A literature survey on missing value imputation methods in data mining

Mrs. J. Sujitha, Mrs. S.R. Lavanya
'Research Scholar, Department of Computer Science, Sri Ramakrishna College of Arts and Science for Women, Coimbatore, Tamil Nadu, India'
'Assistant Professor, Department of Computer Science, Sri Ramakrishna College of Arts and Science for Women, Coimbatore, Tamil Nadu, India',

Abstract: *Preprocessing is one decisive stage in data mining. The most important step in pre-processing is handling missing data. Managing missing data has different strategies such as deletion of incomplete data and imputation (filling) of missing values through depends on statistical and machine learning (ML) procedures. Most of the data mining algorithm cannot work with an incomplete dataset. In such a case, missing values create a problem with real data in the world. Missing value Imputation is a challengeable task in many sectors. Some of them perform their handling imputations in different ways. Missing value imputation determines the quality of datasets. A high accuracy data analysis is essential because it deals with real-world decision making. It performs different kinds of data mining operations. This survey analyses the various missing value imputation methods and its performances. It consists of imputation methods like single and multiple imputations. The imputation algorithms like mean, K-nearest neighbor, Naive Bayes classification, MIAEC are discussed in this survey.*

Keywords: *Missing value, Imputation methods, Mean, K-nearest neighbor, Naive Bayes, Chain of evidence.*

I. INTRODUCTION

Preprocessing is one of the main works in data mining projects. Missing value imputation is an important task in preprocessing. The main problem of a dataset with some missing values leads to an inappropriate result. To overcome the problem, it is necessary to impute the missing values in the dataset. Generally, the problem of missing data emerges in many areas of research such as statistical, environmental, medical, industrial fields, etc. The empty or unanswered values in datasets are called missing value (data). Little and Rubin describe a list of missing mechanisms. The missing data mechanisms classify into two types such as ignorable and non–ignorable. Non-ignorable is where the probability of missing data depends on the value of observation.

Ignorable missing data is where the probability of missing data does not depend on the value of observation [2]. The missing data mechanisms explain the connection between missing value and its variables in the data matrix [2]. Schafer explains given an experimental variable L as L_{obs} and a missing variable L as L_{mis}, it can be said that L= [(L_{obs},L_{mis})][3].The following diagram shows the missing data types,

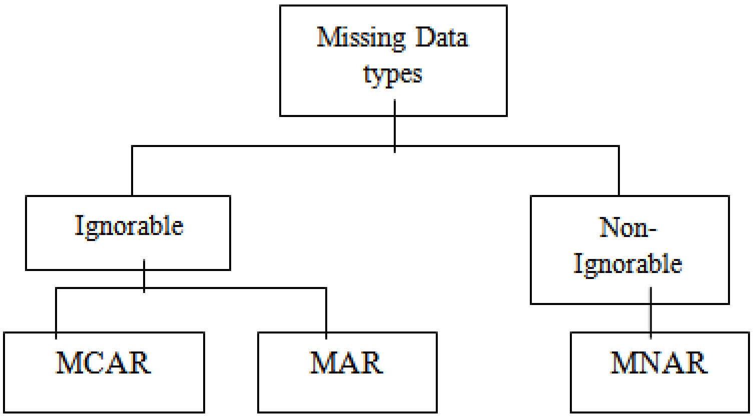

Fig 1: Missing data types

The mechanisms are following,

Missing completely at random (MCAR):

The first type of mechanism is MCAR. It defined missing value occurs at random across the whole data sets. Thus the possibility of missing value is independent of both L_{obs} and L_{mis} [3].

Missing at random (MAR):

The second type of mechanism is MAR. These types of missing data occur if the probability of a record having a missing value for an attribute that does not depend on the value of the missing data itself, but could depend on the observed data [6]. Thus, the possibility of missing value is independent of L_{mis}[3].

Not missing at random (NMAR):

The third type of mechanism is non-ignorable missing at random (MNAR). It occurs if the probability of a record having a missing value for an attribute that depends on the value of the attribute. If missing data are MNAR, valuable information lost from the data and, there is no common method of solving the missing data properly [4].

II. METHODOLOGIES OF MISSING DATA IMPUTATIONS

Jaiwei Han, et al analyzes the preprocessing topics in their third edition book. Imagine that we need to analyze sales and customer data in a company. Note that many rows or columns in a table have no recorded value for several attributes such as customer income. Then how can go about filling in the missing values for this attribute? Look at the following methods.

- ***Avoid the data**:* This actually appears when the data label is missing (assume this mining work involves classification). This method is not effective unless the rows contain several attributes with missing values.
- ***Fill in the missing value yourself:*** This method is time-consuming and may not be feasible given a large data set with more missing values [1].
- ***Global constant use to fill in the missing value*****:** Replace all missing attribute values by the same constant such as a label like 'unknown' or 'null' [1].
- ***Find a central tendency for an attribute to fill in the missing value:*** The Measure of central tendency specify the "middle" value of a data distribution. The mean or median method is used to find the missing values [1].
- ***Use the most possible value to fill in the missing value:*** This may be determined with regression, inference-based tools using a Bayesian formalism, or decision tree induction [1].

The following table has a list of imputation methods

Single imputation	**Multiple imputation**
1. Hot-deck 2. Cold-deck 3. Mean/Mode substitution 4. Regression 5. Group mean 6. Last value carried Forward (LVCF) 7. KNN 8. Naive Bayes	1. Predictive model 2. EM Algorithm 3. Multiple Linear Regression 4. Model based multiple imputation 5. Markov chain Monte Carlo (MCMC)

Table 1: List of Imputations algorithms

Missing value imputation methods are divided into two categories. They are single imputation and multiple imputations.

Mode imputation:

The Mode is the easiest behavior in the case of categorical data is to fill in each missing value with the sample mode. The major disadvantage of mode imputation is that it creates points in the delivery by focused all the imputed values in the mode, as a result, the variance is reduced artificially (Kalton and Kish, 1981). This is a single imputation method since only one value is replacing each missing observation.

Mean imputation:

Mean imputation method is one of the most repeatedly used methods. It includes of exchange the missing data for a given component or attribute by the mean of all known values of that attribute in the class where the instance with missing attribute fit in [16].

K-Nearest Neighbor imputation (KNN):

KNN algorithm is a simple classification algorithm and it is one of the most used learning algorithms.KNN is a non-parametric, lazy learning algorithm. The missing values of this method are an attribute is imputed using the known number of attributes that are mainly related to the attribute whose values are missing. The comparison of the two attributes is determined using a distance function. The N parameter is a number of interested neighbors which generally would be self-determined. Distance measurement accomplishes by several techniques, namely Mahalanobis Distance, Euclidean Distance, and Manhattan Distance. These techniques are parallel neighbors that must apply normalization data. As for Euclidean Distance, it is not required to conduct normalization data [15]. K-Nearest Neighbor imputation has some advantages and disadvantages.

- ***Advantages:*** 1) K-nearest neighbor can calculate both qualitative & quantitative attributes.2) Formation of a predictive model for each attribute with missing data is not required, 3) Attributes with multiple missing values can be easily treated parallel structure of the data is taken into consideration.
- ***Disadvantage:*** 1) KNN algorithm is very time-consuming in analyzing the huge database. It searches through all the dataset looking for the most similar instances.2) Selection of k-value is very significant. A higher value of k would include attributes which are significantly different from what we need whereas a lower value of k implies missing out of significant attributes [15].

K-Means clustering imputation:

K-Means is to categorize or to cluster the objects based on attributes/features into k number of the group. The grouping is finished by minimizing the sum of squares of distances between data and the corresponding cluster centroid. It offered a rapid and specific way of estimating missing values [17].

Fuzzy K-Means clustering imputation (FKMI):

Here membership functions acting an important role. The Membership function is allocated with each data object that represented in what degree the data objects as belonging to the particular cluster. Data objects would not get allotted to a concrete cluster which is indicated by the centroid of a cluster (as in the case of K means), this is because of the various membership degrees of every data with entire K clusters [17].

Regression imputation:

Regression imputation has regression models to calculate missing values. Many forms of regression models can be used for regression imputation such as linear regression, logistic regression, and semi parametric regression. Here, predicted values are used for filling Missing values [14].

Multiple imputations:

The imputed values are described from a distribution, so they basically enclose some variation. Thus, multiple imputations (MI) explain the limitations of single imputation by an additional form of error based on variation in the parameter estimates across the imputation, which is called between imputation errors [14].

EM imputation:

It uses the iterative procedure of the EM algorithm to calculate the sufficient statistics and estimate the parameters [14]. EM Algorithm, which stands for Expectation-Maximization. It is an iterative process in which it uses other variables to impute a value (probability), then checks whether that is the value most likely (Maximization). If not, it re-imputes a more likely value. It repeated until it reaches the most likely value.

EM imputations are superior to mean imputations because they maintain the relationship with other variables, which is important if you go on to, use something like Factor analysis or linear regression.

III. LITERATURE SURVEY

M.N. Norazian Ramli, et al analyzed imputation methods like Single and multiple imputations. Single imputation methods work in short gap length of missing data. Embracing a single imputation method to the long gap of missing data will cause systematically error since the reflection of uncertainty is not covered. Multiple imputations were recognized as the superior method for a missing-at-random data set. It is a comparative study paper of single and multiple imputation techniques [5]. To perform complete data by filling in missing value divided by single imputation and multiple imputation methods. Single imputation is defined as filling in accurately one value for each missing one and multiple imputations are defined as generating multiple fake values for each missing item in order to reflect properly the uncertainty attached to missing data [6].

According to Mertler, C. A. et al, the basic analysis of the missing value mechanism is a prediction. Quantitative values are applied in these prediction and imputation process. There are three major predictions methods to make imputation that is, prior knowledge, regression and average (mean) value imputation [7]. Prior knowledge is used to an imputation of new values into missing values based on previous knowledge [7]. The second one is mean or average value imputation; this method is used to calculate the mean by data obtained and imputing those values for missing values. It is the best way estimation if the researcher does not have other information. The third one is a regression. Here, one or more independent variables are taken and develop an equation. It can be used in imputing the dependent variable's value. A variable that has missing values in the missing value estimation process is leading the dependent variable. Focus that has entire data is used to develop this estimation equation. Once the equation is found, it is used to approximate the missing values independent variable for focus that has missing data [8].

R.S. Somasundaram, et al achieved their work in the following techniques. To calculate the quality of imputation, the imputed data is clustered with fuzzy C means clustering algorithm and the performance of classification is measured with different quality metrics. FC-means was selected to evaluate the imputation performance because in our previous work was observed that FC-means provided better performance. Fuzzy c-means (FCM) is a data clustering method in which each data point belongs to a cluster to a number of degrees that is specified by a membership grade. This technique was originally introduced by Jim Bezdek in 1981 as an improvement on earlier clustering methods [10]. The performance has been measured with respect to different rate or a different percentage of missing values in the data set. To evaluate the performance, the standard WDBC data set has been used [10].

Xiaolong Xu, et al examined several missing data imputation methods. Those methods can be separated into two types based on the probability of statistical analysis and data mining [9]. Linear regression (LR) and Expectation-maximization (EM) are the mainly frequent methods based on probability. For LR and EM parameter methods, if the data allocation of the dataset being processed is not well understood, this will outcome in an estimate of the variation. However, in real life, we have a limited perceptive of the dataset to be processed. If we compare the data distribution of the dataset, choose the correct parameters, and then the missing value of the imputation outcome is quite good. But like the EM method even for the dataset of data allocation to understand, the parameter union is very slow and time-consuming. Bayesian classification is also a better algorithm for imputation of missing data [11]. Naive Bayes classification results can be compared with the decision tree algorithm and neural network classifier [12]. The naive Bayesian classification algorithm used for imputation of the missing value is based on the Bayesian theorem formula to estimate the value of missing data [11].

Xiaolong Xu performs the missing data imputation by using the algorithm named MIAEC. The core task of the algorithm is to calculate the reliability of each estimated missing value in the chain of evidence. Thus, gives the sum of the confidence of all the evidence for the estimated missing data, the maximum estimate of the sum of the confidence values is selected as the imputation value [9]. The experimental data in this works are from the real dataset of UCI. Get the data in this work from American Census database, completed by Barrr Becke et al, which includes 15 different attributes of people's age, work, weight, education, marital status, occupation, etc. This work mainly analyzes the algorithm from two aspects: accuracy of imputation and speedup.

In an experiment, it has five discrete attributes selected: occupation, education, sex, work class, race. The value of these attributes is randomly removed and datasets with different missing rates are obtained. The missing rates were 10%, 20%, 30%, respectively. The following table described the imputation accuracy of missing data under different algorithms [9].

Accuracy				
Missing rates	Mode	KNN	Naïve Bayes	MIAEC
10%	58	59	59	62
20%	58	57	56	61
30%	58	46	53	61

Table 2: Imputation accuracy of mode, KNN, Naïve Bayes and MIAEC

In experimental results, the imputation accuracy of MIAEC is better than other algorithms.

P. Keerin, et al analyzes cluster based KNN missing value imputation. KNN has very high classification accuracy. This work experiments established that the K-nearest neighbor algorithm performs well in the absence of DNA microarray data [13]. DNA microarray is a popular high-throughput technology for the monitoring of thousands of gene expression levels simultaneously under different conditions. The usual purposes of microarray studies are to identify likewise expressed genes under various cell conditions and combine the genes with cellular functions [14]. However, the KNN algorithm has its own failing, when the dataset has a large percentage of missing data, KNN accuracy will be reduced.

IV. CONCLUSION

The accuracy of missing data imputation is different from one task to other tasks. KNN algorithm is one of the famous classifiers for grouping up of data. KNN algorithm has its own weakness when the dataset has a large percentage of missing data, KNN accuracy will be reduced. The established technique such us mean/ median and a standard deviation is responsive to recover the performance of accuracy in missing data imputation. K-Means clustering imputation conveys a rapid and exact way of estimating missing values. MIAEC algorithm performs consistency of all estimated missing value in the chain of evidence. Here, filling missing value done by a sum of confidence of all the evidence. The imputation accuracy of MIAEC is better than other algorithms. But it has a small dataset and little missing values. If the dataset is large, the imputation performance may give an insufficient result and accuracy of imputation will be decreasing. To solve the particular problem, need to find the perfect task to improve accuracy.

REFERENCES

[1]. Jaiwei Han, Micheline Kamber, Jian Pei,"Data mining Concepts and Techniques", Third Edition, Reprinted 2015.

[2]. R. J. Little and D. B. Rubin, "Statistical Analysis with Missing Data.John Wiley and Sons", New York, 1997.

[3]. Schafer, J.L., 1997,"Analysis of incomplete multivariate data.Monographs on Staistics and AppliedProbability", No. 72. Chapman and Hall, London.

[4]. Donders, A.R.T., G.J.M.G. van der Heijden, T. Stijnen, K.G.M. Moons, "Review: A gentleintroduction to imputation of missing values",Journal of Clinical Epidemiology, 59: 1087-1091, 2006.

[5]. M.N. Norazian Ramli, Yahaya, A.S., Ramli, N.A., Yusof, N.F.F.M., Abdullah, M.M.A. "Roles of Imputation Methods for Filling the Missing Values: A Review", AENSI Journals Advances in Environmental Biology, Special Issue for International Conference of Advanced Materials Engineering and Technology (ICAMET 2013), 28-29 November 2013, Bandung Indonesia.

[6]. Junninen, H., H. Niska, K. Tuppurainen, J. Ruuskanen, M. Kolehmainen," Methods for imputation of missing values in air quality data sets", Atmospheric Environment, 38: 2895-2907, 2004.

[7]. Mertler, C. A., & Vannatta, R. A., "Advanced and multivariate statistical methods: Practical application and interpretation (3th ed.)", Glendale, CA: Pyrczak Publishing, 2005.

[8]. Tabachnick, B., & Fidell, L, "*Using multivariate statistics* (3th ed.)", New York: Herper Collins College Publishers, 1996.

[9]. Xiaolong Xu, Weizhi chong, Shancang Li, Abdullahi Arabo, Jianyu Xiao,"MIAEC: Missing Data Imputation Based on the Evidence Chain", IEEE Access, current version March 19, 2018.

[10]. R.S. Somasundaram, R. Nedunchezhian, "Missing Value Imputation using Refined Mean Substitution", IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 4, No 3, July 2012.

[11]. D.H.Yang, N. N. Li, H.-Z.Wang, J. Z, Zhao, and H. Gao, ``The optimization of the big data cleaning based on task merging," Chin. J. Comput.,vol. 39, no. 1, pp. 97_108, 2016.

[12]. M. Zhu and X. B. Cheng, ``Iterative KNN imputation based on GRA for missing values in TPLMS," in Proc. 4th Int. Conf. Comput. Sci. Netw.Technol (ICCSNT), Harbin, China, 2015, pp. 94_99.

[13]. P. Keerin, W. Kurutach, and T. Boongoen, ``Cluster-based KNN missing value imputation for DNA microarray data," in Proc. IEEE Int. Conf. Syst.,Man, (SMC), Seoul, South Korea, Oct. 2012, pp. 445_450.

[14]. Dudoit S, Yang YH, Callow MJ, Speed TP, " Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments", Statistica Sinica,12:111–139, 2002.

[15]. Jarumon Nookhong, Nutthapat Kaewrattanapat,

[16]. "Efficiency Comparison of Data Mining Techniques for Missing-Value Imputation", Journal of Industrial and Intelligent Information, Vol. 3, No. 4, December 2015.

[17]. Swatijain, Mrs. Kalpana Jain, Dr. Naveen Choudry, "A Survey Paper on missing data in data mining", International Journal of Innovations in Engeneering Research and Technology [IJIERT], ISSN: 2394-3696,VOLUME 3, ISSUE 12, Dec-2016.

[18]. N. Poolsawad L. Moore C. Kambhampati and J. G. F. Cleland, "Handling MissingValues in Data Mining - A Case Study of Heart Failure Dataset", l9th InternationalConference on Fuzzy Systems and Knowledge Discovery, 2012.

[19]. A.Rogier T.Donders, Geert J.M.G Vander Heljden, Theo St ijnen, Kernel G.M Moons, "Review: A gentle introduction to imputation of missing values", Journal of Clinical Epidemiology,59:1087-1091, 2012.

[20]. Kin Wagstaff, "Clustering with Missing Values: No Imputation Required", NSF grant IIS-0325329:1-10.