International Journal of Research In Science & Engineering e-ISSN: 2394-8299
Volume: 3 Issue: 2 March-April 2017 p-ISSN: 2394-8280

MISSING VALUES IMPUTATION USING HYBRID APPROACH FOR KNOWLEDGE DISCOVERY

Sweety Baiwal [1], Prof. Abhishek Raghuvanshi[2]

*[*1]Research Scholar, MIT Ujjain, M.P, India.*
[2]Assistant Professor, MIT Ujjain, M.P, India.
**Department of Computer Science & Engineering*

ABSTRACT

The training information for potential discovery in databases (KDD) and information mini ng relies on many factors, but handling missing values is viewed to be a principal element i n total knowledge great. In these days actual world datasets includes lacking values due to human, operational error, hardware malfunctioning and lots of other factors. The first-cla ss of advantage extracted, finding out and determination problems rely directly upon the fir st-class of training information. Through for the reason that the value of dealing with lacki ng values in KDD and information mining duties, in this paper we advocate a novel. Associ ation Rule mining from Data with Missing Values is a novel technique to mine association rules from data with missing values. The proposed algorithm adopts Apriori approach, and uses partitioned databases. It consists of two phases: database scrutinizing phase and cons truction of association rule phase. The association rule generation phase consists of the fol lowing two processes: frequent itemset generation and Association rule construction. Data base scrutinizing phase generates database partition and deletes unnecessary attributes whi ch means attributes are not satisfied for user specified minimum representatively threshold from database. Our results endorse that the method shouldn't be most effective higher in t ime period of accuracy however it additionally take much less processing time as in compa rison with present missing values imputation process based on Apriori procedure, which i ndicates the effectiveness of our missing values imputation manner.

Key-Words: - Quality of training data, missing values imputation, association rules mining , data mining, missing values imputation, association rules, categorical data.

I.INTRODUCTION

To address the issues with lacking values in the knowledge occurs in the guidance of informa tion for analyses. The primary possible solution of this situation is decreasing the data set. Th is way of handling missing values continues to be commonly utilized in apply [1] however co uld cause tremendous lack of usable data. The opposite viable answer is lacking values imput ation. Choice of missing values imputation procedure have got to be completed with regard to the constitution of the info set. Many described ways can be used for lacking values imputati on in numerical information, e.g.: mean substitution [2], linear regression [2], neural network s [3] and nearest neighbor technique [4]. The most commonly used system for lacking values imputation in categorical knowledge is to replacement lacking values of each attribute by usi ng the most common value of the attribute [5]. The disadvantage of this approach is that it do es now not don't forget dependencies amongst attributes values. This paper describes three va riants of new algorithm for lacking values imputation with use of association rules and offers results of assessments.

II. LITERATURE REVIEW

Jing Tian ,Bing Yu , Dan Yu , Shilong Ma proposes a new hybrid missing data completion method named Multiple Imputation using Gray-system-theory and Entropy based on Clusteri ng (MIGEC) to impute the missing value attributes in their paper **"**Missing data analyses: a h ybrid multiple imputation algorithm using Gray System Theory and entropy based on clusteri ng". Here the method firstly separates the non-missing data instances into several clusters. T he second step covers the calculations for imputed values by utilizing the information entropy of the proximal category for each incomplete instance in terms of the similarity metric based on *Gray System Theory* (*GST*). In their experiment they use the dataset of *University of Califo rnia Irvine* (UCI) [6]**.**

Subsequently N. Poolsawad, L. Moore, C. Kambhampati and J. G. F. Cleland investigate the characteristics of a clinical dataset using feature selection and classification techniques to deal with missing values and develop a method to quantify numerous complexities. Here the aim is to find the features that have high effect on mortality time frame, and to design method ologies which will cope with missing values. For Missing value imputation their work includ es the K-means clustering and Hierarchical Clustering approach to reveal similarities and rela tionships between attributes and variables having missing values[7]**.**

In recent Archana Purwar and Sandeep Kumar Singh suggested a new approach of missi ng data imputation based on Clustering. In their work the use of clustering based algorithms n amely K-Means, Fuzzy K-Means and Weighted K-Means provides an efficient technique of i mputation. From the large data set of approximately 22,000 tuples, investment patterns of 611 different patterns were taken. The data taken for experiments was taken incomplete. The reas on for taking entire data for missing value experiments was to check the efficiency of the met hods used in the work and that can be efficiently be done with the comparison with the actual and the estimated values [8].

Anupama A Chavan, Vijay Kumar Verma gives the concept of Missing values and incom plete data are usual occurrences in real datasets [1]. The problem of recovering missing value s from a dataset has become an important research issue in the field of data mining and machi ne learning [2]. With the speedy increase in the use of databases, the difficulty of missing val ues unavoidably arises. The techniques developed to effectively recover these missing values should be highly accurate in order to remove the missing values completely. The association r ules are the popular method that is effectively used to establish the relationship among items i n databases. The discovered association rules are useful to recover the missing values in data bases.

K. Rameshkumar presents Missing values and incomplete data are a natural phenomenon in real datasets. If the association rules mine incomplete disregard of missing values, mistaken r ules are derived. In association rule mining, treatments of missing values and incomplete data are important. This paper proposes novel technique to mine association rule from data with missing values from large voluminous databases. The proposed technique is decomposed into two sub problems: database scrutinizes and rules mining phases. The first phase is used to re examine transactions which are useful to mine frequent itemset. The second phase is to mine frequent itemset amd construct association rules from valid database. This paper uses Apriori based algorithm in which proposed technique.

Liu Peng, Lei Lei The topic of missing data has received considerable attention in the last de cade. More and more missing data treatment methods have sprouted-up. Mainly methods for dealing with missing data are compared in this paper. Missing data is a common problem for data quality. Most real datasets have missing data. This paper analyzes the missing data mech anisms and treatment rules. Popular and conventional missing data treatment methods are intr oduced and compared. Suitable environments for method are analyzed in experiments. Metho ds are classified into certain categories according to different characters..

III. EXISTING METHOD

1) Prediction means matching Imputation

Randomization can be introduced by defining a set of values that are closest to the predicted values and choosing one value out of that set at random for imputation. This Imputation meth od combines the parametric and nonparametric methods which impute the missing values by i ts nearest-neighbor donor in which the distance for the missing values are computed from the expected values of the missing data, instead of directly on the values of the covariance. These expected values are computed by a linear regression model.

Predictive mean matching imputation is hot deck imputation within classes where the classes are defined based on the range of the predicted values from the imputation model .This metho d achieves a more even spread of donor values for imputation within classes, which reduces t he variance of the imputed estimator. Donor values within the classes may be drawn with or without replacement, where without replacement is expected to lead to a further reduction in t he variance. The method of predictive mean matching is an example of a composite method, combining elements of regression, nearest-neighbor and hot deck imputation.

2) Repeated Random imputation Method

In single value imputation only one value imputed for each missing value but here in repeated random imputation several times values are imputed. There are two types of repeated imputa tion methods.

- First one is multiple imputation, in which for one missing values there are several values are estimated values.
- Second one is fraction imputation, in which fraction point is added in estimated value every time in repeated forms to determine the new values or to estimate the new values.

The Repeated imputation method is advantageous because it allows one to get good estimates of the standard errors. Single imputation methods don't allow for the additional error introdu ced by imputation.

IV. PROPOSED SYSTEM:

Stating from imputation process a set of association rules are generated from missing values d ata set. After generating association rules utilize these association rules for missing values im putation. For a case if dataset is empty then missing values are imputed using K-nearest neigh bor method.

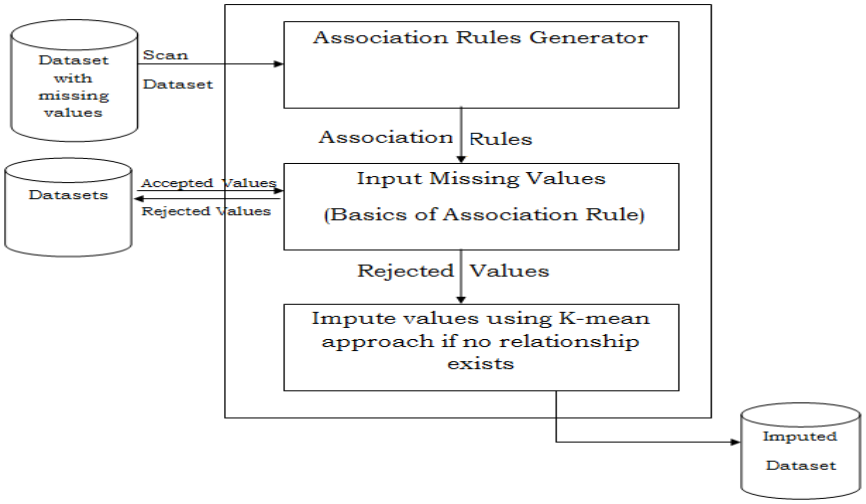

Figure1. Proposed System Architecture

PROPOSED ALGORITHM

Algorithm AssociationRules_Gen(l_k: H_m)

{

// large k-itemset,

//H_m:set of m-item consequents

If(k>m+1)then being

H_{m+1}=apriori-gen(H_m);

For all h_{m+1} E H_{m+1} do being

Conf=support(l_k)/support(l_k- h_{m+1});

If(conf>=minconf) then

output the rule (l_k-h_{m+1})->h_{m+1}

with confidence=conf and support(l_k)

else

delete h_{m+1} from $H_{m+1;}$

The Idea of this work is to apply the K Means, Fuzzy K Means And Weighted K Means Clus tering algorithms and two another clustering algorithms on a adult data seta set. Now the Da ta set is tested for missing values with five different Clustering algorithms and the imputed va lues will be calculated. First we are considering a dataset having 2000 tuples.

V. RESULT:

Comparative Study

Table .1Comparison table for all the methods

Name of Method	Used Concept	Type of Missing Value Handled
EM Method	Partial disables the victim tuples	Handles only one missing value in a Tuples
Multiple Method	Combines at least two sub-frequent item sets to become a frequent combined item set , Recursive RAR to predict multiple missing values	Handles many missing values at a time compared to previous , Handles multiple missing values in a tuples
Regression Method	Establishes the bit-arrays of a relational database table using simple Boolean AND/ OR operations	High & Low Missing Rates
New Approach	Three phases used Roughly assigning values, filling in remaining missing values & adjusting the assigned missing values, Two phases used Database Securitizing & Construction of association rules	Handles maximum missing values at high speed

Association Rule mining from Data with Missing Values is a novel technique to mine association rules from data with missing values. The proposed algorithm adopts Apriori approach, and uses partitioned databases. It consists of two phases: database scrutinizing phase and construction of association rule phase. The association rule generation phase consists of the following two processes: frequent itemset generation and Association rule construction. Database scrutinizing phase generates database partition and deletes unnecessary attributes which means attributes are not satisfied for user specified minimum representatively threshold from database. The association rules are mined by using database partition which is created by the database scrutinizing phase. This phase adopts Apriori technique using transaction reduction based approach. This procedure is divided into two processes as: Generation of frequent itemset and Construction of association rules [1]. The most improved part in this method is that it avoids unwanted database partition process. The method gives good performance even if the missing rate is high or low.

Table 2 Data set of output missing values

0	1	1	3
1	1	2	3.392277
1	2	3	3
1	2	4	3
1.718248	2	5	4
2	2	6	4.392277

e-ISSN: 2394-8299
p-ISSN: 2394-8280

2	3	7	4
2	3	8	4

Table 3 Data set of Missing Values

14.23	1.71	0	15.6	127	2.8	3.06	0.28	2.29	5.64	1.04	3.92	1065	1
13.2	1.78	2.14	11.2	100	0	2.76	0.26	1.28	4.38	1.05	3.4	1050	1
13.16	2.36	0	18.6	101	2.8	3.24	0.3	2.81	5.68	1.03	3.17	1185	1
14.37	1.95	2.5	0	113	3.85	3.49	0.24	2.18	7.8	0.86	3.45	1480	1
13.24	2.59	2.87	21	118	2.8	0	0.39	1.82	4.32	1.04	2.93	735	1
14.2	1.76	2.45	15.2	0	3.27	3.39	0.34	1.97	6.75	0	2.85	0	1
14.39	1.87	2.45	14.6	96	2.5	2.52	0.3	1.98	5.25	1.02	3.58	1290	1
14.06	0	2.61	17.6	121	2.6	2.51	0.31	1.25	5.05	1.06	3.58	1295	1
14.83	1.64	2.17	14	97	2.8	2.98	0.29	1.98	5.2	1.08	2.85	1045	1
13.86	1.35	2.27	16	0	2.98	3.15	0.22	1.85	7.22	1.01	3.55	1045	1
14.1	2.16	2.3	18	105	2.95	3.32	0.22	2.38	5.75	1.25	3.17	1510	1
14.12	1.48	2.32	16.8	95	2.2	2.43	0.26	1.57	5	1.17	2.82	1280	1
14.75	1.73	2.39	11.4	91	3.1	3.69	0.43	2.81	5.4	1.25	2.73	1150	1
14.38	1.87	2.38	12	102	3.3	0	0.29	2.96	0	1.2	3	1547	1
13.63	0	2.7	17.2	112	2.85	2.91	0.3	1.46	7.3	1.28	2.88	1310	1
0	1.92	2.72	20	120	2.8	3.14	0.33	1.97	6.2	1.07	2.65	1280	1
13.83	1.57	2.62	20	115	2.95	3.4	0.4	1.72	6.6	0	2.57	1130	1
14.19	1.59	2.48	16.5	108	3.3	0	0.32	1.86	8.7	1.23	2.82	1680	1
12.93	3.8	2.65	18.6	102	2.41	2.41	0.25	1.98	4.5	1.03	3.52	770	1
13.71	1.86	2.36	16.6	101	2.61	2.88	0.27	1.69	3.8	1.11	4	1035	1
12.85	1.6	2.52	17.8	95	2.48	2.37	0.26	0	3.93	1.09	3.63	1015	1
13.5	1.8	2.6	20	96	2.5	2.6	0.2	1.6	0	1.1	3.8	845	1

	1	1			3	1	8	6		2	2		
13.05	2.05	3.22	25	124	2.63	2.68	0.47	1.92	3.58	1.13	3.2	0	1
13.39	1.77	2.62	16.1	93	2.85	2.94	0.34	0	4.8	0.92	3.22	1195	1
13.3	1.72	2.14	17	94	2.4	2.19	0.27	1.35	3.95	1.02	2.77	1285	1

Table 4. Dataset of Tested Missing Values

A	B	C	D
0	1	1	3
1	1	2	0
1	2	3	3
1	2	4	3
0	2	5	4
2	2	6	0
2	3	7	4
2	3	8	4

Above data set define the missing values data and methods are applied on it and then find the results that results are efficient and flexible. New Method provide better results than other me thods. The ratio of missing values is set as 25%, MinConf threshold is set as 70% and MinRe pr is set as 50%.

CONCLUSION

In this paper we have investigated the exclusive techniques for missing price imputation and dimensionality reduction. We tried to realize and in finding the suitable tactics for establishin g the model for examining the have an impact on of lacking circumstances in a dataset. Besid es this, the key component is to have an understanding of the character of the dataset so as to select the compatible system. The important effects of this wide gain knowledge of will supp ort in selecting the appropriate strategies for lacking knowledge dealing with problems.

Future Scope

Our results recommend that missing values imputation utilizing our manner has just right abil ities in time period of accuracy and can be a excellent procedure in term of processing time. I

n future we enhance this factor by using merging some methods. Hope so they give extra bett er results than this one.

REFERENCES

[1] Sun S.,Yen J., "Information supply chain: A unified framework for information-sharing", Intelligence and Security Informatics. Springer 2005, 422–428.

[2] Shankaranarayan G., Ziad M., and Wang R. Y., "Managing data quality in dynamic decision environments: An information product approach". J. Datab. Manage 2003. 14, 14–32.

[3] Ludmila Himmelspach, Stefan Conrad, "Clustering Approaches for Data with Missing Values: Comparison and Evaluation", IEEE 2010.

[4] Horton N. J., Kleinman K. P., "Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models", Amer. Statist. 2007 61, 79–90

[5] Raquel Mart´ınez, Jos´e M. Cadenas, M. Carmen Garrido and Alejandro Mart´ınez**, "**Imputing Missing Values from Low Quality Data by NIP Tool", IEEE International Conference 2013.

[6] Jing Tian, Bing Yu, Dan Yu, Shilong Ma, **"**Missing data analyses: a hybrid multiple imputation algorithm using Gray System Theory and entropy based on clustering", Springe**r** Science+ Business Media New York 2013.

[7] N. Poolsawad, L. Moore, C. Kambhampati and J. G. F. Cleland, "Handling Missing Values in Data Mining - A Case Study of Heart Failure Dataset" **,** 2012 9th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2012)IEEE 2934-2938.

[8] Archana Purwar, Sandeep Kumar Singh, "Empirical Evaluation of Algorithms to impute Missing Values for Financial Dataset" IEEE 2014.

[9] Bhavisha Suthar, Hemant Patel, Ankur Goswami, "A Survey: Classification of imputation methods in data mining", IJETAE Volume 2, Issue 1, January 2012.

[10] Gabriele B. Durrant, "Imputation Methods for Handling Item- Non response in the Social Sciences: A Methodological Review" ESRC National Centre for Research Methods and Southampton Statistical Sciences Research Institute (S3RI). University of Southampton, 2005.

[11] R. Kavitha Kumar, Dr. R.M. Chadrasekar, "Missing data Imputation in Cardiac Dataset(Survival Prognosis)",International Journal on Computer Science and Engineering Vol. 02, No. 05, 2010, 1836-1840 .

[12] Gustavo E. A. P. A. Batista, Maria Carolina Monard, "A Study of K -Nearest Neighbor as an Imputation Method", University of Sao Paulo USP, 2002.

[13] Jiri Kaiser,'' Dealing with Missing Values in Data", Journal of Systems Integration 2014.

[14] R. Srikant, Q. Vu and R. Agrawal,, "Mining association rules with item constraints," The Third International Conference on Knowledge Discovery and Data Mining, 1997

[15] R. Agrawal and R. Srikant, "Fast algorithm for mining association rules," The International Conference on Very Large Data Bases, 1994

[16] R. Agrawal, T. Imielinksi and A. Swami, "Database mining: a performance perspective," IEEE Transactions on Knowledge and Data Engineering, Vol. 5, No. 6, 1993

[17] R. Agrawal, T. Imielinksi and A. Swami, "Mining association rules between sets of items in large database, " The 1993 ACM SIGMOD Conference on Management of Data, 1993