Volume 2, Issue 11, November 2017 ISSN (online): 2456-0006
International Journal of Science Technology Management and Research
Available online at: www.ijstmr.com

A New and Efficient Approach to Constructs Missing Values for Categorical Attribute

Banesingh Pachlaniya
M. Tech. (CSE) IV Sem
Lord Krishna College of Technology
Indore M.P. India

Vijay Kumar Verma
Asst. Prof. C.S.E. Dept.
Lord Krishna College of Technology
Indore M.P. India

Abstract: ***Missing values and incomplete data are a natural phenomenon in real datasets. The problem of recovering missing values from a dataset has become an important research issue in the field of data mining and machine. With the rapid increase in the use of databases, the problem of missing values inevitably arises.The techniques developed to effectively recover these missing values. There are several methods have been proposed to construct missing value. . In This paper we present a new and more efficient approach to construct missing value for categorical dataset. In this paper we proposed frequent pattern and association rule based approach to construct missing values.***

Keywords: Missing value, frequent pattern Association, Incomplete

I. INTRODUCTION

Discovery in Databases (KDD) attempt to mine hidden and useful information from large databases. Although current computer technologies can handle massive amounts of data, the rapid growth of databases causes some attribute values to be missed or causes inconsistencies in the data gathering process. Before data analysis begins, the data cleaning step deals with errors and inconsistencies from raw data to improve the quality of the following discovered information. Therefore, the problem of recovering missing values has become a top priority and has played an important role in the data mining field one simple method of dealing with missing values is to delete all tuples with missing values. The resulting truncated databases often include too little data to analyse effectively. As an alternative, users can apply some simple statistical methods such as using mean or median to predict missing values. However, the predicted values, which are still inaccurate, become noise and influence the quality of the information. Consequently, to effectively deal with missing values, several researches have been proposed a verity of methods.
Consider a simple example of incomplete dataset in table there are twelve tuples and three attributes, A, B, C.

	Items
TID	A, B, C, D, E
T1	A,B,C,D
T2	?, D,E
T3	A, B, C, D, E
T4	A, B, C, D, E
T5	A,?,D
T6	A, B, C, D, E
T7	?,C, D, E
T8	A, B, C, D, E
T9	A,C,D
T10	A,B,D,E

Table 1 Transactional data set with missing item

Common method for treatment of missing value includes:
(1) Replacing missing attribute values by the most common (most frequent) value of the attribute,
(2) For numerical attributes, missing attribute value may be replaced by the attribute average value,
(3) Assigning all possible values of the attribute. (4) Ignoring cases with missing attribute values. (5) Considering missing attribute values as special values.

II. CLASSIFICATION OF MISSING VALUES

Missingness can be classified into three categories

(1)MCAR (Missing Completely At Random) - Missing completely at random A value is missing completely at random if it is independent of its own value, or the value of any other attribute. The probability that a value of some attribute is missing is always the same. This is random missingness in the intuitive sense of the word.
(2)MAR (Missing At Random) - Missing at random A value is (quite confusingly) called missing at random, if it is dependent of one or more other attributes (i.e. their values in the same tuple.
(3) MNAR (Missing Not At Random) - Missing not at random A value is called missing not at random, if it is dependent of itself. A typical example is a measuring device incapable of detecting extreme values above a certain threshold, and so produces null values. This is the most difficult mechanism to deal with, since it cannot be derived from the data.

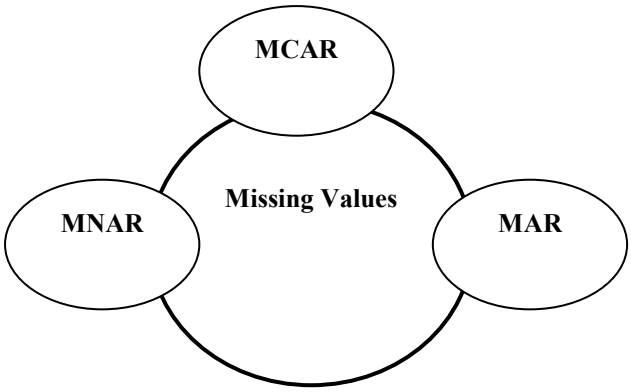


Figure1. Different types of missing values

III. LITERATURE REVIEW

Some of the paper which we have studied related with our topics is
In 2007 Toon Calders et. Proposed "Mining Itemsets in the Presence of Missing Values "Missing values make up an important and unavoidable problem in data management and analysis. They proposed an efficient algorithm, XMiner, for mining association rules and frequent itemsets in databases with missing values. They evaluated XMiner and empirically shows gain over a straightforward baseline-algorithm [1].

In 2008 David C. Howell et al proposed " The Treatment of Missing Data" They introduced treatment of missing data across a range of experimental designs, starting with those designs whose treatment is relatively straightforward (though not necessarily satisfactory) and moving to situations where the optimal solution is elusive. They showed that recent techniques have come far in narrowing the gap between the ideal and the practical [2].

In 2009 John W. Graham et al proposed "Missing Data Analysis: Making it works in the real world". They reviewed and presented a practical summary of the missing data literature; including a sketch of missing data theory and descriptions of normal model multiple imputation (MI) and maximum likelihood methods. They discussed, most notably the inclusion of auxiliary variables for improving power and reducing bias [3].
In 2010 Amanda N. Baraldi et al proposed "An introduction to modern missing data analyses". They proposed a study over recent methodological and focused on two modern missing data analysis methods: maximum likelihood and multiple imputations. These approaches are advantageous to traditional techniques (e.g. deletion and mean imputation techniques) because they require less stringent assumptions and mitigate the pitfalls of traditional techniques. They explain the theoretical underpinnings of missing data analyses, give an overview of traditional missing data techniques, and provide accessible descriptions of maximum likelihood and multiple imputations [4].

In 2011 Tzung-Pei Hong et al. proposed "Mining rules from an incomplete dataset with a high missing rate". They introduced an iterative missing value completion method based on the RAR (Robust Association Rules) support values to extract useful association rules for inferring missing values in an iterative way. It consists of three phases. The first phase uses the association rules to roughly complete the missing values. The second phase iteratively reduces the minimum support to gather more association rules to complete the rest of missing values. The third phase uses the association rules from the completed dataset to correct the missing values [5].

In 2012 S. S. Dhenakaran et. al proposed "A Perspective Missing Values In Data mining Applications". They calculated missing set values and estimate the imputation of missing values in data set. Methods are discussed for learning and application of decision rules for classification of data with many missing values [6].

In 2013 James D. Dziura et al proposed "Strategies for dealing with Missing data in clinical trials: From design to Analysis" They proposed a recommendations operationalize by providing specific guidance for each stage of the trial. In proposed the design stage, researchers should anticipate missing data patterns and causes and consider methods/designs that encourage participant retention. Developing detailed study documentation, training study personnel and testing operational aspects of the trial are important during the planning stage. Regular monitoring of missing data and enhanced participant contact is recommended for the conduct stage. While easy to implement, ad hoc methods such as complete case analysis and last observation carried forward are not advocated as primary analytic strategies [7].

In 2014 Adam Kapelner et al proposed "Prediction with Missing Data via Bayesian Additive Regression Trees". They present a method for incorporating missing data into general forecasting problems which use non-parametric statistical learning. They focus on a tree-based method, Bayesian Additive Regression Trees (BART), enhanced with Missing value. Incorporated in Attributes, an approach recently proposed for incorporating missing value into decision trees. They extend the work to native partitioning mechanisms found in tree-based models and does not require imputation. Simulations on generated models and real data [8].

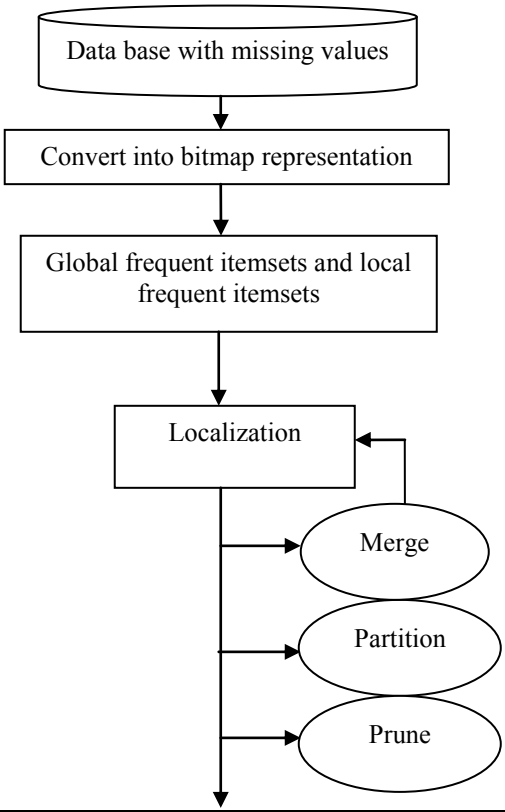
In 2015 Peter Schmitt et al proposed "A Comparison of Six Methods for Missing Data Imputation" . They presented a study and compare 6 different imputation methods: Mean, K-nearest neighbors (KNN), fuzzy K-means (FKM), singular value decomposition (SVD), Bayesian principal component analysis (BPCA) and multiple imputations by chained equations (MICE). Comparison was performed on four real datasets of various sizes (from 4 to 65 variables), under a missing completely at random (MCAR) assumption, and based on four evaluation criteria: Root mean squared error (RMSE), unsupervised classification error (UCE), supervised classification error (SCE) and execution time [9].

In 2016 Barnali Das et al proposed "Missing Data and Imputation" .They discussed about Missing data is a common problems in large data sets, Issues with Missing Data, Approaches to Handle Missing Data, Classifications of Missing Data Examples: MCAR, MAR, MNAR and Imputation Goals[10].

In 2017 Newsom et al proposed "Missing Data and Missing Data Estimation". In the proposed study they discus about List wise Deletion, MAR and MCAR, Determining Whether Missing Values are MAR or MCAR, full information maximum likelihood (FIML), Other Missing Data Approaches and Other imputation methods [11].

IV. PROPOSED APPROACH

Proposed approach is based on following steps



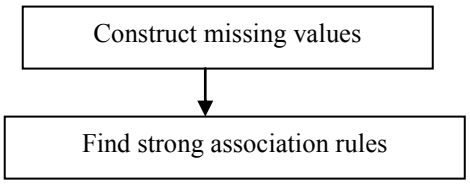


Figure 2. working procedure of proposed approach

Step 1: Find All Possible Frequent Itemsets and Partitions.
Step 2: Finds Global Frequent Item Set
Step 3: Perform Localization
Step 4: Recursively Join the Itemsets in Depth-First Order from the Root

V. EXPERIMENTAL ANALYSIS

We implemented our algorithm in VB Dot Net 2013 and SQL server R2 2010 on an Intel Pentium i3 Processor personal computer with 4 GB RAM on Windows 7 having 400 GB Hard disk. We have taken 100,500 and 1000 transactions with 10 items, transaction containing missing value.

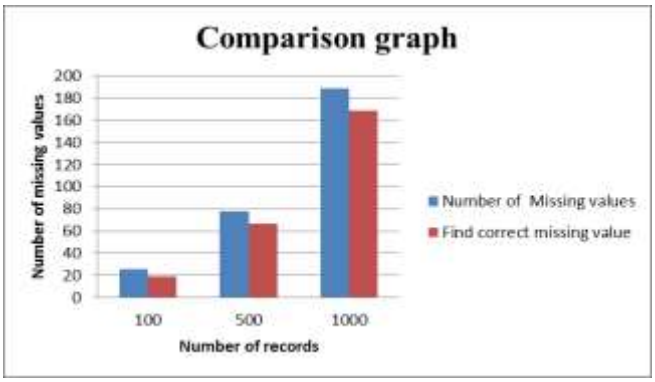


Figure 3. find number of missing values correctly

CONCLUSION

There are several traditional and new approaches are exit to constructs or treatment of missing values. The proposed approach based global frequent value and strong association rules. Proposed approach construct missing value correctly up to 90%.

REFERENCE

1. Toon Calders "Mining Itemsets in the Presence of Missing Values" SAC'07, March 1115, 2007, Seoul, Korea. Copyright 2007 ACM
2. David C. Howell "The Treatment of Missing Data" (Howell, D.C. (2008) The analysis of missing data. In Outhwaite, W. & Turner, S. *Handbook of Social Science Methodology*. London: Sage.)
3. John W. Graham" Missing Data Analysis: Making It Work in the Real World" First published online as a Review in Advance on July 24, 2008
4. Amanda N. Baraldi "An introduction to modern missing data analyses" 0022-4405/$ - see front matter © 2009 Society for the Study of School Psychology. Published by Elsevier Ltd. All rights reserved.
5. Roderick J. Little" Some methods for handling missing values in outcome variables" university of Michigan.
6. Dr. S. S. Dhenakaran" A Perspective Missing Values In Data mining Applications" International Journal of Engineering Trends and Technology- Volume3Issue3- 2012
7. James D. Dziuraa "Strategies for dealing with Missing data in clinical trials: From design to Analysis" YALE Journal of biology and medicine 86 (2013), pp.343-358. Copyright © 2013.
8. Adam Keener "Prediction with Missing Data via Bayesian Additive Regression Trees" The Wharton School of the University of Pennsylvania February 14, 2014
9. Peter Schmitt "A Comparison of Six Methods for Missing Data Imputation Peter Schmitt*, Jonas Mandel and Mickael Guedj Department of Bioinformatics and Biostatistics, Pharnext, Paris, France.
10. Barnali Das "Missing Data and Imputation" NAACCR Webinar May 2016.
11. Newsom, winter "Missing Data and Missing Data Estimation". PSY 510/610 Structural Equation Modeling Little (1988) has a test for MCAR, however, and Enders offers a macro to conduct the test.