A NEW TECHNIQUE FOR PROTECTING SENSITIVE DATA AND EVALUATING CLUSTERING PERFORMANCE

S.Vijayarani[1], Dr. A.Tamilarasi[2], N.Murugesh[3]

[1]Assistant Professor, School of Computer Science and Engineering.
Bharathiar University, Coimbatore, Tamilnadu, India
vijimohan_2000@yahoo.com

[2]Professor and Head, Department of MCA, Kongu Engineering College,
Erode, Tamilnadu, India
drtamil@kongu.ac.in

[3]M.Phil Research Scholar, School of Computer Science and Engineering.
Bharathiar University, Coimbatore, Tamilnadu, India
murugeshfeb20@gmail.com

ABSTRACT

Data mining researchers and policy makers have need of raw data collected from organizations and business companies for their analysis. Any transmission of data to third parties and the organizations outsourcing their wok should satisfy the privacy requirements in order to avoid the disclosure of sensitive information. In order to maintain privacy in databases, the confidential data should be protected in the form of modifying the sensitive data items. In statistical disclosure control, masking methods are used for modifying the confidential data. Most of the perturbative masking techniques existing in the literature are general purpose ones. In this work, a new perturbative masking technique called as modified data transitive technique (MDTT) is used for protecting the sensitive numerical attribute(s). The performance of the proposed technique (MDTT) is compared with the existing masking techniques additive noise, rounding and micro aggregation. The experimental result shows the MDTT technique has produced better results than existing techniques.

KEYWORDS

Privacy, Micro Aggregation, Rounding, Additive Noise, Modified Data Transitive Technique (MDTT).

1. INTRODUCTION

Huge volumes of detailed individual information are frequently collected and analyzed by the various applications with the help of data mining. Such data include shopping habits, criminal records, medicinal history, credit records and etc., On the one hand, data is an important asset to big business organizations and governments for decision making processes and to provide social benefits, such as medical research, crime reduction, national security, etc. On the other hand, analyzing such data opens a new threat of misusing of individual's privacy.

The significant growth of the computational power makes the researchers and decision makers for analyzing the large volume of raw data. Also, companies, that generate a huge burden of data, often need to transmit these data to third parties for their study. As data usually contains sensitive information about people and corporations their release to third parties requires the application of mechanisms to make sure data privacy [4].

Statistical database security focuses on the safety of private individual values stored in so-called *statistical databases* and used for statistical purposes [5]. Statistical agencies have to take good care of the data they have composed. In the case of register-based data, it is easy to have access to all kinds of data the agency needs, but when it comes to surveys, the respondents' enthusiasm determines whether they give the information or not. If respondents can faith that their data will be used appropriately by statistical agencies and there will be no risk of disclosure, they are more willing to provide accurate information. If respondents suspect their information is at risk of disclosure, it is only natural for them to refuse to answer or to provide mistaken information [3].

Often, statistical agencies disseminate information only in the form of tables. But, micro data records which contain information about individuals or establishments offers far greater flexibility for statistical research, especially of an exploratory nature, than tables. As a result, there has been an increasing require from users for such data, and agencies would like to be able to this require, provided that confidentiality is not compromised. In particular, there is well-recognized need to avoid both identity and attribute disclosure [4].

The rest of the paper is organized as follows. Section 2 describes related works. In section3 various masking techniques are discussed. Section 4 gives the objective of the problem and the proposed solution. In section 5, existing and the proposed masking techniques are discussed. Section 6 discusses the experimental results and performance analysis. We conclude the paper in Section 7.

2. RELATED WORKS

The security of governmental databases has received substantial consideration in the literature in recent years. This can be attributed to a concurrent increase in the amount of data being stored in databases, the analysis of such data, and the desire to protect confidential data. Data perturbation methods are frequently used to defend top secret data from unauthorized queries while providing greatest access and accurate information to legitimate queries. To supply precise information, it is advantageous that perturbation does not result in a change in relationships between attributes.

Most techniques for privacy computations use some form of renovation on the data in order to perform the privacy preservation. Typically, such methods diminish the granularity of demonstration in order to lessen the privacy. This decrease in granularity results in some loss of efficiency of data management or mining algorithms. This is the usual trade-off between information loss and privacy. A host of techniques are offered for protecting numerical data from disclosure. These consist of sampling, local suppression, random noise, rounding, micro-aggregation and etc.

Ruth Brand [11] described micro data protection by adding noise. Several algorithms were developed that have different characteristics. The simplest algorithm consists of adding white noise to the data. More sophisticated methods use more or less complex transformations of the data and more complex error-matrices to improve the results. He gives an overview of different algorithms and discusses their properties in terms of analytical validity and level of protection.

Domingo Ferror et.al [5] discussed micro aggregation, a technique for statistical confidentiality that uses aggregation operators. They described the goals of statistical confidentiality and its application to continuous and categorical data. They showed the application of the method to a small publicly available data set.

Krishnamurthy Muralidhar et.al [6], described a new method (General Additive Data Perturbation) that does not change relationships between attributes. All existing methods of additive data perturbation are shown to be special cases of this method. When the database has a multivariate normal distribution, the new method provides maximum security and minimum bias. For non normal databases, the new method provides better security and bias performance than the multiplicative data perturbation method.

Muralidhar and Sarathy [17] provide a comprehensive discussion of the different techniques for protecting numerical data. With the exception of swapping and shuffling, most other data masking techniques involve the modification of the original values of the confidential variables. Many users find such modification of values to be objectionable and hence are less likely to use the modified data. By contrast, by transforming the original values leaves the original data unmodified. Hence, this type of transformation techniques are more likely to be accepted by users who find "data modification" objectionable.

3. MASKING TECHNIQUES

Protecting sensitive data is a very important issue in the government, public and private bodies. Masking techniques are used to prevent confidential information in the table. Masking techniques can operate on different data types [3]. Data types can be categorized as follows.

- *Continuous variables* (also referred to as cardinal, metric and scale variables): differences between values are meaningful so that arithmetic operations (calculating mean, standard deviation, etc.) are performed (e.g. income, age).
- *Categorical variables* (also referred to as non-metric variables): values are set of categories, and standard arithmetic operations cannot be performed. There are two types of categorical data.
    - *Nominal variables*: values are only names, and have no meaning in or of themselves (e.g. sex, address, marital status, religion)
    - *Ordinal variables*: the order of values is relevant, in that the higher the value the more (or less) of construct exists, but the deference between the values has no meaning (e.g. educational status if measured as none, primary, high school, bachelors, masters)

Masking techniques are classified into two categories

- Perturbative - the original data are modified.
- Non-Perturbative – the original data are not modified, but some data are suppressed and/or some details are removed.[3]

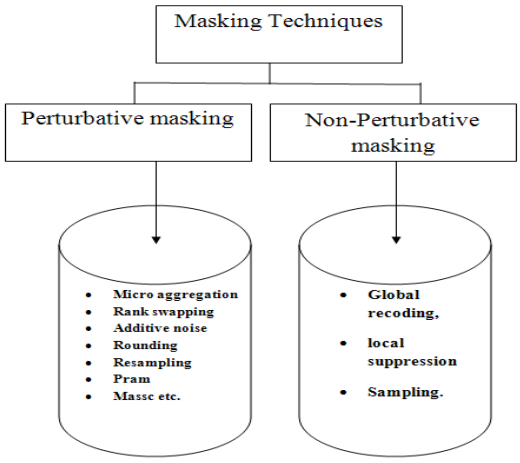

Figure1. Classification of masking techniques

3.1 Perturbative Masking Techniques

Perturbation is nothing but altering an attribute value by a new value. The data set are distorted before publication. Data is distorted in some way that affects the protected data set, i.e. it may contain some errors. In this way the original dataset may disappear and new unique combinations of data items may appear in the perturbed dataset; in perturbation method statistics computed on the perturbed dataset do not differ from the statistics obtained on the original dataset [3].

- Micro aggregation
- Rank swapping
- Additive noise
- Rounding
- Resampling
- PRAM etc.

3.2 Non Perturbative Masking Techniques

Non-Perturbative methods do not modify the values of the variables rather they produce a reduction of detail in the original data set i.e some data are suppressed or removed.

- Sampling
- Local suppression
- Global recoding
- Top-coding
- Bottom-coding
- Generalization etc.

4. OBJECTIVE OF THE PROBLEM

The main objective of this research work is protecting the confidential numeric attribute(s) in the micro data table. From this micro data, the confidential numeric attribute is selected and it can be

perturbed by the perturbative masking techniques. After perturbation, the perturbed data can be released to data mining researchers or any agency or firm for data analysis. The data mining techniques such as clustering, classification, etc are applied to this modified data, the modified table does not affect the result. In this work, a new perturbative masking technique modified data transitive technique is proposed and the performance of this technique is compared with the existing techniques additive noise, rounding and micro aggregation.

4.1 Proposed Solution

1) Identify the confidential numeric attribute from the micro data table.

2) Modification – confidential numeric attribute

- 2.1 Proposed Technique
  - 2.1.1 Modified Data Transitive Technique(MDTT)
- 2.2 Existing Technique
  - 2.2.1 Additive Noise
  - 2.2.2 Rounding
  - 2.2.3 Micro Aggregation

3) Performance Analysis

4) Finding the best technique

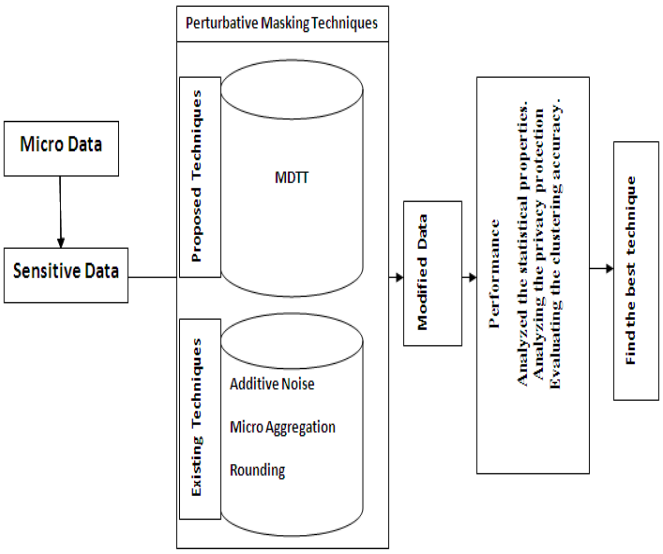

Figure 2. System Architecture

5. MASKING TECHNIQUES

5.1 Existing Techniques

5.1.1 Micro Aggregation

Micro aggregation is a Perturbative masking technique. Micro aggregation is a family of statistical disclosure control techniques for protecting continuous data.

- ***Partition:*** The set of original records is partitioned into several clusters.
- ***Aggregation:*** An aggregation operator is computed for each cluster and is used to replace the original records.

Table 1. Micro Aggregation

1. Consider a database D consists of T tuples. $D=\{T_1,T_2,\ldots T_n\}$. Each tuple in T consists of set of attributes $T = \{A_1,A_2,\ldots A_p\}$ where $A=\{A_1,A_2,\ldots A_p\}$, $A \in T$ and $T \in D$
2. Identify the sensitive or confidential numeric attribute SA_R. $\forall SA_R \in A$.
3. Consider the sensitive data item sen_data where sen_data $\in SA_R$
4. Partition
4.1 The original records are partitioned into several clusters
4.2 The records in the same cluster are similar to each other
4.3 The number of records in each cluster is atleast k
5. Aggregation
5.1 an aggregation operator, the mean for continuous data is computed for each cluster
5.2 This mean value and is used to replace the sensitive attribute of the original records.
6. End process

5.1.2 Rounding

Rounding methods replace original values of attributes with rounded values. For a given attribute Z_i, rounded values are chosen among a set of rounding points defining a rounding set [1]. In a multivariate original dataset, rounding is usually performed one attribute at a time. Rounded values are chosen from a set of rounding points ri each of which defines a rounding set.

Table 2. Rounding

1. Consider a database D consists of T tuples. D={t_1,t_2,…t_n}. Each tuple in T consists of set of attributes T={A_1,A_2,…A_p} where $A_i \in T$ and $T_i \in D$
2. Identify the sensitive or confidential numeric attribute SA_R
3. Get the base value b
4. For every sensitive data item sen_data_i in the sensitive attribute SA_R b
5. Calculate b'=(b+1)/2
6. Calculate d = (sen_data_i) mod b
7. if d=0 then "no modification "
else if d < b' then // round down
$mod\_sen\_data_i = sen\_data_i - d$
else // round up
$mod\_sen\_data_i = (sen\_data_i)+(b-d)$
}
8 : Repeat the step 5 to 7 for all the data items.

5.1.3. Additive Noise

The basic idea of the additive-noise-based perturbation technique is to add random noise to the actual data) [1]. The noise being added is typically continuous and with mean zero, which suits well continuous original data.

Table 3. Additive Noise

1. Consider a data base D with n tuples t = {t_1, t_2….t_n}.
Each tuples contains Set of attribute A = {A_1, A_2…..A_m} A € t_i.
2. Find the sensitive attribute SA_R for all $SA_R \in A_i \in A$
(i=1,2…m).
3. Calculate Average for all $SA_{R(i).}$
4. Do
Initialize the value i = (1, 2….n).
Check if (Average ≥ $SA_{R(i.}$) to count the all values C_1.
*Calculate M_1 = (2*Average)/C_1*
Replace $SA_{R(i).}$ With M_1.
Check if (Average ≤ $SA_{R(i.}$) to count the all values C_2.
*Calculate M_2 = (2*Average)/C_2*
Replace $SA_{R(i).}$ With M_2.
Increment the value of i.
5. While (i ≥ n)
6. End.

5.2 PROPOSED TECHNIQUES

5.2.1 Modified Data Transitive Technique (MDTT)

The basic idea of the new technique -based new perturbation masking technique is to modify the actual confidential data. The MDTT technique algorithm is given below.

Table 4. Modified Data Transitive

1: Create the Database D = {A_1, A_2, A_3……A_n}
2: find the sensitive attribute SA_n .for each SA_n € A_1 and A_1€ D
3: In such SA_n find out the Length (L).
4: For i = 1 to n to get the value i
If (i = 0) {
Go to the step4;
} else {
Begin
{
$T = (T_i = SA_i / L) * 2$
Replace T_i with SAi
$MT = (MTv_i = SAi – T * 2)$
Replace MTv_i with SAi
}
}
5: the modify value MT is found.
End.

6. RESULTS

The data set is taken from UCI repository (Website http://archive.ics.uci.edu/ml/datasets/Statlog+ (German + Credit + Data). It is a German credit dataset. This dataset includes 1000 instances. The number of attribute is 20 in which 7 are numerical and 13 are categorical. The German credit dataset contains bank customer details. In this research work, the proposed and existing techniques are compared and find the best masking technique the following performance factors are considered.

1. Statistical Performance
2. Privacy Protection
3. K-means clustering accuracy

6.1 Statistical Properties

6.1.1 Mean

The Figure 2 shows the statistical mean accuracy of the proposed and the existing techniques. In this performance measure, first the original data item mean value is calculated, secondly the modified data item mean value is calculated and then these two mean values are verified. The results show that the MDTT technique has given more accuracy than other techniques.

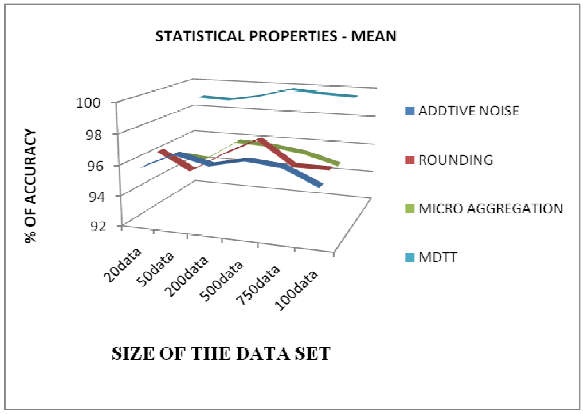

Figure 3. Statistical Calculation For Mean

6.2 Privacy Protection

Privacy Protection performance factor is used to find whether all the confidential data items of the original table are modified or not. All the existing and the proposed techniques have modified all the confidential data items 100% accurately.

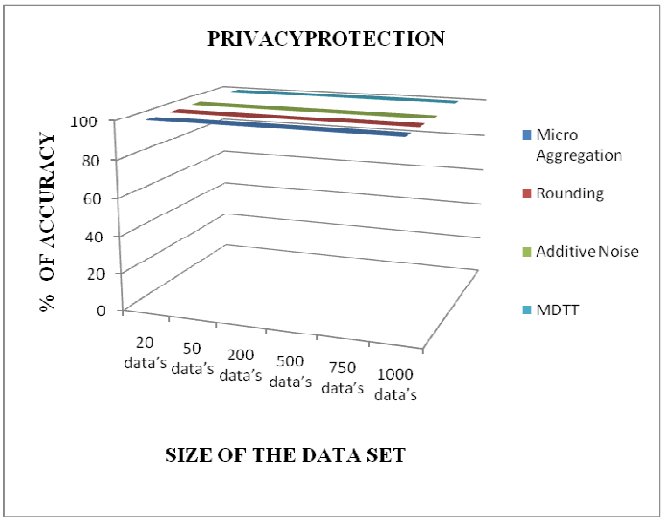

Figure 4. Accuracy of Privacy Protection

6.3 K-Means Clustering Accuracy

With the help of this performance factor, we have analyzed whether the modified data can be used for data analysis or not. In this work, we have used k-means clustering algorithm to the modified data set.

Table 5. K-Means Clustering Algorithm

```
Input
    K : the number of clusters
    D : a data set containing n objects
Output: Set of k clusters
    Method
        Arbitrarily choose k objects from D as the initial cluster centers;
    Repeat
        (re)assign each object to the cluster to which the object is the most similar, based
        on the mean value of the objects in the cluster
        Update the cluster means, i.e. calculate the mean value of the objects for each
cluster
        Until no change
        Apply the k-means clustering for both original and modified data set to get the
clusters
```

Different sizes of data sets are used for evaluating the clustering performance of the modified data. Figure 5 shows the k-means clustering accuracy of MDTT, additive noise, rounding and micro aggregation. The MDTT clustering accuracy is better than other techniques.

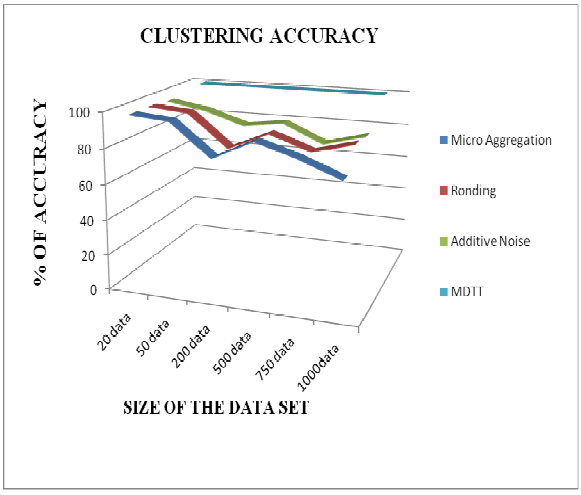

Figure 5. Clustering Accuracy for MDTT

7. CONCLUSION

Protecting sensitive raw data in the large database and the knowledge extraction is an important research problem in the field of privacy preserving data mining. In this paper, we have protected the sensitive numerical data item in the form of modifying the original data item using the proposed technique modified data transitive. This proposed technique is compared with the existing perturbative masking techniques such as additive noise, rounding and micro aggregation. The performances of the existing and the proposed techniques are analyzed. From these results

we found that the modified data transitive technique has produced better results than the existing techniques. In future, new masking techniques are to be proposed for protecting sensitive categorical attributes.

REFERENCES

[1] Stanley R. M. Oliveira Osmar R. Za¨ıane "Privacy Preserving Clustering by Data Transformation" Stanley Oliveirawas partially supported by CNPq (Consuelo National de Desenvolvimento Cient´ıfico e Technologic) of Ministry for Science.

[2] Vicenc¸ Torra, Yasunori Endo and Sadaaki Miyamoto "On The Comparison Of Some Fuzzy Clustering Methods For Privacy Preserving Data Mining: Towards The Development Of Specific Information Loss Measures" Kybernetika — Volume 45 (2009).

[3] V.Ciriani, S.De Capitan di Vimercati, S. Forest, and P. Samurai "Micro data Protection" Springer US, Advances in Information Security (2007)

[4] Charu C.Aggarwal IBM T.J. Watson Research Center, USA and Philip S. "Privacy Preserving Data mining: Models and algorithms" Yu University of Illinois at Chicago, USA.

[5] Domingo-Ferrer, J & Torra, V (2002), "*Aggregation Techniques for Statistical confidentiality*". In: Aggregation operators: new trends and applications, pp. 260-271. Physica-Verlag GmbH, Heidelberg (2002).

[6] Krishnamurty Muralidhar, Rahul Parsa, Rathindra Sarathy, "*A general Additive Data Perturbation Method for database Security*", management science, Vol. 45, No. 10, October 1999, pp. 1399-1415
DOI: 10.1287/mnsc.45.10.1399

[7] Janice Konnu "The use of protected micro data in tabulation: A case of SDC-methods, micro aggregation and PRAM "Statistics Finland, P.O. Box 5V, FI-00022 Statistics Finland, Finland.

[8] Jodi Castro "Statistical Disclosure Control In Tabular Data" Statistics and Operations Research Universidad Polit`Ecnica De Cataluña Report Dr 2009-11 November 2009.

[9] "Micro data protection" V.Ciriani, S. De Capitani di Vimercati, S.Foresti, and P.Samarati Universities degli Study di Milano, 26013 Crema, ltalia.

[10] S.Vijayarani, Dr.A.Tamilarasi, "Bit Transformation Perturbative Masking Technique for Protecting Sensitive Information in Privacy Preserving Data Mining", International Journal of Database Management Systems (IJDMS), Vol.2, No.4, November 2010.

[11] S.Vijayarani, Dr.A.Tamilarasi, "Data Transformation Technique for Protecting Private Information in Privacy Preserving Data Mining", Advanced Computing: An International Journal (ACIJ), Vol.1, No.1, November 2010.

[12] Vesalius S.Veryhios, Elisa Bettino, Igor Nai Fovino Lording Parasiliti Potenza, Yael Saygin, Yannis eodoridis, "State-of-the-art in Privacy Preserving Data Mining" , SIGMOD Record, Vol. 33, No. 1, March 2004.

[13] Brand R (2002). "Micro data protection through noise addition". In Domingo-Ferrier J, editor, Inference Control in Statistical Databases, vol. 2316 of LNCS,pp. 97{116. Springer, Berlin Heidelberg.

[14] Stanley R. M. Oliveira Osmar R. Za¨ıane "Privacy Preserving Clustering By Data Transformation" Stanley Oliveirawas partially supported by CNPq (Consuelo National de Desenvolvimento Cient´ıfico e Technologic) of Ministry for Science

[15] Janice Konnu "The use of protected micro data in tabulation: A case of SDC-methods, micro aggregation and PRAM "Statistics Finland, P.O. Box 5V, FI-00022 Statistics Finland, Finland.

[16] Manchester "Work session on statistical data confidentiality" Methodologies and Working papers 17-19 December 2007

[17] Rathindra Sarathy, Krishnamurty Muralidhar, "*The Security of Confidential Numerical Data in Databases*", information systems research,Vol. 13, No. 4, December 2002, pp. 389-403 DOI: 10.1287/isre.13.4.389.74.

Authors

Mrs. S.Vijayarani has completed MCA and M.Phil in Computer Science. She is working as Assistant Professor in the School of Computer Science and Engineering, Bharathiar University, Coimbatore. She is currently pursuing her Ph.D in the area of privacy preserving data mining. She has published papers in international journals and national/ international conferences.

Dr. A.Tamilarasi is a Professor and Head in the Departmentof MCA, Kongu Engineering College, Perundurai. She has supervised a number of Ph.D students. She has published a number of research papers in national and international journals and conference proceedings.

Mr. N.Murugesh is a M.Phil Research Scholar in School of Computer Science and Engineering, Bharathiar University, Coimbatore. He did his research work in the area of privacy preserving data mining. He has presented two papers in national conferences and published one paper in international journal.