Survey paper on Big data imputation and Privacy algorithms

G.Swetha[1], G.Ramya[2]

[1,2] *Professor,CSE,CVRCE,India*

Abstract - *Big data is a collection of large data sets that traditional processing methods are inadequate to deal with them. however , the fast growth of such large data generates both opportunities and problems. This paper presents the literature review about issues, data creation ,data protection and also different algorithms to deal with the issues.*

Key Words: **Big Data, Imputation, nearest neighbour, data protection , Data Distortion, data blocking.**

1.INTRODUCTION

Goods and Services tax was introduced in India from July 1st2017.People from all over the nation have given their feedback on it. Some people have given positive feedback and some have given negative feedback on it. If we can summaries all types of opinions including updated ones, we can consider it as a good example for Big data. Maximum percentage of the data in the world were produced within the last few years[2].Data is coming from various sources and in various formats. Especially social networking sites are producing large amount of data every hour and handling this large data is very difficult.

Big data challenges [7] include Capturing, data storage, data analysis, search, sharing, transfer, visualization, querying, and updating and information privacy

The paper is organized as follows. Chapter II gives an introduction to data imputation and algorithms for missing data replacement. chapter III gives an introduction to privacy protection and algorithms for privacy protection. IV. Conclusion ,

2. Data Imputation

Normally when we preprocess data in data mining, we miss some of the attribute values. But we can extract knowledge from the data only if the data has good quality that is without missing values. But if we have missing data we cannot get good quality data. Missing data may occur because of a detained student in a class, not responding to the questions in a survey and so on. If we can handle missing data carefully, then we can increase the quality of the knowledge. So we need to replace the missing data with some other reasonable data. This is known as data Imputation.

If we have knowledge on that data we can predict the missing value, but it is very complicated. Data may be missed in columns or rows or in both. Data which is missed can be replaced before Data mining starts or after it starts. This paper is a survey on 2 methods for handling missing data. First method is Refined Mean Substitution and Second method is K-Nearest Neighbor for missing data.

2.1 Data Imputation Algorithms:

The paper[1],proposed an algorithm for missing data. Here missing data is estimated by using an Euclidean distance of the missing instances or attributes and remaining records. In this method distance(d) is calculated between approximately imputed data set and rows of the data set. Now we need to find data whose value is greater than the mean of d. Now name this data as I. That is I is the index elements whose distance is higher than mean(d).Now we need to find mean (μ_j)of elements $D_{new}(I,n)$.Now for all the missing values we need to replace μ_j in rows of missing data.By calculating for every row like this and by substituting in every missing place ,finally the imputed data set will be generated. This algorithm was evaluated with five different metrics. The performance is evaluated in terms of RAND INDEX, Performance in terms of Accuracy, Performance in terms of Specificity, Performance in terms of sensitivity, and performance in terms of Mean Square Error. According to [1],in almost all the cases this algorithm performed better than MC/mean value substitution method.
The second algorithm for[8] imputation is K-Nearest Neighbors.

Features of k-Nearest Neighbor are:

1).All the values of the attributes correlate with in an n-dimensional Euclidean space.
2).When a new attribute value is entered, then classification begins.
3).Different points' feature vector is compared for doing classification.
4).Here we don't use any particular function, it may be discrete or real valued.
5).Euclidean distance between any two values will be calculated. Mean value of the k-nearest neighbors will be taken.

According to [4], classes for missing data randomness are:

(1).Missing completely at random: Here probability of the missing value does not depend on existing value or itself. So, we can do imputation with any data.

(2).Missing at random: Here probability of missing data depends on known values but not itself.
(3).Not missing at random: Here probability of the missing data depends on itself.

According to [4], Missing data handling methods are:

(1).We can completely delete all instances of missing values or attributes or we can check whether any particular attribute or instance is missing in higher levels also then we can delete it.
(2).We can use algorithms which can handle estimation of parameter in the presence of missing data.
(3).we can replace with some reasonable value in the missing data, which is known as imputation.

Imputation using k-nearest neighbor[3]:

According to [3],the main advantages of this method are:
(1).k-nearest neighbor can predict the missing value by considering the most frequent value among the k-nearest neighbors, and it can find mean among the k-nearest neighbors and substitute it.
(2).Here it is not required to have a model which guesses the value of the missing attribute , thats why here we can use any attribute as class because we are not using any specific model.

The main drawback of this model is: As we need to see for the most frequent instance, the algorithm searches all the data set, As the database is very large it will be difficult for KDD.

3.Privacy Protection:

In recent years, the privacy and personal data protection has become an issue especially in the context of social networking and online advertisement. personal data means any kind of data which identifies an individual person. examples are person name, address, phone number, identity number, date of birth. the way data is growing exponentially, it will change the world that scarcely imagine today. that is why the protection of personal data is very important. Safeguards are necessary to give citizens and consumers trust in administration, business and other private entities.

Data Privacy Algorithms:

Privacy preservation using association rule hiding:

Association rule hiding algorithms are used to hide sensitive data. Suppose a database ‘D’ is available with minimum support and confidence and set of rules ‘R’ are mined. A subset ‘$R_{S'}$ set of sensitive association rules where’ R_S ‘is subset of ‘R’. The aim of association rule hiding algorithms are to change the database in such a way that it will be difficult to mine sensitive association rules by maintain remaining rule unaffected[5],

Classification of privacy preserving association rule hiding algorithm:

1. Heuristic –Based Techniques
2. Border Approach
3. Exact Approach
4. Reconstruction based association Rule
5. Cryptography based Techniques
6. Hybrid technique approach

3.1 Heuristic- based techniques:

Heuristic based techniques directly modify the data to hide sensitive information. Based on the modification of data, this technique is divided into two groups: Data distortion method and Data Blocking method.

a. Data distortion method:

Data distortion methods works by adding some noise or unknown values. These distortion methods must preserve the privacy and at the same time must keep the utility of data after distortion. The classical data distortion methods are based on random value perturbation. Below functions are two random value perturbation functions.

i. Uniformly distributed noise:

In this method a noise matrix is added to the original matrix. And noise[6] matrix is generated with the uniform distribute function in a given interval of values.

ii. Normally distributed noise:

This method is same as previous method but here noise matrix is generated with the help of normal distribution function[6] using mean and standard deviation.

b. Data blocking method:

Data blocking method works by reducing degree of support and confidence [6] of association rule. To get less value this method replaces the attribute values with the values that give low support count.

4. CONCLUSIONS

Big data is collection of large amount of structured, unstructured form of data coming from different sources.It has both advantages and disadvantages.

In order to solve problems of big data challenges, many researchers proposed a different system models, techniques for big data In this paper we discussed about the two issues related to big data mining. The two issues are problems while collecting the data and data protection. we also discussed algorithms like k-nearest neighbor for missing data and association rule hiding for data privacy protection.

REFERENCES

[1] R.S. Somasundaram1 and R. Nedunchezhian2."Missing Value Imputation using Refined Mean Substitution"IJCSI International Journal of computer science issues,vol.9,issue 4,No 3,July 2012 ISSN(online):1694-0814.

[2]]"IBM What is Big Data:Bring Big Data to the Enterprise ,"http://www01.ibm.com/software/data/bigdata/,IBM, 2012.

[3] ."A Study of *K*-Nearest Neighbour as an Imputation Method" Gustavo E. A. P. A. Batista and Maria Carolina Monard.*University of S˜ao Paulo – USP,*Institute of Mathematics and Computer Science – ICMC, Department of Computer Science and Statistics – SCE, Laboratory of Computational Intelligence – LABIC, P. O. Box 668, 13560-970 - S˜ao Carlos, SP, Brazil, *{*gbatista, mcmonard*}*@icmc.usp.br

[4] R. J. Little and D. B. Rubin. *Statistical Analysis with Missing Data*. John Wiley and Sons, New, York, 1987

[5] Mohamed Refaat Abdellah ,H. Aboelseoud M , Khalid Shafee Badran , M. Badr Senousy ,"Privacy Preserving Association Rule Hiding Techniques: Current Research Challenges ",International Journal of Computer Applications (0975 – 8887) Volume 136 – No.6, February 2016 .

[6] Jun Zhang and Jie Wang, University of Kentucky, USA Shuting Xu, Virginia State University, USA ,Matrix "Decomposition-Based Data Distortion Techniques for Privacy Preservation in Data Mining ."

[7] Jaseena K.U.1 and Julie M. David2, "Issues,Challenges and Solutions : Big Data Minig."

[8] G. E. A. P. A. Batista and M. C. Monard. K-Nearest Neighbour as Imputation Method: Experimental Results (in print). Technical report, ICMC-USP, 2002. ISSN-0103-2569.