Missing Value Imputation Methods and Algorithms Evaluation

Hiteshkumar Patel[1] Mr. Nimesh Patel[2]
[1]ME Scholar [2]Assistant Professor
[1,2]Department of Computer Engineering
[1,2]LDPR Institute of Technology and Research, Gandhinagar, India

Abstract— Data mining has achieved spectacular success in practically every sector, including finance and banking, retail sector, insurance and healthcare, scientific analysis, telecommunication industry, research and so on. Since, real-world data are highly susceptible to noisy, missing, and inconsistent data due to their typically huge size and their likely origin from multiple, heterogenous sources. Low-quality data will lead to low-quality mining results. To improve the quality of data number of data preprocessing techniques are applied to clean the data and as consequence data mining results. One of the major irritating issues with real world data is missing values (MVs). Attributing missing values of data improves classification accuracy. This paper presents many methods and techniques for dealing with missing data. The paper also sheds insight on potential limitations and research needs. According to the results of a survey, the most commonly utilized algorithm performance indicators are assumptions, accuracy and time complexity.

Key words: Data Mining; Imputation; Missing Values; Machine Learning

I. INTRODUCTION

Missing data is one of the most vexing parts of handling data collections of varying sizes. There are several reasons why data may be absent from data sets. Some of the most prevalent explanations are as follows:

- Data is integrated from many sources: one set of data did not catch some value, while another set did not contain another value. As a result, there will be data gaps.
- Attributes of interest may not always be available
- Some values may not be included simply because it was not considered important at the time of entry
- Data acquired first, chronologically, may lack attributes obtained later.
- Relevant data may be not included due to a misunderstanding or lost due to device failure.
- Data that were inconsistent with other recorded data may have been deleted.
- Data may not be collected due to technological limitations.
- Data are not gathered properly because data entry problems or due to data transmission problems.
- Data is not collected due to ethical issue like data pertaining to religion, race, and ethnicity should no longer be collected.

II. METHODS TO DEAL WITH MISSING VALUES

Mainly there are four frequent missing valuetechniques for handling missing values: [2]

1) Records removal: If a missing data emerges on any of the components towards the data, discard the entire perspective. usually done when class label is missing. But, not effective when the percentage of missing values per attribute varies considerably.
2) Fill in the missing value manually: This technique is time consuming and impractical for a large data collection with a large number of missing values.
3) Use a global constant: replace any missing attribute data with an equivalent consistent, such as "unidentified" or “unknown” or “-∞“.
4) Use the attribute mean: Impute in attribute missing value comparisons. Replace with the central tendency calculation such as Mean, Mode, Median, and Midrange, (Max + Min)/2.

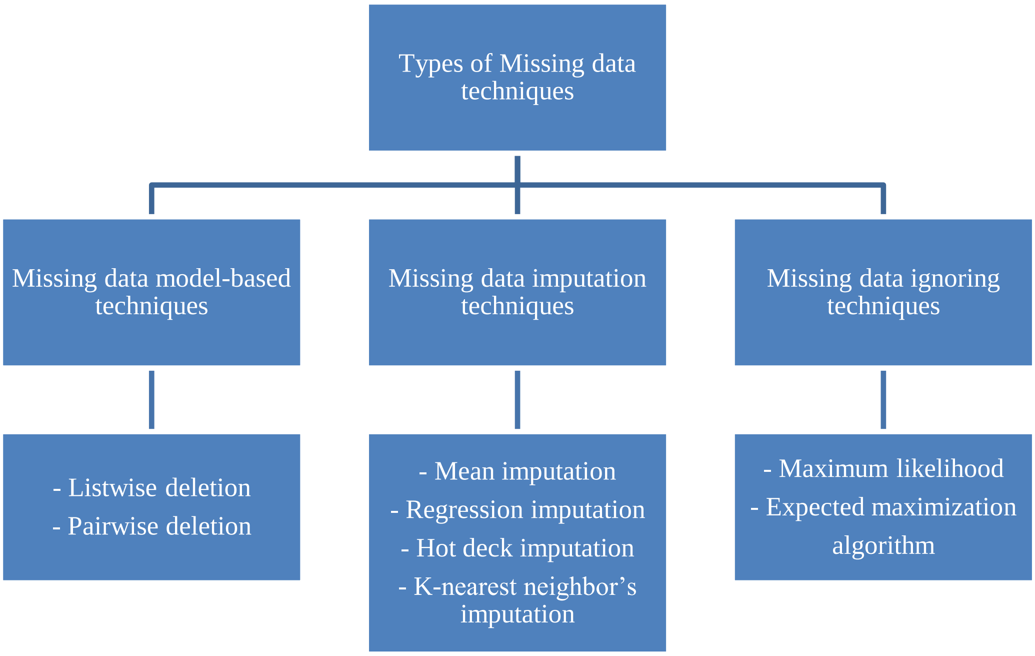
Types of Missing data techniques
Missing data model-based techniques
Missing data imputation techniques
Missing data ignoring techniques
- Listwise deletion
- Pairwise deletion
- Mean imputation
- Regression imputation
- Hot deck imputation
- K-nearest neighbor's imputation
- Maximum likelihood
- Expected maximization algorithm

Fig. 1: Different types of missing data methods

III. MISSING VALUE REMOVING METHODS

1) LD ("Listwise Deletion"): If an item has a missing value for a couple of the criteria, keep a strategic space from that point in the research. In statistical bundles, it is usually the evasion.
2) PD ("Pairwise Deletion"): Pairwise deletion (available-case analysis) seeks to reduce the loss caused by listwise deletion. Consider a correlation matrix to see how pairwise deletion works. The strength of the link between two variables is measured by a correlation. The correlation coefficient will take into consideration each pair of variables for which data is available. As a result, paired elimination maximises all accessible data on an analysis per analysis basis. This strategy has the advantage of increasing the power of your analysis. Though this method is usually preferable over listwise deletion, it also presupposes that the missing data are MCAR.

IV. MISSING VALUE IMPUTATION MODEL

The imputation model is a collection of systems that attempt to replace missing data with evaluated ones. The goal is to use identified relations that may be found in large numbers of data tuples to aid in determining missing characteristics. This domain is concerned with the attribution of missing values.

1) Mean Value Substitution Technique: This is one of the most often utilised imputation techniques. It consists of substituting the misplaced data for a known segment or value using the average of every single supplied estimation of that attribute in the class where the quantity with lost feature has been placed.
2) HD ("Hot Deck Imputation"): Given a deficient pattern, HD replaces the lost information with a values construction read information vector that is closest to the known attributes in both designs. HD operations to ensure the distribution by substituting typical observed values for each missing variable. Cold deck imputation is a comparable methodology for HD that gets diverse information resources where present dataset.
3) KNN ("K-Nearest Neighbour Imputation"): K-Means is to layout or to group the articles needy on attributes or features into 'k' number of group. The meeting is doneby restrictive the sum of squares of distance amongst data and the relating cluster centroid. It gives speedy and exact method for evaluating lost qualities.
4) "K-Means Clustering Method": The purpose of K-Means is to arrange or group articles based on traits or features into 'k' number of groups. The meeting is accomplished by limiting the sum of squares of distance between data and the corresponding cluster centroid. It provides a quick and accurate approach for evaluating lost attributes.
5) FKMI ("Fuzzy K-Means Clustering Imputation"): In this FKMI, membership algorithm assumes a significant job. Membership function is dispensed with every data point that delineates in which degree the information point is having a position with the group. The information objects would not get dispensed to solid cluster which is demonstrated by centroid of group (as on account of K means), this is direct result of the dissimilar membership degrees of every data with entire K clusters[1].
6) "Regression Imputation": Utilizing regression strategy for the imputation, the qualities from the highlights are watched and then anticipated qualities are used for filling out lost values[1].

7) “Multiple Imputations”: Because the imputed values are derived from an appropriation, they are naturally variable. Multiple imputations (MI) illuminate the restrictions of each imputation by adding a new sort of mistake type on range in the argument that appraises more than the imputation, which is known as amongst imputation errors. It returns all lost items with at least two appropriate attributes, showing a range of possible outcomes.

Methods	Remarks
Mean Valueimputation Method	-Exchange MVs with the valuemean of ensuing data. - Standard Deviation and mean value after attribution can be a lot more than that of the actual.
HD “Hot Deck”Imputation	All missing worth is exchanged with a watched reaction from a "comparative" element. The comparative technique forHot Deck is Cold deck imputation.
KNN “K-NearestNeighbor” Imputation:	-This technique applies k-nearestneighbor model to approximateand substitute missing value. -This method can appraise all the subjective values and quantitative properties.
K-Meansclusteringmethod:	- At that point KMI utilizes calculation called closest neighbor to attribute the MVs similarly like KNNI.
FKMI “Fuzzy K-Means clusteringImputation”:	- Unreferenced characteristics foreach uncompleted values are replaced by FKMI dependent on partisanship degrees and cluster centroid attributes.
RegressionImputation:	- Supplant MVs with the qualitiesanticipated from watched attributes. Regression Equation: $S = \alpha 0 + \alpha 1\ T$
MI “MultipleImputation”:	-Multiple imputations (MI) illuminate the restraints of on itsown imputation. - It returns every missing data withtwo suitable merits.

Table 1: Overview Of Imputation Methods[1]

V. Missing value algorithm evaluation

We investigated three alternative techniques from the literature that have been found to perform well at imputation of missing values in data sets: Bayesian principle component analysis (BPCA) [8, 9], and Radial basis Function Networks [me]. We examine the efficacy of these approaches on three distinct types and sizes of data sets with varying percentages of missing data.

We selected data from two distinct data sets: time series (TS) and non-time series (NTS). In general, each of these data sets has a different sort of expression pattern. We used air pollutant data from Calen [10] and data from Lim [11, 12] as examples of TS experiments. We also used Oba et al [8] data sets as an example of a TS experiment. We used data sets from Lim and Yogan [13] for the NTS experiment. A brief summary of various data sets is provided in below Table II.

DATASET	DIMENSION	CATEGORY
LIM & YOGAN	100 x 3	NTS
CALEN	480 x 9	TS
OBA	474 x 14	TS

TABLE 2: Three Data Sets Used In Our Analysis

1) “Local Least Square”(LLS): The LLS approach of [9] picks neighbours based on Pearson correlation, like in Ordinary Least Squares [14,15], but instead of weighted univariate regressions, it does multiple regression with all k closest neighbours. The MVs are computed using the pseudo-inverse of the k nearest neighbours expression matrix and the least squares estimations. If the fraction of MVs is modest, neighbour data containing MVs is omitted from least squares systems; otherwise, MVs are calculated using the row average.
2) “Bayesian Principle Component Analysis” (BPCA): This method uses Bayesian estimation to fit a probabilistic PCA model [8]. A variational Bayes algorithm is used to iteratively estimate the posterior distribution of the model parameters and the MVs until convergence is reached. The key feature of this approach is that principle axes with small signal to noise ratios are shrunk toward zero, so that the algorithm automatically screens for those axes that are the most relevant. MVs are initially imputed by row average.
3) “Radian Basis Function”(RBF) Network”): According to Haykin [16], the RBF network is a three-layer network. The input values are passed to the connecting arcs via the input nodes. The internal units are organised into a single layer of L RBF nodes. The nodes are just weighted sums of the replies. The training method is composed of a collection of input-output pairs [z(ij, y(i)], I — 1,...,K, where x(ij is the N-dimensional input vector, y(i) is the matching goal or intended M-dimensional output vector, and K is the number of training samples. The information base is a set of input-output examples that is used to calculate the values of the unknown parameters, which are the hidden node centres and radii, as well as the connection weights between the hidden and output layers.

VI. Assessment of performance

Normalized root mean squared error (NRMSE) is used to assess the performance of missing value estimation:

$$\text{NRMSE} = \sqrt{\frac{Mean[(y_{guess} - y_{answer})^2]}{Variance[y_{answer}]}}$$

Where the mean and variance are computed across all missing items in the matrix. Because the missing entries are fictitious, we know yq, q„. When the estimation is correct, NRMSE reaches its smallest value of 0.00. NRMSE approaches 1.00 when the estimation is comparable to a random guess, which occurs when the estimation is too poor or the noise involved is too significant.

VII. Conclusion

To reduce the detrimental impact of incomplete data on mining jobs, several methodologies for missing value imputation outlined above are used. However, the majority of these techniques have some restrictions. For starters, the majority of them are assumption based that the different qualities of data records are independent of one another. Real-world data, on the other hand, may include intrinsic connections across numerous variables. This might result in improper filling of missing data. Second, the majority of missing value imputation algorithms are exclusively relevant to numerical data. However, in the actual world, data might be numerical, categorical, or mixed. Most approaches fail in such situations. However, an essential problem such as noise, which may be present in the data set, has been overlooked in current research.

We conducted a thorough examination of existing approaches for data imputing missing values. Our research shows that the imputation methods are very competitive with one another. In our simulation analysis, the BPCA imputation approaches fared the best overall. Although the LLS method is based on neighbor selection for imputation, it also has aspects that are similar to global based imputation. LLS also enables the selection of a very high number of data points (up to several thousand) for imputation [9]. Over the data sets we examined, the BPCA is more consistent than the other approaches.

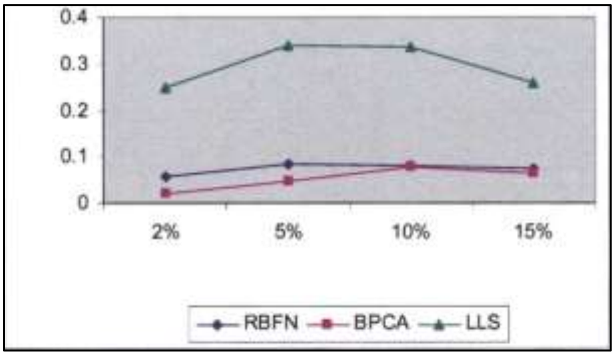

Fig. 2: NRMSE values for different percentages of missing values in Lim & Yogan data set

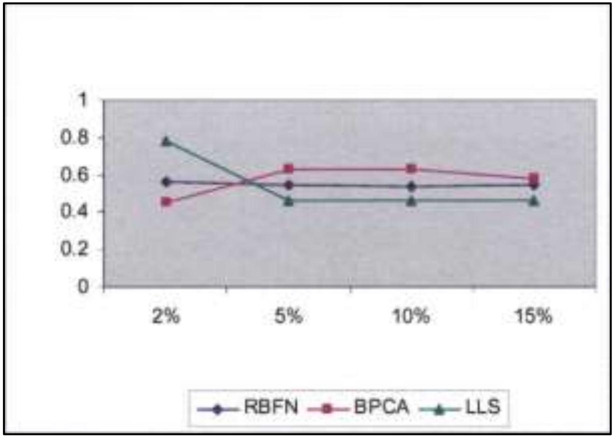

Fig. 3: NRMSE values for different percentages of missing values in Oba data set

Method	Data Set			
	Lim & Yogan	Calen	Oba	Overall
RBF Network	0.0736	0.6122	0.5479	0.4112
BPCA	0.0533	0.0197	0.5759	0.1163
LLS	0.2960	0.0189	0.5405	0.2851

Table 3: Average NRMSE values of MV for all data sets

References

[1] B S Panda and Rajesh Kumar Adhikari, “A Method for Classification of Missing Values using Data Mining Techniques,” International Conference on Computer Science, Engineering and Applications (ICCSEA), March 2020 doi:10.1109/ICCSEA49143.2020.

[2] Graham, J. W. “Missing data analysis: Making it work in the reality. Annual review of psychology”, 60:549-576, 2014, 2009.

[3] Archana Purwar and Sandeep Kumar Singh, “Issues in Data mining: A comprehensive survey,” IEEE International Conference on Computer Science, Engineering and Applications (ICCSEA), Dec 2014.

[4] Elenita T. Capariño, Ariel M. Sison and Ruji P. Medina, “Application of the Modified Imputation Method to Missing Data to Increase Classification Performance,” IEEE 4th International Conference on Computer and Communication Systems (ICCCS) Feb 2020 doi: 10.1109/ICCCS46626.2019.

[5] Dr. A.Sumathi “Missing Value Imputation Techniques Depth Survey to Improve the Efficiency of Imputation”. IEEE-Fourth International Conference on Advanced Computing, ICoAC, 2012.

[6] S. Jain, M. K. Jain, and D. N. Chodhary, “A SURVEY PAPER ON MISSING DATA IN DATA MINING,” International Journal of Innovations in Engineering Research and Technology (IJIERT), vol. 7301, no. 12, pp. 11–13, 2016.

[7] Abu-Soud S and Sufyan Al Majali. “ILA-3: An Enhanced Version of ILA with a New Feature Selection Approach “, WSEAS Transactions on Systems and Control, vol. 13, pp. 171-185, ISSN: 19918763, 2018.

[8] S. Oba et al, “A Bayesian missing value estimation method for gene expression profile data”, Bioinformatics 2003, 19(16): 2088-2096.

[9] T.H. Bo, B. Dysvik, I. Jonassen, “LSimpute: accurate estimation of missing values in microarray data with least squares methods”, Nucleic Acids Res 2004, 32(3): e34.

[10] W. Calen, Radial Basis Function Neural Networks, Master Dissertation, USM, Pulau Pinang, 2005.

[11] E.A. Lim, Z. Zainuddin, “A Symmetry-Based Fuzzy C- Means Clustering for training Radial Basis Function Networks”, Proceeding of the Third International Conference on Research and Education in Mathematics 2007 (ICREM III), pp. 319-323.

[12] Z. Zainuddin, E.A. Lim, W. Calen, “Subtractive clustering for training Radial Basis Function Networks”, Proceeding of International Conference on Quantitative Sciences and Its Application 2005 (ICOQSIA 2005), Fuzzy System.

[13] E.A.Lim, J.K. Yogan, “A Study of Neuro-Fuzzy system in Approximation Problems”, Journal of MATEMATIKA, Volume 24(1).

[14] D.V. Nguyen, N. Wang, R.J. Carroll, “Evaluation of missing value estimation for microarray data”, Journal of Data Science 2004, 2:347-370.

[15] T.H. Bo, B. Dysvik, I. Jonassen, “LSimpute: accurate estimation of missing values in microarray data with least squares methods”, Nucleic Acids Res 2004, 32(3): e34.

[16] S. Haykin, Neural Networks: a comprehensive foundation, Prentice Hall, New York, 1994.