International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor:6.887
Volume 5 Issue IX, September 2017- Available at www.ijraset.com

A Hybrid approach for Privacy Preservation of Big data

Gayatri[1], Rajneesh Narula[2]
[1, 2] *Department of Computer Science and Engineering, Maharaja Ranjit Singh Punjab Technical University*

Abstract: ***Now a days privacy preservation gains attention in order for secure transmission of data over the internet. A large amount of data is published every year for analytical and research purposes thus, it is important to use data mining. There are various privacy preserving data mining techniques as perturbation, anonymization, randomization, cryptography but still leave gaps in maintain privacy. Thus, it is necessary to propose a new technique termed as hybrid technique that fills the gaps and maintains privacy.***
Keywords: anonymization, randomization, cryptography, perturbation, dependency score

I. INTRODUCTION

A large amount of data is published over the internet every year which can be used for various data mining purposes. *Benjamin C.M Fung[1]* explained in his paper that the published data is usually in the form of table, a row can have many attributes which are categorized as (1)Explicit identifiers: the attributes that distinctly identify an individual as *name*, *address* and so on. (2)Quasi identifiers: the attributes when combined together can intelligibly identify an individual such as combination of *age, zipcode and gender* and so on. (3) Sensitive attributes are those that need not to be revealed such as *Salary* and *disease*.We can apply various privacy preserving data mining techniques as Randomization, Perturabation, Anonymization, and Cryptography on different categories of attributes to achieve privacy. We will discuss all these techniques in detail in the following section.

II. PRIVACY PRESERVATION DATA MINING TECHNIQUES

This section provides the basic concepts behind the various privacy preservation data mining techniques:

A. Anonymization

Data anonymization can be described as a process to remove personally recognizable information from the data sets to preserve privacy. It can further be categorized as k-anonymity, l-diversity and t-closensess. According to census summary data 1991, 87% of U.S population can be uniquely recognizable based on the set of three attributes as *5-digit Zip code*, *Birthdate* and *Gender*. *Latanya Sweeney [2]* introduced the concept of k-anonymity. The data is said to be k-anonymized if the record can not be distinguishable from atleast k-1 records being present in the dataset.

Table 1: Original Data Table

Name	Age	Gender	Race	Disease
Amit	23	Male	White	Cancer
Mrunal	19	Female	Black	Viral-infection
Kriti	18	Female	White	Heart related
Irfan	27	Male	White	Heart related
Arijit	29	Male	Black	No illness

Table 2: Anonymized Data Table

Name	Age	Gender	Race	Disease
*	20<Age<=30	Male	*	Cancer
*	Age<=20	Female	*	Viral-infection
*	Age<=20	Female	*	Heart related
*	20<Age<=30	Male	*	Heart related
*	20<Age<=30	Male	*	No illness

Table 1 shows the original data table and Table 2 shows anonymized data table in which the attributes name and race are suppressed and replaced by '*' whereas age is generalized to the broader values. Although k-anonymity is able to provide privacy but still it is vulnerable to two types of attacks named as homogenity attack and background knowledge attack. Homogenity attack refers to release of records where sensitive records in equivalence class lack diversity whereas in background knowledge attack the attacker has background information regarding record. To counter the attacks of k-anonymity the concept of *l-diversity* is introduced by *A.Machanavajjhala[3]*, states that an equivalence class has atleast l "well-represented" values for the sensitive attribute. This is further vulnerable to attacks known as skewness attack and similarity attack. Further work is done on the privacy preservation of datasets and a new concept of *t-closeness* is introduced by *Ninghui Li[4]*. It states that an equivalence class is said to be t-closeness if the distance between the distribution of a sensitive attribute in this class and the distribution of the attribute in the whole table is no more than a threshold t. A table is said to have t-closeness if all equivalence classes have t-closeness. But the limitations with t-closensess also exists. (1) It lack the adaptability of determining different protection levels of sensitive attributes. (2) EMD provides the improper way of attribute linkage on numerical sensitive attributes. Benjamin C. Fung et al [] also explained in his paper that t-closeness is not suitable for publishing data of privacy preservation. It damages the correlation between quasi-identifiers and sensitive attributes.

B. Perturbation

Perturbation is one of the relevant privacy preserving data mining technique used for data protection. It has following types (1) Probability distribution type (2) Value distortion type

Probability distribution approach perturbs the data by replacing it from same distribution sample or from the distribution itself. Value distortion approach takes the data and replaces it by additive or multiplicative noise or by some other random processes. The latter provides better results than the former one. The latter one builds decision tree classifiers in which each element is assigned random noise from pearson distribution, for example. *Keke Chen[5]* desribes in his work the creation of random matrices and how the geometric distributions of random matrices are considered to increase the data privacy.

Zhenmin Lin[6] described in this paper the disadvantage of random rotation perturbation that it cannot hold the geometric properties of random rotation matrix in order to preserve privacy. . So here the data matrix is partitioned vertically and for each partitioned data sub-set matrix the random rotation matrix is used to perturb the original value of the data.

Xiao-Bai Li[7] describes in his paper the perturbation approach for preserving the categorical data. The proposed technique limits the disclosure of confidential data and also attempts to preserve the statistical properties of data before releasing them for data mining analysis.

C. Randomization

Randomization is the process of adding extra noise and masks the attribute value of records. The level of noise should be chosen at optimum level so that the randomized values cannot be recovered as described by *R. Agrawal[8]*. The process of randomization needs to be implemented in two steps (1) data providers randomize their data and provide randomized values to the data receivers (2) the information received should be estimated by original distribution of data using data reconstruction algorithm. *Alexandre Evfimievski [9]* describes in his paper a randomized algorithm which perturbs the original values and preserve privacy of the data. In this method, the aggregate properties of data can be recovered with some precision but the individual properties of record are being distorted, the amount of distortion required to protect privacy is also a measure being discussed in this paper.

Ling Guo [10] proposed a new framework in his paper which investigates data utility and privacy of randomization based models for categorical data. This also investigates the accuracy of association rules in data mining. A theoretical analysis has been done on how the randomization process affects the accuracy of various measures in categorical data analysis.

D. Cryptography

Cryptography can be described as a method for storing and transmitting data in a particular form so that it is revealed to those only for whom it is intended. *Murat Kantarcioglu[11]* states cryptography as an efficient technique which can preserve tough privacy and moreover can be implemented practically through various cryptographic protocols.

Benny pinkas[12] discusses the construction of generic functions that can be implemented with other functions to get the desired output of function. The generic constructions of two-party and multi-party cases are demonstrated. Further it is stated that the implementation of two-party case generic construction is easier than multi-party cases. In the end the computational drawback is stated that any change in the discussed protocol directly affects the secured computation of data.

Anand Sharma[13] describes in his paper the various types of cryptographic techniques as association rules, clustering or classification etc which are being implemented to achieve remarkable results. Special functions are being computed to store the sensitive data and providing access to stored data based on individual's role, thus the privacy is taken as main concern.

III. PARAMETERS FOR EVALUATION OF TECHNIQUES

There is some base required which can be used to find out that which technique is efficient in preserving the data. Some parameters required for evaluation are listed below:

A. Computational Time

For a technique to be efficient, the computational time should be minimum.

B. Complexity

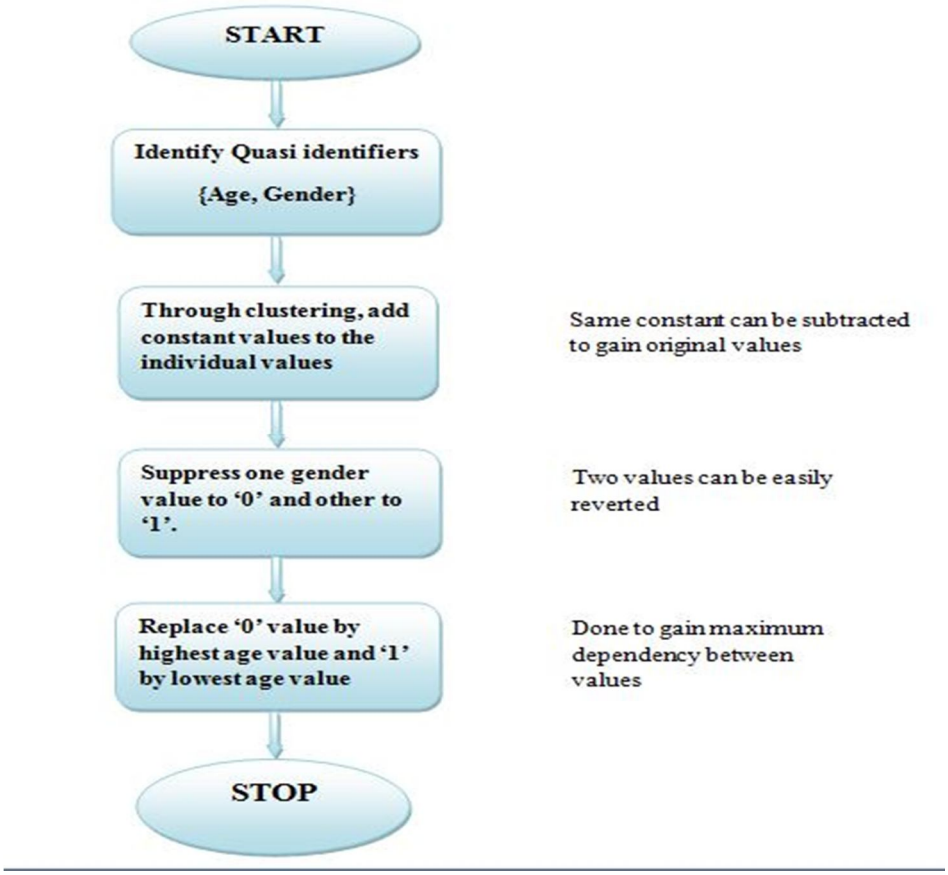
The complexity of technique should be minimum in order to be efficient.

C. Dependency Score

For technique to be efficient the dependency score between different attributes should be maximum. For example, in our proposed algorithm, there is direct linking between attribute age and gender which results in increasing dependency between them and hence results in privacy preservation of data.

D. Proposed Algorithm

All the individual approaches for privacy preserving data mining has been discussed in the previous sections. All approaches have some gaps that lacks in attaining the privacy of data published for data mining purposes. We have proposed a new technique termed as hybrid approach for data mining that preserves different quasi identifiers with different privacy preserving data mining techniques and hence results in providing better privacy results. We call this hybrid because it comprises of suppression and randomization. It works as follows:



The above diagram shows the step by step procedure of hiding the data by using hybrid approach.

IV. CONCLUSIONS AND FUTURE SCOPE

All the individual privacy preserving data mining techniques are discussed in the paper. All the techniques along with their research work and their respective drawbacks are also discussed. We can conclude that all the techniques lack in mounting the privacy of data at some point thus hybrid approach is proposed which fills the gap of preserving privacy of individual techniques and thus provides us with new framework that helps in attaining the privacy of data. All the parameters used for evaluation are achieving their respective values. The hybrid approach can further be extended to large set of quasi identifiers and multiple sensitive attributes.

REFERENCES

[1] B. C. M. Fung, K. Wang, and P. S. Yu. Anonymizing classification data for privacy preservation. IEEE Transactions on Knowledge and Data Engineering (**TKDE**), 19(5):711-725, May 2007. IEEE Computer Society.

[2] L. Sweeney, "k-anonymity: A model for protecting privacy," International Journal on uncertainty, Fuzziness and Knowledge-based Systems, vol. 10, no. 5, pp. 557–570, 2002

[3] A.Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "ℓ-diversity: Privacy beyond A.Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "ℓ-diversity: Privacy beyond k- anonymity", In Proc. 22nd Intel international Conference on data engineering. (ICDE), 2006, pp24.Y.T. Yu, M.F. Lau, "A comparison of MC/DC, MUMCUT and several other coverage criteria for logical decisions", Journal of Systems and Software, 2005, in press.

[4] Ninghui Li, Tiancheng Li, Suresh Vengakatasubramaniam, "t-Closeness: Privacy Beyond k-Anonymity and ℓ-Diversity", International Conference on Data Engineering, 2007, pp106-115.

[5] Keke Chen, Ling Liu, "Privacy Preserving Data Classification with Rotation Perturbation", Fifth IEEE International Conference on Data Mining, 2005

[6] Zhenmin Lin, Lexington, KY, Jie Wang, Lian Liu , Changjiang Zhang, "Generalized Random Rotation Perturbation for Vertically Partitioned Data Sets", IEEE Symposium on Computational Intelligence and Data Mining, 2009.

[7] Xiao-Bai Li, Sumit Sarkar, Privacy Protection in Data Mining: A Perturbation Approach for Categorical Data, Information Systems Research, 2006.

[8] Alexandre Evfimievski and Tyrone Grandison, ―Privacy Preserving Data Mining‖ at IBM Almaden Research Center, 2007.

[9] R. Agrawal, & R. Srikant, Privacy Preserving Data Mining. In Proc. of ACM SIGMOD‖, Conference on Management of Data (SIGMOD'00), Dallas, TX, 2000.

[10] Ling Guo, Randomization Based Privacy Preserving Categorical Data Analysis, Charlotte 2010

[11] Murat Kantarcioglu, Chris Clifton, "Privacy Preserving Distributed Mining of Association Rules on Horizontally Partitioned Data", IEEE Transactions on Knowledge and Data Engineering, Vol. 16, No. 9, pp. 1026 – 1037, 2004.

[12] Benny Pinkas, Cryptographic techniques for privacy-preserving data mining, HP labs.

[13] Anand Sharma and Vibha Ojha , Implementation of cryptography for privacy preserving data mining, International Journal of Database Management Systems (IJDMS) Vol.2, No.3, August 2010.