Shivani Rawal et al. International Journal of Recent Research Aspects ISSN: 2349-7688, Vol. 4, Issue 3, Sept 2017, pp. 34-38

Predicting Missing Values in a Dataset: Challenges and Approaches

Shivani Rawal[1], Dr. S.C Gupta[2], Mr. Shekhar Singh[3]

[1]M.tech Student Department of CSE PIET, Samalkha
[2]Professor Department of CSE PIET, Samalkha
[3]Assistant Professor Department of CSE PIET, Samalkha

Abstract— **Data serves as an asset in the present era of information in almost all fields of study. The quality of the knowledge extracted, learning and decision problems depends upon the quality of data gathered. But most of the real world datasets tends to be incomplete. The problem of missing data can have a significant effect on the conclusions that can be drawn from this data. Data mining has made a great progress in recent year but the problem of missing data or value has still remained a great challenge for data mining. Missing data or value in a dataset can affect the performance of classifier which leads to difficulty of extracting useful information from datasets. In this paper we present the major challenges in predicting missing data in a data set and the existing approaches dealing with the missing data imputation**.

Keywords- Dataset, Data mining, missing value, data imputation.

I. INTRODUCTION

Missing data is an inherent problem in data collection especially when dealing with large real world data sets. Values in data sets can be missed because of many possible reasons like the data was not available or not applicable. The value can also be missed while entering into the dataset. Missing values in the data set create uncertainty for the analyst and the information consumer because decisions need to be made without having the full picture. Hence knowing how to predict those missing values is important. Data mining refers to the process of extracting knowledge from large volumes of data. This data may be multimedia data, spatial data, text data, time series data or the web data. Data mining helps in obtaining interesting, implicit, nontrivial, previously unknown and useful patterns from unorganized data [1].

Missing values can be categorized in three types:

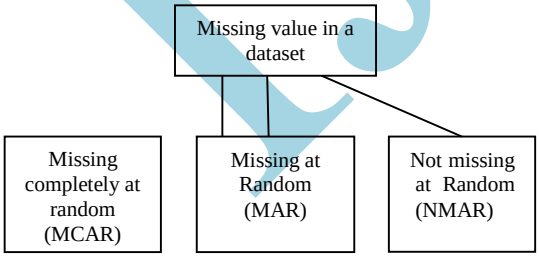

Fig. 1 Types of missing values in a dataset

A. Missing Completely at Random (MCAR)

Values in a data set are said to be missing completely at random (MCAR) if the events that lead to any particular data item being missing are independent of both the observable variables and unobservable parameters of interest, and hence occurs entirely at random. Unbiased analysis is performed on the MCAR data. The data sets rarely have MCAR data. It signifies the maximum level of randomness [9].

B. Missing at Random (MAR)

When the missing values of some attribute are not randomly distributed across the observations but are distributed within one or more samples, they are said to be missing at random (MAR). Or in other MAR does not depend on that particular attribute but depends on the values of other attribute. MAR is more common than the previous MCAR type.

C. Not Missing at Random (NMAR)

NMAR is also known as non ignorable missing value. In this case the missing data is dependent on the values of the attribute. NMAR signifies the least level of randomness [13]. It is the most problematic form as it involves missing values that are not randomly distributed across the observations. The only way to deal with NMAR data is to attain an estimate of the parameters by modelling the miss.

II. CHALLENGES IN PREDICTION OF MISSING VALUES

Missing data directly impacts the data quality [14]. If the missing data in a dataset is less than 1% of its total data then it does not cause a significant problem for knowledge discovery in database. Also 1%-5% missing data is manageable to some extent while 5-15% needs sophisticated methods for handling. But if the missing

data exceeds the 15% of total data, it severely affects the interpretation and the mining tasks in a negative way [6]. Handling of missing data through imputation methods have their own issues like loss of information or reduced efficiency, difficulty in data handling because of irregular data and systematic difference in the data. Thus such issues make the prediction task a challenging one.

A. Challenges in predicting missing values

The following challenges arise when we apply any technique to predict the missing values:

i. The missing data prediction method should not alter the distribution of data.
ii. The relationships between the attributes of the data set must be retained by the prediction method deployed.
iii. The prediction method should not be too complex and should not have high time cost factor.
iv. The missing values should be predicted and replaced in such a way that all the data mining analytical procedures can be applied to the newly completed dataset easily.

III. LITERATURE REVIEW

Missing data might occur because the value is not relevant to a particular case, could not be recorded when the data was collected, or is ignored by users because of privacy concerns. The problem of missing data has been investigated in last two decades and a number of strategies have been devised for dealing with it.

Missing values lead to the difficulty of extracting useful information from that data set. Missing data are the absence of data items that hide some information that may be important [1]. Most of the real world databases are characterized by an unavoidable problem of incompleteness, in terms of missing or erroneous values.

The simplest solution is to discard the data instances with the missing value but it leads to loss of information [2]. Another simple approach is to assign all possible values and then use k-means of the attribute [3]. To address the missing values data handling problem, a Column-wise Guided Data Imputation method (CGDI) is proposed by I. Jordanov et al. [4]. They have put forth a new scheme based on learning from the known values in the data. CGDI selects the most efficient model from a range of available imputation methods for each individual feature based on learning. Their experiments conducted on various datasets have shown the superiority of CGDI over four existing imputation techniques in term of estimation accuracy.

Dao Lam et al. have applied unsupervised feature learning (UFL) to the mixed-type data to achieve a sparse representation, which makes it easier for clustering algorithms to separate the data [5]. They have demonstrated their research on noisy data from the petroleum industry which is also mixed type and have obtained better clustering results than the existing schemes.

J. kraiser [8] have proposed an algorithm for imputation of missing value of categorical type of data. The algorithm is based on association rules and provides three different variants. J. Luengo et al. [10] have analyzed the existing data imputation methods using a group based approach to classify them in three classification method categories namely rule induction learning techniques, approximate techniques and lazy learning techniques. Through their analysis they have suggested that any particular imputation method should be deployed based on the conditioning to the group they belong. S. Thirukumaran et al. [11] have also analyzed the existing techniques for data imputation on various parameters. Based on their research they have proposed a Clustering Imputation algorithm specifically for the medical databases to minimize the error rate at which missing value is predicted.

D.V Patil et al. [12] have combined genetic algorithms techniques and decision tree learning to propose a new technique for predicting the missing data values. "Survival of the Fittest" is the basic principle of the genetic algorithm used which uses the domain values as the pool of solutions for the attribute with missing values. The solution uses the global search technique with its fitness function as the classification accuracy to achieve an optimal solution.

IV. COMPARISON OF EXISTING APPROACHES FOR PREDICTION

The methods for the treatment of missing data can be classified in certain categories based on characteristics like the phase KDD process in which missing data is handled, their basic approach, the kind of attributes available or the time of treatment of missing data [16].

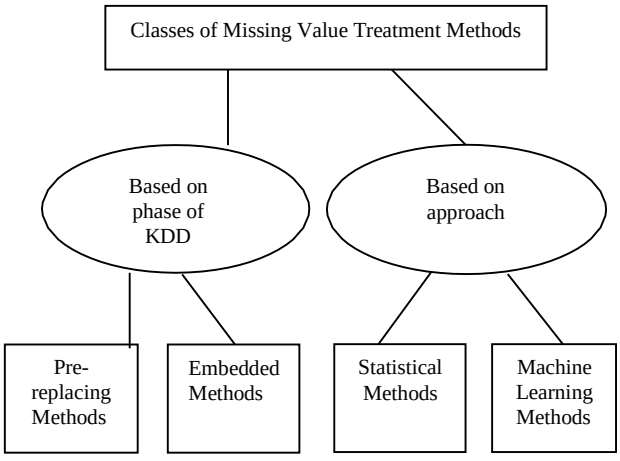

Fig. 2a Classes of Missing Value Treatment Methods

Figure 2a describes that on the basis of phase of KDD process in which the missing data is handled; there are two types of missing data treatment methods: Pre replacing methods and embedded methods. Pre replacing methods works in the data preparation phase to deal with the missing data while embedded methods work in data mining phase of the KDD process. Pre replacing methods are more flexible while embedded methods are cost effective.

On the other hand, on the basis of basic approach the missing data treatment methods are categorized as statistical methods and machine learning methods. Statistical methods are simple to apply but machine learning methods provide greater accuracy.

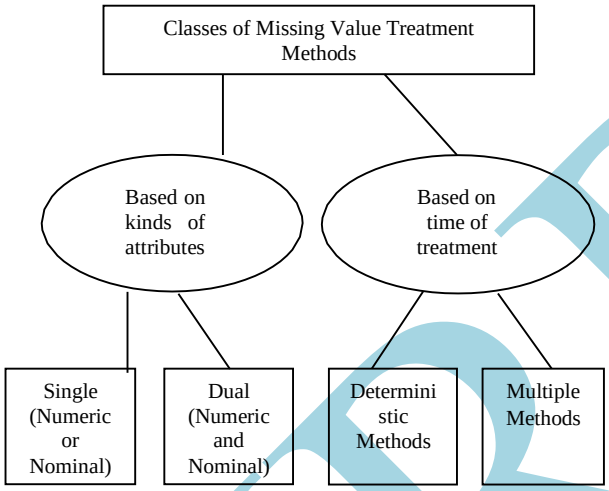

Fig. 2b Classes of Missing Value Treatment Methods

Above Figure 2b shows that on the basis of types of attributes values, the treatment methods can be either dealing with single type attribute (numeric or nominal) or dual type attributes (both numeric and nominal). Finally on the basis of time of treatment of the missing data the treatment methods are either deterministic methods or multiple methods. Deterministic methods do not tackle the uncertainty of replaced values as they are confined to single value imputation, where as multiple methods impute multiple values.

Many methods for above classes have been proposed by various authors. We describe in detail and compare the popular existing techniques or approaches for missing data imputation in a data set.

A. Instance Deletion

The method of instance deletion is the most primitive approach used for handling the missing values in a dataset. It involves the complete deletion of the instances with missing data and analysing the remaining complete data of the dataset [2]. Though it is easy to implement but it has several consequences. First, it leads to reduction in the size of data set available for analysis which in turn gives inaccurate results for mining. Secondly, deleting entire instance causes biasing in the distribution of data and its statistical analysis because the data is not always missing at random. An improved version of this method can be deleting the attributes with high missing rate but only after running the relevance analysis.

B. Mean/Mode Imputation

In this method of missing data imputation, the numeric type missing data is replaced by the means of rest of the values for that attribute while the nominal type of missing data is replaced by the mode of the rest of values for that attribute [3].

C. Hot Deck Imputation

This method involves a two stage processing of data set with missing values [11]. First the data set is partitioned into clusters. Then within every cluster the missing value are replaced with predicted values [14].

D. Prediction using K- nearest neighbour algorithm (KNN)

The KNN prediction algorithm searches for the most similar instances of the data for predicting and replacing the missing values. It can be used for both numeric and nominal values. The value selected for 'k' has a huge impact on results.

E. Regression Imputation

In Regression based imputation, it is assumed that a variable changes its value linearly with other variables or there is a linear relationship between the attributes of the data set. So the missing data can be replaced by the linear regression function. But the drawback of this method is that generally the relationship among the attributes is not linear. Support Vector Machine Imputation (SVMI) is an example of regression based imputation technique that takes condition and decision attributes [7]. It is then applied for the prediction of values for the missed condition.

F. Prediction using Bayesian Iteration

The Naive Bayesian Classifier is a simple and popular classifier which gives good performance in terms of accuracy. Prediction using Bayesian Iteration also consists of two phases. Initially the order of attribute with missing values is decided based on parameters such as information gain, weighted index etc. Thereafter the Naive Bayesian Classifier is used to predict the missing data in an iterative process. In first iteration, the algorithm replaces the missing value in the first attribute in the order and then goes to next attributes in further iterations.

G. Fuzzy k- means clustering imputation

In Fuzzy k- means clustering imputation method, every data object is assigned a membership function [8]. This membership function signifies the degree of belongingness of that data object to any particular cluster. Then on the basis of these membership function and cluster centroid values, this method substitutes the missing values in the dataset [13].

H. Prediction using C4.5 algorithm internal methods

C4.5 is a tree based classifier used widely [15]. It has been further improved for handling missing data by developing some internal algorithms. C4.5 uses the probabilistic approach to treat missing data. It first selects the attribute from the dataset based on correctional gain ratio and then distributes all the missing data instances into subsets based on probability of size of subset. The decision tree classifies the instances and searches all possible paths. Finally a classification result is obtained in terms of probability.

The following table summarizes the advantages and disadvantages of the above discussed techniques for missing data imputation.

TABLE I

COMPARISON OF APPROACHES FOR MISSING DATA IMPUTATION

Prediction Technique	Advantages	Disadvantages
Instance Deletion	Convenient to apply	a) Reduction of size of dataset b) Induction of bias c) Reduction of accuracy for data mining
Mean/Mode Imputation	Ease of Application	a) Replacing all missing values with same mean changes the characteristic of original data set b) Induces bias
Hot deck imputation	a) Predicts realistic values b) Avoids Distortion in Imputation	Difficult to predict in non related samples available for prediction
KNN prediction	a) Can predict both quantitative and qualitative attributes b) Doesn't require a predictive model for each attribute with missing data	With the increase in size of data set, the selection of value of K becomes a critical issue.
Regression imputation	Works well with large size datasets	Performance degrades if the samples taken are less than the features in the dataset
Prediction using Bayesian Iteration	Provides good performance by using Naïve Bayesian classifier	Can deal with only the nominal attributes for prediction
Fuzzy k means clustering imputation	Better performance output than simple K means prediction as data objects can belong to more than one cluster	a) High implementation cost in terms of computation time b) Highly sensitive to noise
C 4.5 internal prediction	a) It can be applied to both nominal and numeric attributes. b) Searches all possible paths to give result in form of classification	Computation time increases significantly with increase in size of dataset

The above table shows that though all of the existing approaches predict the missing values in a dataset but still they have some significant drawbacks and scope of improvement. They can be further improvised to predict values with higher accuracy so that the data mining tasks applied on them can yield better and accurate conclusions.

V. CONCLUSION AND FUTURE WORK

Missing values present in a dataset can create huge classification and data mining errors. So missing values must be predicted and replaced before analyzing that dataset. In this work we discussed the types of missing data and the major challenges faced while missing data imputation. Further we listed the categories of imputation techniques based on parameters like phase of the KDD process, approach, time etc. Finally we described eight prominent existing techniques for predicting the missing values and evaluated their advantages and disadvantages. In future, we would work to devise a novel technique for the missing data imputation that deals with the flaws of existing techniques and provides higher accuracy results.

REFERENCES

[1] D.J. Prajapati, J.H. Prajapati, "Handling Missing Values: Application to University Data Set", Issue 1, Volume 1, ISSN 2249-6149, 2011.

[2] G. Ssali, T. Marwala, "Estimation of missing data using computational intelligence and decision trees", Proceeding of IEEE International Joint Conference on Neural Networks, Hong Kong.

[3] S. Sugana, K.G. Thanushkodi, "Predicting missing values using K-means Clustering", Journal of Computer Science, Issue 2, Volume 7, Pages 216-224, 2011.

[4] A. Petrozillo and I. Jordanov, "Coulmn wise guided data imputation", Elsevier B.V, 2017.
[5] D. Lam, M. Wei, D. Wunsh, "Clustering data of mixed categorical and numerical type with unsupervised feature learning", IEEE, 2015.
[6] E. Acuna and C. Rodriguez, "The treatment of missing values and its effect in the classifier accuracy", Springer-Verlag Berlin-Heidelberg, Pages 639-648, 2004.
[7] S.K Singh, A. Purwar "Empirical Evaluation of Algorithms to impute Missing Values for Financial Dataset", IEEE International Conference, 2014.
[8] J. Kaiser, "Algorithm For Missing Values Imputation In Categorical Data With Use Of Association Rules", The Ninth International Conference on Web-Age Information Management IEEE, 2012.
[9] C.K Enders, "Applied missing data analysis", Guidford: Guildford Press, 2010
[10] J. Luengo, S. Garca, F. Herrera, "On the choice of the best imputation methods for missing values considering three groups of classification methods", Knowledge and Information Systems 32, Pages 77-108, 2012.
[11] A.Sumathi, S. Thirukumaran, "Missing Value Imputation Techniques Depth Survey and an Imputation Algorithm to Improve the Efficiency of Imputation". IEEE- Fourth International Conference on Advanced Computing (ICOAC), 2012.
[12] D.V. Patil, R.S Bichkar, "Multiple Imputation of Missing Data with Genetic Algorithm based Techniques" IJCA Special Issue on "Evolutionary Computation for Optimization Techniques" (ECOT), 2010.
[13] T.V Rajnikanth, "An Enhanced Approach On Handling Missing Values Using Bagging K-NN Imputation", International Conference on Computer Communication and Informatics (ICCCI), 2013.
[14] N. Poolsawad, L. Moore, C. Kambhampati, J. G. F. Cleland "Handling Missing Values in Data Mining - A Case Study of Heart Failure Dataset", 9th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), 2012.
[15] J.W. Grzymala, M. Hu "A Comparison of Several Approaches to Missing Attribute Values in Data Mining", (Eds.): RSCTC 2000, LNAI, Pages 378−385, 2005.
[16] Yoshikazu Fujikawa, "Efficient Algorithms for Dealing with Missing values in Knowledge Discovery", Master Degree Thesis, Japan Advanced Institute of Science and Technology, 2001

