International Journal for Rapid Research in Engineering Technology & Applied Science
ISSN (Online): 2455-4723
Vol 2 Issue 7 August 2016

"Imputation of Missing Values using Hybrid Approach with Association Rule and KMean Clustering Algorithm "

Neelesh Shrival*[1],Kapil Sahu[2]
*[1]Research Scholar, Sanghvi Institutes of Management and Science Indore , M.P,India.
[2]Assistant Professor, Sanghvi Institutes of Management and Science Indore, M.P,India.
*Department of Computer Science & Engineering.

Neeleshshrival19@gmail.com*[1], Sahukapil@gmail.com*[2]

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 thesis presenting the Comparative study of the different method employed for Imputation or Replacement of the missing values. These methods can work with text dataset, Boolean dataset and with numeric dataset. We have discussed 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 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.

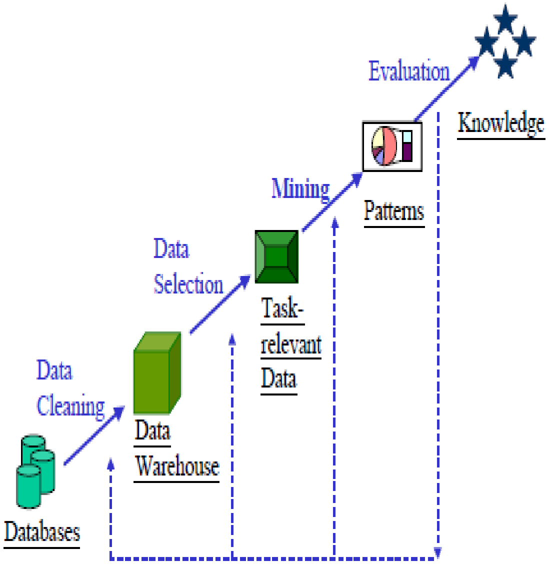

Figure 1 The process of knowledge discovery in databases

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

2. 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 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

3. EXISTING SYSTEM

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].

Regression imputation

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

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.

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.

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].

Prediction mean matching Imputation

Defining a set of values that are closest to the predicted values and choosing one value out of that set at random for imputation can introduce randomization. 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.

4. PROPOSED WORK:

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.

```
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
```

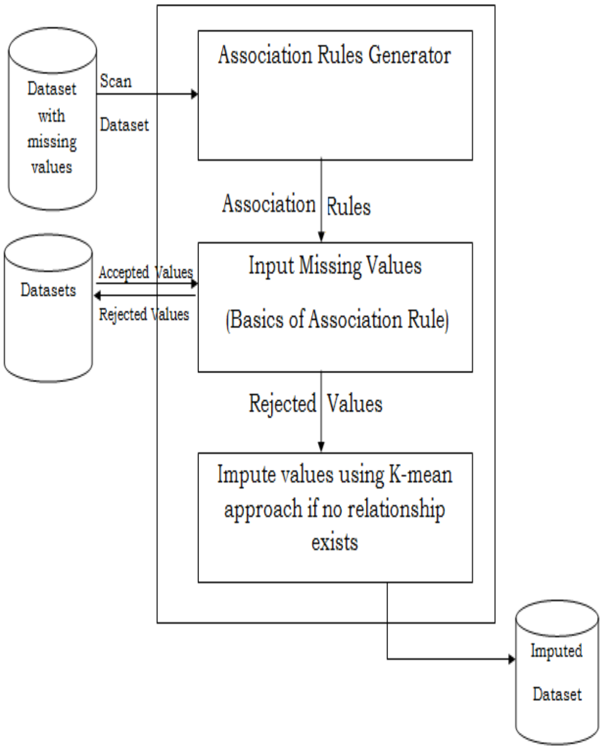
Dataset with missing values
Scan
Dataset
Association Rules Generator
Association Rules
Datasets
Accepted Values
Rejected Values
Input Missing Values
(Basics of Association Rule)
Rejected Values
Impute values using K-mean approach if no relationship exists
Imputed Dataset

Figure 2 Proposed System Architecture

PROPOSED ALGORITHM

Figure 3. Proposed Algorithm for Association Rules

The Idea of this work is to apply the K Means, Fuzzy K Means And Weighted K Means Clustering algorithms and two another clustering algorithms on a adult data seta set. Now the Data set is tested for missing values with five different Clustering algorithms and the imputed values will be calculated.

5. CONCLUSION:

In this Thesis 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.

Our results suggest that missing values imputation using our technique has good potential in term of accuracy and is also a good technique in term of processing time. In future we enhance this thing by merging some methods. Hope so they give more better results than this one.

REFERENCE:

[1]Peter Mell and Tim Grance, “The NIST Definition of Cloud Computing”, NIST, 2010.

[2]Akhil Behl, “Emerging Security Challenges in Cloud Computing**”**, in Proc. of World Congress on Information and communication Technologies ,pp. 217-222, Dec. 2011.

[3]Srinivasarao D et al., “Analyzing the Superlative symmetric Cryptosystem Encryption Algorithm”, Journal of Global Research in Computer Science, vol. 7, Jul. 2011

[4] Tingyuan Nie and Teng Zhang “A study of DES and Blowfish encryption algorithm‟, in Proc. IEEE Region 10 Conference, pp. 1-4 ,Jan. 2009.

[5] Jitendra Singh Yadav et al.,“ Modified RSA Public Key Cryptosystem Using Short Range Natural Number Algorithm” , International Journal of Advanced Research in Computer Science and Software Engineering ,vol. 2,Aug. 2012.

[6] Manikandan.G et al., “A modified cryptographic scheme enhancing data”, Journal of Theoretical and Applied Information Technology**,** vol. 35, no.2, Jan. 2012.

[7] Nilesh Mangtani and Sukhada Bhingarkar, “ The appraisal and Judgment of Nimbus, OpenNebula and Eucalyptus”, International Journal of Computational Biology , vol. 3, issue 1, pp 44-47, 2012.

[8] A. Juels and B. S. Kaliski, Jr., (2007) ―*Pors: proofs of retrievability for large files,”* in CCS ‘07: Proceedings of the 14th ACM conference on Computer and Communications security. New York, NY, USA: ACM, 584–597.

[9] Cody, Brian; Madigan, Justin; MacDonald, Spencer; Hsu, Kenneth W.;, "*High speed SOC design for blowfish cryptographic algorithm,*" Very Large Scale Integration, 2007. VLSI SoC 2007. IFIP International Conference on , vol., no., pp.284-287, 15-17 Oct. 2007.

[10] Govinda.K1 Mythili and Geetha Priya(2014),‖ *Data Security in Cloud using Blowfish Algorithm*‖, International Journal for Scientific Research & Development| Vol. 2, Issue 09.

[11] J. Guo, S. Ling, C. Rechberger, and H. Wang, ―*Advanced Meetin-the-Middle Preimage Attacks: First Results on Full Tiger, and Improved Results on MD4 and SHA-2*,‖ pp. 1–20.

[12] Gurpreet Kaur and Manish Mahajan (2013), ―*Analyzing Data Security for Cloud Computing Using Cryptography Algorithms*‖, International Journal Of Engineering Research and Application, Vol.-3,782-786.