NOVATEUR PUBLICATIONS
INTERNATIONAL JOURNAL OF INNOVATIONS IN ENGINEERING RESEARCH AND TECHNOLOGY [IJIERT]
ISSN: 2394-3696
VOLUME 3, ISSUE 12, Dec.-2016

A SURVEY PAPER ON MISSING DATA IN DATA MINING

SWATI JAIN
Department of computer science and engineering, MPUAT University/College of technology and engineering, Udaipur, India, swati.subhi.9@gmail.com

MRS. KALPANA JAIN
*Assistant professor of CSE, College of technology and engineering, Udaipur, India,*kalpana_jain2@rediffmail.com

DR. NAVEEN CHODHARY
Head of Department in CSE, College of technology and engineering, Udaipur, India, naveenc121@yahoo.com

ABSTRACT
Missing data are characterized as a portion of the qualities in the data set are either lost or not watched or not accessible because of natural or non natural reasons. Data with missing qualities confuses both the data examination and the accommodation of an answer for new data. Numerous specialists are working on this issue to present more modern techniques. Despite the fact that numerous strategies are available, investigators are confronting trouble in seeking an appropriate technique because of absence of information about the strategies and their applicability. This research paper additionally directs a formal review of the missing data strategy. It talks about the strategies that are analyzed in the written works and perceptions that the authors have made.

INTRODUCTION
A large portion of this present datasets experience suffers from the issue of missing data. It might lead information mining examiners to end with wrong inferences about data under review. Data mining is the process which provides a concept to attract attention of users due to high availability of huge amount of data and need to convert such data into useful information. Data preparation is a principal phase of data investigation[5][7]. Three types of mean imputation methods presented on missing data [1]. Rubin investigated about inference and missing data and multiple imputation for non-reaction in the overview[3]. Allison explored estimates of linear models with incomplete data and on missing data[4]. Myrtveit et al. connected missing data strategies to a software project data set, and assessed four missing data procedure are list wise deletion (LD), mean imputation (MI), similar response pattern imputation (SRPI) and full information maximum likelihood (FIML)[11].Junninen et al. evaluated and compared univariate and multivariate methods for missing data imputation in air quality data sets[18].

Types of Incomplete Data are little and Rubin characterize a list of missing mechanisms, which are generally acknowledged by the community. There are three mechanisms under which missing data can happen [3]:

1) Missing completely at random (MCAR): MCAR is the probability that a observation (Xi) is missing, is unrelated to the estimation of Xi or to the estimation of some other variable and the explanation behind missing is completely random. This circumstance is uncommon in real world and is generally talked about in statistical theory.

2) Missing at random (MAR): MAR is the probability of the observed missingness design, given the observed and unobserved data, does not depend on the values of the unobserved

data. This component is normal in practice and is normally considered as the default kind of missing data.

3) Not missing at random (NMAR). If the probability that an observation is missing depends on information that is not observed, this kind of missing data is called not missing at random. This circumstance is generally confused and there is no universal solution.

REVIEW OF TECHNIQUES FOR MISSING DATA:

Four Common missing value methods for dealing with missing values [3]:

1. Removing the tuples: If a missing value transpires on any of the factors in the data, eliminate the entire observation.
2. Filling the missing value manually: This strategy is time consuming and not plausible for a vast information set with many missing values.
3. Use global constant: Replace all missing attribute values by a similar consistent, for example, a lable "unknown "or - ∞.
4. Use the attribute mean value: Filling in the missing estimations of a variable. Substitution with a measure of central tendency, Mean, Median, Midrange, (Max + Min)/2 and Mode.

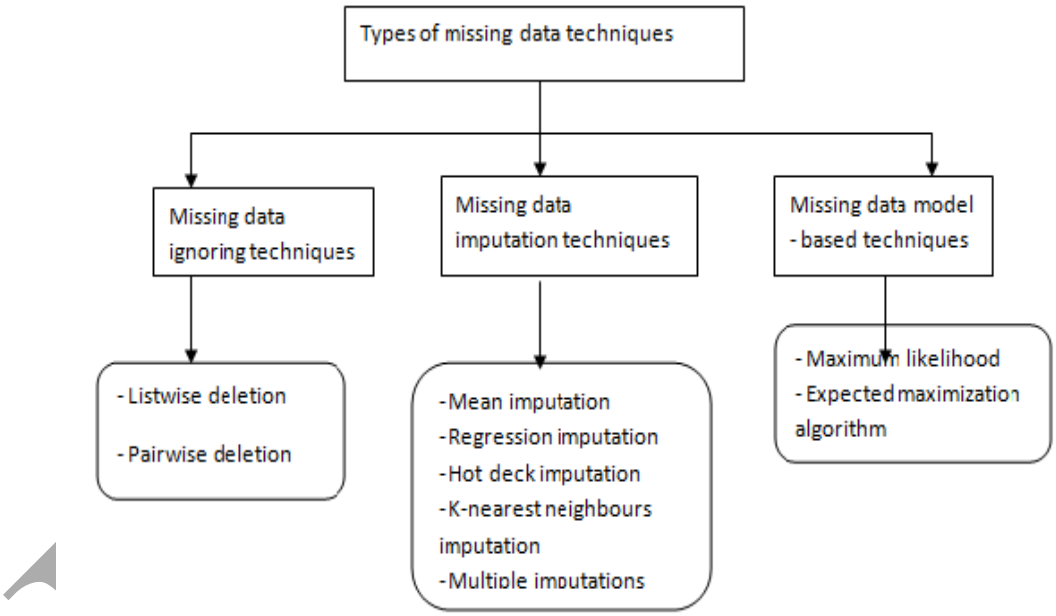

Figure 1: Types of missing data technique

There are several strategies for treating missing data, a few techniques are described below. Missing data treatment techniques can be isolated into three classes, as proposed in [6]

1.1 MISSING DATA IGNORING TECHNIQUES

Listwise deletion (or complete case analysis): If a case has missing data for any of the variables, then essentially avoid that case from the analysis. It is typically the default in statistical packages[19].

PAIRWISE DELETION (PD): is referred to as the available case method. This technique considers every feature independently. For each feature, all recorded values in each observation are considered [12] and missing data are overlooked.

Table 1: overview of ignoring and discarding method

Methods	Remark
Listwise deletion (or complete case analysis)	-Deletion of all cases containing missing values. -High Loss of information
Pairwise deletion (PD):	-Deletion of records only from column containing missing values. -Less loss of information by keeping all available values.

1.2 MISSING DATA IMPUTATION TECHNIQUES

Imputation method is a class of techniques that expects to fill in the missing values with estimated ones. The goal is to employ known connections that can be distinguished in the valid values of the data set assist in estimating the missing values. This field focuses on imputation of missing data.

MEAN VALUE IMPUTATION METHOD: Mean imputation method is one of the most frequently used methods [3]. It comprises of replacing 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 belongs.

HOT DECK IMPUTATION(HD): Given an incomplete pattern, HD replaces the missing data with values form input data vector that is nearest in terms of the attributes that are Known in both patterns. HD attempts to protect the distribution by substituting different observed values for each missing .The similar method of HD is Cold deck imputation method which takes other data source than current dataset[13]

K-NEAREST NEIGHBOR IMPUTATION (KNN): This strategy uses k- nearest neighbor algorithms to estimate and replace missing data. The main advantages of this strategy are a) it can estimate both qualitative attributes and quantitative attributes; b) It is not important to construct a predictive model for each attribute with missing data [15].

K-MEANS CLUSTERING METHOD: K-Means is to classify or to group the objects based on attributes/features into k number of group [15] [16]. The grouping is finished by minimizing the sum of squares of distances between data and the corresponding cluster centroid.It provides quick and precise way of estimating missing values.

FUZZY K-MEANS CLUSTERING IMPUTATION (FKMI): In FKMI, membership function plays an important role. Membership function is allocated with every data object that depicts in what degree the data object is belonging to the particular cluster. Data objects would not get allotted to concrete cluster which is indicated by centroid of 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: Using regression method for imputation, the values from the features are observed and then predicted values are used for filling Missing values [14].

MULTIPLE IMPUTATIONS: The imputed values are draws from a distribution, so they inherently contain some variation. Thus, multiple imputations (MI) illuminates the limitations of single imputation by presenting an additional form of error based on variation in the parameter estimates across the imputation, which is called between imputation error. It

replaces each missing item with two or more acceptable values, representing a distribution of possibilities [4].

Table 2: overview of imputation method

Methods	Remarks
Mean Value imputation Method	-Replace MVs with the mean of Resultant data. -Mean and SD after imputation may be much higher than that of original.
Hotdeck (HD) Imputation	-each missing value is replaced with an observed response from a "similar" unit. - The similar method of HD is Cold deck imputation.
K-Nearest Neighbor Imputation (Knn):	-This method uses k-nearest neighbor algorithms to estimate and replace missing data. -It can estimate both qualitative attributes and quantitative attributes.
K-Means clustering method:	-Then KMI uses algorithm called nearest neighbour to impute the MVs in the same way as KNNI .
Fuzzy K-Means clustering Imputation (FKMI):	-Unreferenced attributes for every uncompleted data are substituted by FKMI on the basis of membership degrees and cluster centroid values.
Regression Imputation:	-Replace MVs with the values predicted from observed values Regression Equation: $Y = \alpha 0 + \alpha 1\ X$
Multiple Imputation:	-Multiple imputation (MI) illuminates the limitations of single imputation. -It replaces each missing item with two or more acceptable values.

1.3 MISSING DATA MODEL-BASED TECHNIQUES

Maximum likelihood techniques are utilized to guess the parameters of a model defined for the all data. Maximum likelihood procedures that use variants of the Expectation-Maximization algorithm can deal with parameter estimation within the sight of missing data.

MAXIMUM LIKELIHOOD: We can utilize this technique to get the variance-covariance matrix for the variables in the model based on all the available data points, and after that obtained variance- covariance matrix to estimate our regression model [10].

EXPECTATION-MAXIMIZATION (EM) ALGORITHM: It depends on an expectation step and a maximization step, which are steady several times until maximum likelihood estimates are obtained. It requires a large sample length and that the data are missing at random (MAR) [4] and [2].

Table 3: overview of model based method

Methods	Remarks
Maximum Likelihood	-It is a method of estimating the parameters of a statistical model given observations. - finding the parameter values that maximize the likelihood of making the observations given the parameters.
Expectation-maximization(EM)algorithm	-Iterative method, finds maximum likelihood -Two steps: Expectation (E step), Maximization (M step) Iteration goes on until algorithm converges.

CONCLUSION

This survey mainly focuses on the study of missing data handling method in data mining consequently, imputation methods are broadly used to fill the missing values of various kinds of datasets. In this survey, the overall views on the handling missing methods and their classes are discussed. In this manner it can be clearly seen that numerous strategies are

proposed for handling missing values present in the dataset. Further, these imputation methods are compared along with their advantages and disadvantages.

REFERENCES

1) Noor, M. N., Yahaya, A. S., Ramli, N. A., & Al Bakri, A. M. M. 2014. *Mean imputation techniques for filling the missing observations in air pollution dataset Key Engineering Materials 594-599:902-908 Trans Tech Publications*

2) Graham, J. W. 2009. *Missing data analysis: Making it work in the real world. Annual review of psychology, 60:549-576.*

3) Rubin, D. B. 1976. *Inference and missing data. Biometrika, 63(3):581-592.*

4) Allison, P. D. 1987. *Estimation of linear models with incomplete data. Sociological methodology, 71-103.*

5) Smyth, P. 2001. *Data mining at the interface of computer science and statistics. In Data mining for scientific and engineering applications 35-61. Springer US.*

6) *A Case Study of Heart Failure Datasetl"*2012, 9th International Conference *on Fuzzy Systems and Knowledge Discovery (FSKD 2012).*

7) Zhang, S., Zhang, C., & Yang, Q. 2003. *Data preparation for data mining.Applied Artificial Intelligence, 17(5-6):375-381.*

8) Song, Q., & Shepperd, M. 2007. *A new imputation method for small software project data sets. Journal of Systems and Software, 80(1): 51-62.*

9) Grzymala-Busse, J. W. 2004. *Data with missing attribute values: Generalization of indiscernibility relation and rule induction. Transactions on Rough Sets 1:78-95. Springer Berlin Heidelberg.*

10) Schafer, J.L., 1997. *Analysis of incomplete multivariate data. Monographs on Statistics and Applied Probability No. 72.*

11) Myrtveit, I., Stensrud, E., & Olsson, U. H. (2001). *Analyzing data sets with missing fata: An empirical evaluation of imputation methods and likelihood-based methods. IEEE Transactions on Software Engineering, 27:999–1013.*

12)]Strike, K., El Emam, K., Madhavji, N., 2001. *Software cost estimation with incomplete data. IEEE Transactions on Software Engineering ,27 (10):890–908.*

13) Dr. A.Sumathi 2012.*Missing Value Imputation Techniques Depth Survey And an Imputation Algorithm To Improve The Efficiency Of Imputation. IEEE- Fourth International Conference on Advanced Computing, ICoAC.*

14) Y. Kou, C.-T. Lu, and D. Chen 2006. *Spatial weighted outlier detection. In Proceedings of the Sixth SIAM International Conference on Data Mining,614–618,Bethesda, Maryland, USA.*

15) N. Poolsawad L. Moore C. Kambhampati and J. G. F. Cleland 2012.*Handling Missing Values in Data Mining - A Case Study of Heart Failure Datasetl.9th International Conference on Fuzzy Systems and Knowledge Discovery.*

16) A.Rogier T.Donders, Geert J.M.G Vander Heljden, Theo St ijnen, Kernel G.M Moons 2006, *Review: A gentle introduction to imputation of missing values, Journal of Clinical Epidemiology,59:1087-1091.*

17) Kin Wagstaff, *Clustering with Missing Values: No Imputation Required,NSF grant IIS-0325329:1-10.*

18) Heikki Junninen, Harri Niska, Kari Tuppurainen, Juhani Ruuskanen, Mikko Kolehmainen 2004, *Methods for imputation of missing values in air quality data sets, Atmospheric Environment 38:2895–2907*

19) Briggs, A., Clark, T., Wolstenholme, J., Clarke, P., 2003. *Missing.... presumed at random: cost-analysis of incomplete data. Health Economics,12:377–392.*

20) Han, J., Pei, J., & Kamber, M. 2011. *Data mining: concepts and techniques. Morgan Kaufmann Publishers, 225 Wyman Street, Waltham, USA pp. 83-91*