension of Support vector machine. In Support vector machine generally missing values are ignored first and then rest of data is feed to train the system and then missing values are filled with the trained system (Irfan Pratama, 2016). By using regression with support vector machine classifier efficiency will increase (Alireza Farhangfara, 2008).

7. **Fuzzy mean imputation:** It is the technique which uses fuzzy in the calculation of missing value with the help of clustering in the known value and finding which missing value belong to which cluster there are two different ways to calculate fuzzy mean one is K-mean and other is C-mean. C-mean is better than K-mean in most of cases (Schmitt P, 2015).
8. **Reinforcement Programming:** It is generally used for dynamic approach for the calculation of missing values by using machine learing approaches. It has capability of convergence and to solving imputation problem by using exploration and exploitation (Irene Erlyn Wina Rachmawan, 2015).
9. **Nonparametric Iterative Imputation algorithm (NIIA):** It is an iteratively imputing the missing values in a dataset. It works as follows:
   Identify some missing values and then compute the values of all complete values used to estimate these incomplete values. Then these missed value imputed is used for further analysis of other incomplete instances and the process repeat until the values of dataset are completely filled. (Shichao Zhang, 2011)
10. **Multilayer Perceptrons:** Multilayer perceptrons is the technique to develop by using artificial neural networks. It runs as on multilayer and also use different learning processes to train the network (Esther-Lydia Silva-Ramírez, 2011).

III. DISCUSSION

In this paper we review different techniques and with different dataset and analysis that which technique is giving best result in which type of dataset. This collaborative information is represented with the help of following table.

Table 1. List of various papers on missing value imputation techniques.

Research paper	Datasets	Techniques	Remarks
Geeta Chhabra et. al. (2017)	Iris	1. Predictive Mean Matching 2. Multiple Random Forest Regression Imputation. 3. Multiple Bayesian Regression Imputation 4. Multiple Classification and Regression Tree (CART). 5. Multiple Linear Regression using Non-Bayesian Imputation. 6. Multiple Linear Regression with Bootstrap Imputation.	A comparison of diff erent approaches of MICE methods on iris datasets. Efficiency gain with Multiple Imputation combined with Bayesian Regression is that it is able to make better use of the available information by accommodating non linearities among the predictors.
Ibrahim Berkan Aydilek et al (2013)	1. Iris 2. Haberman 3. Glass 4. Musk1 5. Wine 6. Yeast	1. SvrFcmGa (proposed) 2. FcmGa 3. SvrGa 4. ZeroImpute	Dataset inconsistence can be ranged from 10% to 25% and analyze that SVRFCMGA (Fuzzy C-mean with Support vector Regression and Genetic algorithm) perform better than other.

Sasi et al. (2016)	1. Iris 2. Credits 3. Adults	1. Mean/Mode. 2. Hot Deck. 3. Expectation Maximization. 4. K neighbor nearest.	In this paper, authors compare C5.0 with this new developed technique known as IITMV and also show its performance on different data sets.
Esther-Lydia Silva-Ramírez et al. (2011)	1.Cleveland 2.Heart 3. Zoo 4. Buhl1-300 5.Glass 6.Ionosphere 7.Iris 8.Pima 9. Sonar 10.WaveForm2 11.Wine 12.Hayes-Roth 13. Led7 14.Solar 15. Soybean	1. mean/mode 2. Regression. 3. Hot deck. 4. ANN.	Result shows that Multilayer perceptrons (MLP) with different learning rules show better results with quantitative datasets as compared with classical imputation methods. In this paper, type of missing value is missing completely at random (MCAR) is taken.
Schmitt P et al.(2015)	1. Iris 2. E. coli 3. Breast cancer 1 4. Breast cancer 2	1. Mean 2. K-nearest neighbors(KNN) 3. Fuzzy K-means (FKM) 4.Singular value decomposition(SVD) 5.Bayesian principal component analysis (bPCA) 6.Multiple imputations by chained equations (MICE).	Results show that different techniques are best at different datasets and different size. MICE is useful for small datasets but for big datasets bPCA and FKM are better one.

In the above table, different researchers compare different techniques on the bases of RSME and calculate difference between correct dataset with incorrect datasets, also predict the efficiency of particular techniques. In Geeta Chhabra et al(2017), discuss various techniques regarding MICE and concluded that is with Multiple Imputation combined with Bayesian Regression gives better efficiency than other techniques, where as in Schmitt P et al.(2015), compare different techniques and concluded that MICE technique are useful for small dataset error replacement. In Ibrahim Berkan Aydilek et al(2013), research made a comparsion between different hybrid techniques on different datasetswith change in missing values (ranges from 10% to 25%), and concluded that SvrFCmGA gives better preformance than other techniques (FcmGa, SvrGa, ZeroImpute). In Schmitt P et al. (2015), for big datasets bPCA and Fuzzy k mean gives better result. In Sasi et al. (2016), author compute different types of datasets on different techniques and gives classification of dataset that which technique suits what kind of datasets and also proposed and test his technique with C5.0 technique.Now a days, the development of new method is done with combining different techniques together.

IV. CONCLUSION AND FUTURE WORK

Missing value is one of the challenge in the fields of data analysis. In this paper, we discussed various techniques dealing with the imputation depending on different datasets and different missing value type (MCAR, MAR) and study the behaviour of different techniques with different percentage of missing values (10%,20%,40%, etc.), find out that there is no such one technique to deal with all datasets. In study, we reach over the conclusion that many research are trying to combine many techniques together to implement intelligently on different datasets and uses a decision algorithm to pick one out of them.

In future work, we need to develop techniques for unclassified datasets (such as, estate estimation problem for nonlinear stochastic timedelay systems with missing measurements) having better efficiency and accuarcy. Moreover, while analysing we found that there is need of intelligent system which make decision regarding which techniques is suitable for which type of datasets.

V. REFERENCES

[1]. Alireza Farhangfara, L. K. (2008). Impact of imputation of missing values on classification error for discrete data. Pattern Recognition , 3692-3705.

[2]. Esther-Lydia Silva-Ramírez, R. P.-M.-C.-D.-d.-l.-V. (2011). Missing value imputation on missing completely at random data using multilayer perceptrons. Neural Networks , 121–129.

[3]. Geeta Chhabra, V. V. (2017). A Comparison of Multiple Imputation Methods for data with Missing Values. Indain Journal of Science and Technology , 1-7.

[4]. Ibrahim Berkan Aydilek, A. A. (2013). A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm. Information Sciences , 25–35.

[5]. Irene Erlyn Wina Rachmawan, A. R. (2015). Optimization of Missing Value Imputation using Reinforcement Programming . International Electronics Symposium (IES), (pp. 128-133).

[6]. Irfan Pratama, A. E. (2016). A Review of Missing Values Handling Methods on Time-Series Data. International Conference on Information Technology Systems and Innovation (ICITSI) (p. 6). Bandung-Bali : IEEE.

[7]. Jason Van Hulse, T. M. (2008). A comprehensive empirical evaluation of missing value imputation in noisy software measurement data. Journal of System and Software , 691-708.

[8]. Julián Luengo, S. G. (2012). On the choice of the best imputation methods for missingvalues considering three groups of classification methods. , Knowledge Information System , 77–108.

[9]. Sasi, T. A. (2016). Intelligent Imputation Technique for Missing Values . International Conference on Advances in Computing, Communications and Informatics (ICACCI), (pp. 2441-2445). Jaipur, India.

[10]. Schmitt P, M. J. (2015). A Comparison of Six Methods for Missing Data Imputation. Journal of Biometrics and Biostatistics , 2-6.

[11]. Shichao Zhang, Z. J. (2011). Missing data imputation by utilizing information within incomplete instances. The Journal of Systems and Software , 452–459.