International Journal of Database Management Systems (IJDMS), Vol.2, No.4, November 2010

BIT TRANSFORMATION PERTURBATIVE MASKING TECHNIQUE FOR PROTECTING SENSITIVE INFORMATION IN PRIVACY PRESERVING DATA MINING

S.Vijayarani[1] and Dr.A.Tamilarasi[2]

[1]School of Computer Science and Engineering, Bharathiar University, Coimbatore

vijimohan_2000@yahoo.com

[2]Department of MCA, Kongu Engg. College, Erode

drtamil@kongu.ac.in

ABSTRACT

The goal of data mining is ascertaining novel and valuable knowledge from data. In many situations, the extracted knowledge is highly confidential and it needs sanitization before giving to data mining researchers and the public in order to address privacy concerns. There have been two types of privacy in data mining. The first type of privacy is that the data is altered so that the mining result will preserve certain privacy. The second type of privacy is that the data is manipulated so that the mining result is not affected or minimally affected. The aim of privacy preserving data mining researchers is to develop data mining techniques that could be applied on data bases without violating the privacy of individuals. Many techniques for privacy preserving data mining have come up over the last decade. Some of them are statistical, cryptographic, randomization methods, k-anonymity model, l-diversity and etc. In this work, we propose a new perturbative masking technique called bit transformation technique for protecting the sensitive information. An experimental result shows that the proposed technique gives the better result compared with the existing micro-aggregation technique.

KEYWORDS

Privacy, Sensitive data, Bit transformation, Micro-aggregation, K-means clustering.

1. INTRODUCTION

Privacy preserving data mining (PPDM) refers to the area of data mining that seeks to safeguard sensitive information from unsolicited or unsanctioned disclosure. Most traditional data mining techniques analyze and model the dataset statistically, in aggregation, while privacy preservation is primarily concerned with protecting against disclosure of individual data records. The main objective in privacy preserving data mining is to develop algorithms for modifying the original data in some way, so that the private data and private knowledge remain private even after the mining process.[11]

The need for protecting numerical data from disclosure has gained considerable importance in recent years. Government agencies which release data have always been interested in this problem. However, with the increase in the ability of organizations to gather, store, analyze, disseminate, and share data, there has also been a growing demand for commercial

organizations to secure sensitive data from disclosure. Recent legislation worldwide has made this an important issue for all organizations that gather and store any sensitive information.

The advancement of information technologies has enabled various organizations (e.g., census agencies, hospitals) to collect large volumes of sensitive personal data (e.g., census data, medical records). Due to the great research value of such data, it is often released for public benefit purposes, which, however, poses a risk to individual privacy. A typical solution to this problem is to anonymize the data before releasing it to the public. In particular, the anonymization should be conducted in a careful manner, such that the published data not only prevents an adversary from inferring sensitive information, but also remains useful for data analysis.

Most methods for privacy computations use some form of transformation on the data in order to perform the privacy preservation. Typically, such methods reduce the granularity of representation in order to reduce the privacy. This reduction in granularity results in some loss of effectiveness of data management or mining algorithms. This is the natural trade-off between information loss and privacy. A host of techniques are available for protecting numerical data from disclosure. These include sampling, local suppression, random noise, rounding, micro-aggregation and etc. Muralidhar and Sarathy [9] 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.

2. STATISTICAL DISCLOSURE CONTROL

2.1. Micro Data

Micro data contain a set of attributes relating to single respondents in a sample or in a population. Microdata can be represented as tables composed of tuples (records) with values from a set of attributes. The attributes in an initial micro data table are usually classified as follows.[12]

- Identifiers**:** Attributes that uniquely identify a micro data respondent. For instance, attribute SSN uniquely identifies the person with which is associated.

- Quasi-identifiers**:** Attributes that, in combination, can be linked with external information to re-identify, all or some of the respondents to whom information refers or reduce the uncertainty over their identities. For instance, attributes DOB, ZIP, and SEX are quasi-identifiers: they can be linked to external public information to reveal the name and address of the corresponding respondents or to reduce the uncertainty to a specific set of respondents.

- Confidential attributes **-** Attributes of the micro-data table contains sensitive information. For instance, attribute disease can be considered sensitive.

2.2. Classification of Micro data Disclosure Protection

The micro data protection techniques can be classified into two main categories: masking techniques, and synthetic data generation techniques.

- Masking techniques **-** The original data are transformed to produce new data that are valid for statistical analysis and such that they preserve the confidentiality of respondents. Masking techniques can be classified as

  - Non-perturbative - the original data are not modified, but some data are suppressed and/or some details are removed; Non-perturbative techniques produce protected micro data by eliminating details from the original micro data. Some of the non-perturbative techniques are Sampling, Local suppression, Global recoding, Top-coding, Bottom-coding and Generalization.

  - Perturbative **-** the original data are modified. With perturbative techniques, the micro data table is modified for publication. Modifications can make unique combinations of values in the original table disappear as well as introduce new combinations. Resampling, Lossy compression, Rounding, PRAM, MASSC, Random noise, Swapping, Rank swapping and Micro-aggregation are some of the perturbative masking techniques.

- Synthetic data generation techniques **-** The original set of tuples in a micro data table is replaced with a new set of tuples generated in such a way to preserve the key statistical properties of the original data. The generation process is usually based on a statistical model and the key statistical properties that are not included in the model will not be necessarily respected by the synthetic data. Since the released micro data table contains synthetic data, the re-identification risks are reduced. Note that the released micro data table can be entirely synthetic (i.e., fully synthetic) or mixed with the original data (i.e., partially synthetic).[12]

3. PROPOSED SYSTEM

3.1. Objective of the Problem

The sensitive attribute can be selected from the micro data and it can be modified by a bit transformation perturbative masking technique. After modification, the modified data can be released to data mining researchers or any agency or firm. If they can apply data mining techniques such as clustering, classification, etc for data analysis, the modified table does not affect the result. In this work, we have applied k-means clustering algorithm to the modified data and verified the result. The steps involved in this work are,

A. Sensitive Attribute Selection
B. Bit transformation perturbative masking technique for modifying the sensitive attribute
C. Applying k-means algorithm for original and modified data
D. Compare the results

A. Sensitive Attribute Selection

From the micro data table select the sensitive numeric attributes. For example, an employee database the attributes are employee number, employee name, date of birth, salary, account no, qualification, designation and etc... The attributes salary and account number are considered as sensitive attributes.

B. Bit Transformation perturbative masking technique

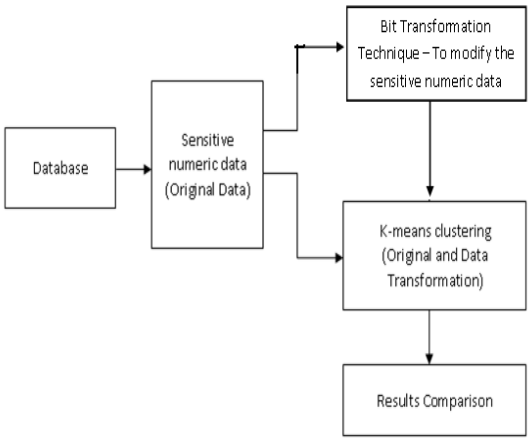

Figure 1. Proposed System Architecture

Algorithm for Bit Transformation

1. Consider a database D consists of T tuples. D={t_1,t_2,…t_n}.
2. Each tuple in T consists of set of attributes T={A_1,A_2,…A_p} where A_i Є T and T_i Є D
3. Identify the sensitive numeric attribute S_R where S_R Є T
4. Consider the sensitive attribute S_R,
5. Verify if(MSB of S_R = 1 or 2 or … 9) and (MSB-1, MSB-2,…LSB = 0)
6. Then (i) add the constant value to S_R
7. (ii) replace S_R with a new value
8. Else if (all the position of the number contains same number)
9. Then (i) add constant value to S_R
10. (ii) replace S_R with a new value
11. Else
12. Retain the MSB of S_R as it is
13. Generate new numbers by transforming the remaining bits $new_i(S_R)$ and count how many numbers can be generated and assign this value into count
14. For (i=1 to count)
15. { Find out the difference between S_R with $new_i(S_R)$ i.e.
16. $Diff_i(S_R) = (S_R) - new_i(S_R)$
17. }
18. Replace S_R with a minimum difference value
19. Repeat the steps 4 to 17 for all the sensitive data.

C. K-means Clustering Algorihm

The k-means algorithm for partitioning, where each cluster's center is represented by the mean value of the objects in the cluster
Input

- K : the number of clusters
- D : a data set containing n objects

Output: Set of k clusters

Method

(1) arbitrarily choose k objects from D as the initial cluster centers;
(2) Repeat
(3) (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
(4) Update the cluster means, i.e. calculate the mean value of the objects for each cluster
(5) Until no change

Apply the k-means clustering for both original and modified data set to get the clusters

D. Comparing the results

The data items found in the clusters are verified in both original data and modified data.

4. MICRO-AGGREGATION

Micro aggregation is a statistical disclosure control technique for microdata. Raw microdata (i. e. individual records) are grouped into small aggregates prior to publication. Each aggregate should contain at least k records to prevent disclosure of individual information. Microaggregation is a family of statistical disclosure control techniques for microdata which belong to the data modification category. The rationale behind microaggregation is that confidentiality rules in use allow publication of microdata sets if the data vectors correspond to groups of *k* or more individuals, where no individual dominates (*i. e.* contributes too much to) the group and *k* is a threshold value. Strict application of such confidentiality rules leads to replacing individual values with values computed on small aggregates (microaggregates) prior to publication. This is the basic principle of micro aggregation.[3]

To obtain micro aggregates in a microdata set with *n* data vectors, these are combined to form *g* groups of size at least *k*. For each variable, the average value over each group is computed and is used to replace each of the original averaged values. Groups are formed using a criterion of maximal similarity. Once the procedure has been completed, the resulting (modified) data vectors can be published. [1].

5. EXPERIMENTAL RESULTS

In order to conduct the experiments, synthetic employee dataset can be created with 500 records. From this dataset, we select the sensitive numeric attribute. i.e. income. Bit transformation perturbative masking technique is used to modify this sensitive attribute.

The following performance factors are considered for evaluating the technique

A. Statistical performance of the original data and modified data

In order to calculate the statistical properties such as mean, variance and standard deviation for original data and modified data. The table shows that, after the modification also the statistical properties are same as the original. Microaggregation technique returns only the mean value is same as the original. But other statistical property such as variance and standard deviation does not produce the same results. We have applied different size of data sets for verification.

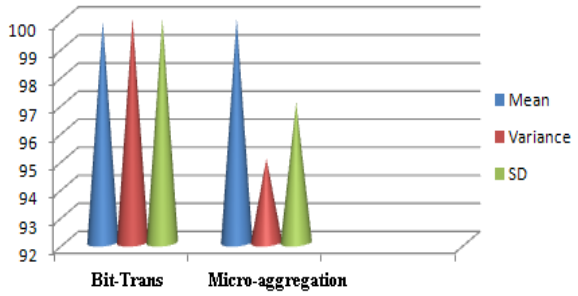

Fig.1. Statistical performances

Verifying the statistical properties the bit-transformation technique will produce accurate results compared to micro-aggregation.

B. Privacy protection

To verify the privacy protection, we test whether all the original data items are modified using bit transformation approach or not. All the data items are modified then we get 100% privacy protection. The following chart depicts this. In the given data set both methods would produce 100% of privacy protection.

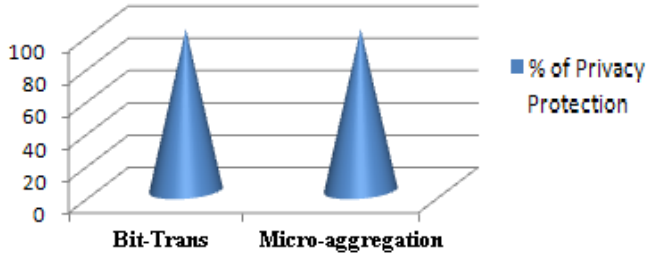

Fig.2. Privacy Protection

C. Accuracy of data mining algorithm

The following chart shows the percentage of accuracy is obtained from the bit transformation and microaggregation. The results show that the data items found in the original data items clusters are same as the bit transformation approach. Comparing bit-transformation and micro-aggregation the accuracy is higher in bit-transformation.

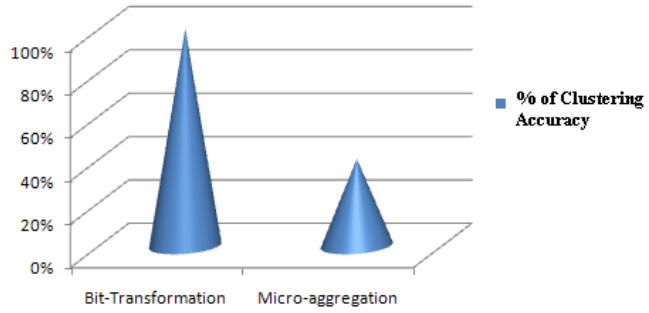

Fig.3. Accuracy of clustering algorithm

6. CONCLUSION AND FUTURE WORK

Protecting the sensitive data and also extracting knowledge is a very complicated problem. Based on the above experimental results we come know that the proposed bit transformation technique is a good technique for protecting and modifying the sensitive data. After modification, the data could be used for data mining also. And also it is very easy to get the original data. After modification we need the original data only two steps are required to get the original data. In the proposed work, the bit transformation technique is used for numerical attributes. In future, we would develop new masking techniques for protecting the categorical attributes.

ACKNOWLEDGEMENT

I would like to thank **"The UGC, New Delhi"** for providing me the necessary funds.

REFERENCES

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

[2] 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.

[3] 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).

[4] Domingo-Ferrer, J & Mateo-Sanz, J. M. (2002), "*Practical data-oriented microaggregation for statistical disclosure control*", IEEE Transactions on Knowledge and Data Engineering, vol. 14, no. 1, pp. 189-201, 2002.

[5] Domingo-Ferrer, J & Torra, V (2005), Ordinal, continuous and heterogenerous *k*-anonymity through microaggregation, *Data Mining and Knowledge Discovery*, vol. 11, no. 2, pp. 195-212, 2005.

[6] Defays, D & Nanopoulous, P (1993), "*Panels of enterprises and confidentiality: the small aggregates method*", in Proc. of 92 Symposium on Design and Analysis of Longitudinal Surveys. Ottawa: Statistics Canada, 1993, pp. 195-204.

[7] J.m. mateo-sanz, j. Domingo-ferrer, "*A comparative study of microaggregation methods"*, question, vol. 22, 3, p. 511-526, 1998.

[8] 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

[9] 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.

[10] Samarati, P (2001), "*Protecting respondents' identities in microdata release*", IEEE Transactions on Knowledge and Data Engineering, 13(6):1010-1027. 2001.

[11] Vassilios S. Veryhios, Elisa Bertino, Igor Nai Fovino Loredana Parasiliti Provenza, Yucel Saygin, Yannis eodoridis, *"State-of-the-art in Privacy Preserving Data Mining"* , SIGMOD Record, Vol. 33, No. 1, March 2004.

[12] V.Ciriani, S.De Capitani di Vimercati, S.Foresti, and P.Samarati Universitua degli Studi di Milano, "*Micro data protection*" 26013 Crema, ltalia., Springer US, Advances in Information Security (2007)

Authors

Mrs. S.Vijayarani has completed MCA and M.Phil in Computer Science. She is working as an Assistant Professor in the School of Computer Science and Engineering, Bharathiar University, Coimbatore. She is currently pursuing her Ph.D in the area of data mining. Her fields of interest are data mining and knowledge discovery, privacy and security issues in data mining. She has published two papers in international journal and presented six research papers in international and national 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.