Imputation of Missing Values using Association Rule Mining & K-Mean Clustering

[1]Sweety Baiwal, [2]Abhishek Raghuvanshi

M.Tech. Scholar, Assistant Professor
Mahakal Institute of technology, Ujjain

***ABSTRACT*: The data mining architecture works on facts and figures which are used for any type of decision making. To perform any analysis and decision making, these facts must be complete so that the analyst can make a strategy for decision making. In fact the most important problem in knowledge discovery is the missing values of the attributes of the Dataset. The presence of such imperfections usually requires a preprocessing stage in which the data are prepared and cleaned, in order to be useful, and sufficiently clear for the knowledge extraction process. In this paper we are created hybrid approach for imputation or Replacement of the missing values. In Hybrid approach we use association rules and K-Nearest Neighbor methods. These methods can work with text dataset, Boolean dataset and with numeric dataset. We also analysis the parametric, non-parametric and semi-parametric imputation methods.**

***Keywords*: Data Mining, Missing Values, Imputation, Feature Selection, Parametric, Non Parametric, Semi Parametric.**

1. INTRODUCTION

With access to vast volumes of data, decision makers frequently draw conclusions from data repositories that may contain data quality problems, for a variety of reasons. In decision making, data quality is a serious concern. The incidence of data quality issues arises from the nature of the information supply chain [1], where the consumer of a data product may be several supply-chain steps removed from the people or groups who gathered the original datasets on which the data product is based. These consumers use data products to make decisions, often with financial and time budgeting implications. The separation of the data consumer from the data producer creates a situation where the consumer has little or no idea about the level of quality of the data [2], leading to the potential for poor decision-making and poorly allocated time and financial resources. Missing data, that is, fields for which data is unavailable or incomplete, is particularly important problem, since it can lead the analysts to draw inaccurate conclusions. When data is extracted from a data warehouse or database (a common occurrence when aggregating data from multiple sources), it typically passes through a cleansing process to reduce the incidence of missing or noisy values as far as possible. Some missing data attributes cannot be fixed because the database manager may not have any way of knowing the value that is missing or incomplete, as would happen. At this point, the database manager may choose to remove the records with missing data, at the loss of the power of the other data attributes that contained in these records, or to provide the dataset with missing data records included.

Data values may be missing for a variety of reasons, falling into two general categories: Missing At Random (MAR) and Missing Not At Random (MNAR) [3]. In the MAR case, the incidence of a missing data value cannot be predicted based on other data, while in the MNAR case, there is a pattern among the missing data records. For MAR scenarios, there exist methods for estimating, or imputing the missing values [4]. The incidence of missing data, however, often falls into the MNAR category; that is, there is bias in the occurrence of null or missing values. This missing may occur for a variety of reasons. For example, survey respondents may refuse to answer questions that they feel reveal information that is too personal, often based on religious, cultural, or gender norms. Bias may also occur due to a natural human tendency to suppress unfavorable information. For example, in health care informatics, patients may choose not to report unhealthy behaviors that increase risks for certain illnesses. In MNAR scenarios, if an analyst assumes there is no bias in the pattern of the missing data, the conclusions drawn may be inaccurate. Clearly, for MNAR scenarios, it would be useful to uncover the patterns in the incidence of missing data in a dataset. Once patterns are uncovered, several steps can be taken, depending on the context in which the bias is found.
If the decision maker is the creator or owner of the dataset, preventative steps can be taken to prevent the occurrence of missing data by improving data collection methods given the information on possible biases.

The rest of the paper is organized as follows: In section 2, we have briefly discussed the literature review of the past work. Section 3 shows the different approaches for missing value handling. The different types of imputation techniques are then elaborated in section 4. Finally the conclusion is presented in section 5.

2. LITERATURE REVIEW

The data set which contains the raw data may not have any type of bonding between the data items. It means that there is no relevance between two data items. Such kind of dataset is called perfect dataset. But there is also a probability that the data items may have strong relationship among themselves. This type of dataset is called as Dependent dataset. So the problem is to identify the bias patterns in the case of Dependent dataset.
To identify these patterns Raquel Mart´ınez, Jos´e M. Cadenas, M. Carmen Garrido and Alejandro Mart´ınez proposes a technique to carry out the imputation of missing values within the data set that may be of low quality. They

had applies a predictive model to identify and impute the missing attribute values. The imputation method is incorporated into the software tool NIP increasing its functionality of imputation/ replacement of low quality values. The algorithm uses the *K-means Nearest Neighbor* as an imputation method. The K-Means Nearest Neighbors imputation has been incorporated into the NIP tool in order to increase the functionality of replacement/imputation [5]**.**

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 Clustering (MIGEC) to impute the missing value attributes in their paper **"**Missing data analyses: a hybrid multiple imputation algorithm using Gray System Theory and entropy based on clustering". Here the method firstly separates the non-missing data instances into several clusters. The 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 California 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 methodologies which will cope with missing values. For Missing value imputation their work includes the K-means clustering and Hierarchical Clustering approach to reveal similarities and relationships between attributes and variables having missing values. They had applied the programming methods to optimize and compute the matrices of the clinical data set which are having missing information for certain diseases. Three feature selection techniques; *t*-Test, entropy ranking and nonlinear gain analysis (NLGA) are employed to identify the most common features within the data set [7]**.**

In recent Archana Purwar and Sandeep Kumar Singh suggested a new approach of missing data imputation based on Clustering. In their work the use of clustering based algorithms namely K-Means, Fuzzy K-Means and Weighted K-Means provides an efficient technique of imputation. 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 reason for taking entire data for missing value experiments was to check the efficiency of the methods used in the work and that can be efficiently be done with the comparison with the actual and the estimated values. The experimentation is divided into two modules. One module compares above three imputation methods with 100 instances. To check the accuracy; missing values were artificially introduced in the dataset. Experiments were repeated for 1 %, 3%, 6%, 9%, and 12% missing values to increase the accurate analysis. 1 % of missing values gives the lowest RMSE in case of K-means. But K- means is performing worst when 2 % of the missing values were generated. When the percentage of missing values increases, weighted k-means and fuzzy k-mean gives almost similar results where as k-means gives the lowest RMSE value. Second module repeats the first experiment with 600 instances for 2%, 4%, 6%, 10%, 12%, and 14% missing values. It was observed in the second module that Fuzzy k-means and Weighted distance K-mean gives the same RMSE and underperformed as compared to K- means imputation. Hence K-means imputation is best in the second experiment also [8].

3. MISSING DATA HANDLING APPROACHES

In this section we are discussing the handling of missing attributes and instance values within the raw data set. Here the relevance factor between data items is also taken into account so that any kind of relationship that exit between data items is identified. The serious problem that arises to handle missing values is, in the case of Dependent data set where, there is a strong bonding between the past and the present data of any historical data set. On the basis of the type of relationship the Missing data handling approaches lies in the following three categories [9]**.**

i) Avoiding and discarding data: These methods determine the missing data on each instance and delete those instances. Other thing is determining each attribute or instances variable and remove that whole attribute or instances value which is having high level of missing data. This method is applicable only when the dataset is MCAR type where the data items are missing randomly.

ii) Estimation of Parameters: This method is used to find out the parameters for the complete dataset. This method is use the Expectation-maximization algorithm for handling the parameter estimation of the missing data attributes.

iii) Imputation techniques: This is one kind of procedure in which replacement of the missing values is done based on some estimated values derived from the existing set. The missing attributes are derived from the present set of entries and the complete data set is formed for any kind of analysis or knowledge discovery purpose.

4. TYPES OF IMPUTATION TECHNIQUES

Imputation is a term for the replacement of the missing values by some predictable set of values which said to be plausible values in the dataset [10]. It makes use of observed supporting information for cases with non-response maintain high accuracy.

4.1 Simple imputation or mean Imputation

Different approaches for imputation, like unconditional mean and mode imputation in which respectively in continuous dataset and discrete dataset missing values are replaced with its mean or mode of all known values of that attributes [10]. Simple imputation methods are C4.5 and KNN method, and so on. That is why this imputation is also known as mean imputation [11].

4.2 Regression imputation

This is one of the broad methods for imputing missing values. There are different regression techniques [10].

4.2.1 The Predictive Regression

In this, the linear regression which is used for numeric variables and logistic regression is used for categorical data. The linear regression works with linear functions based on probability, and the logistic regression works on logistic functions based on probability but it has only two possibilities for probability. In regression method predictive regression imputation which uses auxiliary variable to find the missing values which relates missing values Yi to

auxiliary variable Xi and the predicted values used for the missing values in Y.

4.2.2 The Random Regression

The random regression imputation method is used to find the missing values for any variable based on conditional distribution. It imputes the value based on conditional distribution of Y given X. It is more effective for numeric datasets.

4.2.3 Hot deck and cold deck imputation

This is the method which is applied when the components of the dataset are skewed (twisted). The imputed values have the same distributional shape as observed data. Hot deck is implemented in two stages. In First stage the dataset is divided into clusters. In second stage the instances with missing data in the dataset is associated with one cluster. This calculates the mean or mode of the attributes within the cluster. In this method the donor came from the same data source. Cold deck imputation is contrast than the hot deck imputation because it select donor from the other data source [9].

4.2.4 *Prediction mean 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 method combines the parametric and nonparametric methods which impute the missing values by its 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 method achieves a more even spread of donor values for imputation within classes, which reduces the 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 the variance. The method of predictive mean matching is an example of a composite method, combining elements of regression, nearest-neighbor and hot deck imputation.

4.2.5 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 imputation 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 introduced by imputation.

4.2.6 K-Nearest Neighbor imputation method

The k-nearest neighbor algorithm also called as KNN is used to estimate and substitute the missing values. It works as a lazy learner. It outperforms internally C4.5 and CN2 for handling missing values and also mean and median for handling missing values [12].

The main three tasks needed in KNN method are:

- An integer K (decide the how many neighbor taken for estimating the missing values in each iteration)
- A set of labeled example (training data)
- A metric to measure "closeness"

In this method the main factor is Distance metrics. In 1NN imputation method we can replace the missing values with the nearest neighbor. But if the value of K is greater than one then replace the missing values with the mean of K-nearest neighbors. Here we discuss the one imputation method which is said to be regression imputation which we already discussed above. The imputation methods for missing values include parametric regression imputation methods and non-parametric regression imputation methods. Non-parametric imputation models are k-nearest neighbor or kernel regression. In which there is no relationship between depended and independent variables.

Another category of this is the parametric imputation model in which we know the part of relationship between depended and independent variables.

4.2.7 Imputation Using Neural Networks

Neural networks constitute a class of predictive modeling system that works by iterative parameter adjustment. The network structure, also called as topology or architecture, which includes the neural framework (number of neurons, number of layers, neuron model type, etc.) and the interconnection structure. Single-layer network has only an input and output layer. In a multilayer network, one or more hidden layers are inserted between the input and the output layer [13].

Gupta, A. & Lam, M. (1998) introduced following procedure for reconstruction of missing values using multilayered networks and back propagation algorithm:

Step 1: Collect all training cases without any missing value and call them the complete set.

Step 2: Collect all training and test cases with at least one missing value and call them as incomplete set.

Step 3: For each pattern of missing values, construct a multi-layered network with the number of input nodes in the input layer equal to the number of non-missing attributes, and the number of output nodes in the output layer equal to the number of missing attributes. Each input node is used to accept one non-missing attribute, and each output node to represent one missing attribute.

Step 4: Use the complete set and the back propagation algorithm to train each network constructed in step 3. Since the complete set does not have missing values, different patterns of input-output pairs can be obtained from the complete set to satisfy the input-output requirements for different networks from step 3. As the output of a network is between 0 and 1, data have to be converted to values between 0 and 1 for this reconstruction procedure.

Step5: Use the trained networks generated from step 4 to calculate the missing values in the incomplete dataset.

4.2.8 Imputing missing values Using Association Rules

An association rule is a simple probabilistic statement about the co-occurrence of certain events. For binary variables association rule takes the following form:

IF A=1 AND B=1 THEN C=1

Here the input for the missing values imputation is the incomplete data set. Algorithms for association rules generation are usually unable to handle missing values [13]. Some association rules extraction methods may ignore rows with missing values (conservative approach) or handle missing values as they are supporting the rules (optimistic approach) or are in contrary with the rules (secured approach).
Useable association rule for missing values imputation has the consequent containing value of attribute whose value is being searched and the antecedent correspond to values of other given case attributes. If complete data set for association rules generation was obtained by replacing missing values by special values association rules with these special values in consequent must be omitted. It is possible to use only association rules with consequent length equal to 1 and impute missing values one by one in cases with more than one missing values. Another possibility is using association rules with consequent length equal to the count of missing values in given case.

Support and confidence of the rule are two criteria that should be maximized during making decision about using the association rule for missing value imputation. For missing value imputation can be used the rule with maximum confidence along with required support. It is also possible to ignore support of the rule and use only confidence.

5. PROPOSED SYSTEM:

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

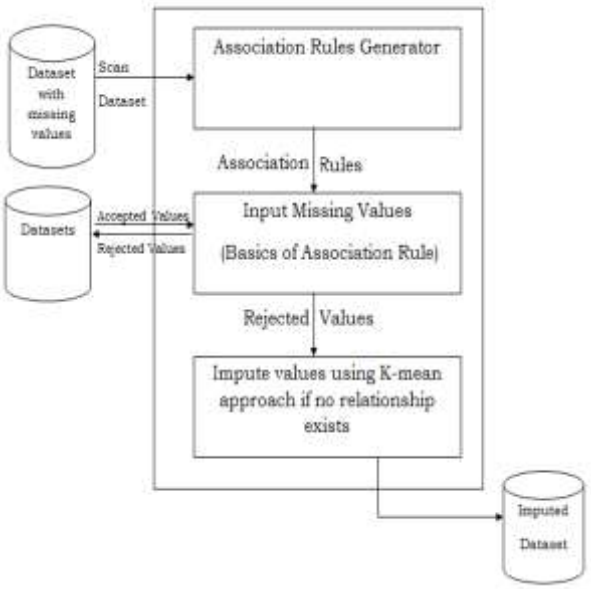

Figure 1. Proposed System Architecture

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 $H_{m+1;}$

end if

call gen-ap(l_k , H_{m+1});

end for

end if

end procedure

Figure 2. Proposed Algorithm for Association Rules

6. CONCLUSION

In this paper we have investigated the different techniques for missing value imputation and dimensionality reduction. We attempted to understand and find the suitable techniques for developing the model for analyzing the impact of missing instances in a dataset. Besides this, the key factor is to understand the nature of the dataset in order to choose the suitable technique. The important outcomes of this extensive study will help in choosing the appropriate techniques for missing data handling problems.

7. 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.