# Data Transformation Technique for Protecting Private Information in Privacy Preserving Data Mining

S.Vijayarani[#1], Dr.A.Tamilarasi[*2]

[#1]*Assistant Professor, School of Computer Science and Engineering, Bharathiar University, Coimbatore*

[*2]*Prof&Head, Department of MCA, Kongu Engg. College, Erode*

[1]vijimohan_2000@yahoo.com, [2]drtamil@kongu.ac.in

***Abstract -*** Data mining is the process of extracting patterns from data. Data mining is seen as an increasingly important tool by modern business to transform data into an informational advantage. Data Mining can be utilized in any organization that needs to find patterns or relationships in their data. A group of techniques that find relationships that have not previously been discovered. In many situations, the extracted patterns are highly private and it should not be disclosed. In order to maintain the secrecy of data, there is in need of several techniques and algorithms for modifying the original data in order to limit the extraction of confidential patterns. 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 known as data transformation technique can be used for protecting the sensitive information. An experimental result shows that the proposed technique gives the better result compared with the existing technique.

***Keywords-*** Privacy, Sensitive data, Data transformation, Micro-aggregation, K-means clustering.

## I. INTRODUCTION

With the speedy development in database, networking, and computing technologies, a large amount of individual data can be integrated and investigated digitally, leading to an increased use of data-mining tools to infer trends and patterns. This has lifted worldwide concerns about protecting the privacy of individuals. Privacy preserving data mining (PPDM) refers to the area of data mining that seeks to safeguard sensitive information from unsolicited or unsanctioned disclosure. 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 shielding numerical data from leak has gained significant importance in recent years. Government organizations 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.

DOI : 10.5121/acij.2010.1101 

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. A host of techniques are available for protecting numerical data from disclosure. These include rounding or coarsening, perturbation, micro-aggregation, data swapping, and more recently, data shuffling. 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 data transformation techniques are more likely to be accepted by users who find "data modification" objectionable.

The rest of this paper is organized as follows. In Section 2, we present an overview of micro data and masking techniques. Section 3 we discuss about the data transformation perturbative masking technique. Micro-aggregation technique is discussed in section 4. Section 5 gives the experimental results of data transformation and micro-aggreagation. Conclusions are given in Section 6.

## II. STATISTICAL DISCLOSURE CONTROL

Inference control in statistical databases, also known as Statistical Disclosure Control (SDC) or Statistical Disclosure Limitation (SDL), seeks to protect statistical data in such a way that they can be publicly released and mined without giving away private information that can be linked to specific individuals or entities

### *A. Micro Data*

A microdata set is a set of records containing data of individuals being studied, who can be persons, companies, etc. The individual records of a microdata set are stored in a microdata file. Each individual j is assigned a data vector Vj, also called data record or data set. A data vector is formed by several variables or attributes. The attributes in an initial micro data table are usually classified as follows.[12]

1) Identifiers **-** Attributes that exclusively recognize a micro data respondent. For instance, attribute employee number uniquely identifies the employee with which is associated.
2) 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 confidential information. For instance, attribute salary can be considered as confidential.

### *B. 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

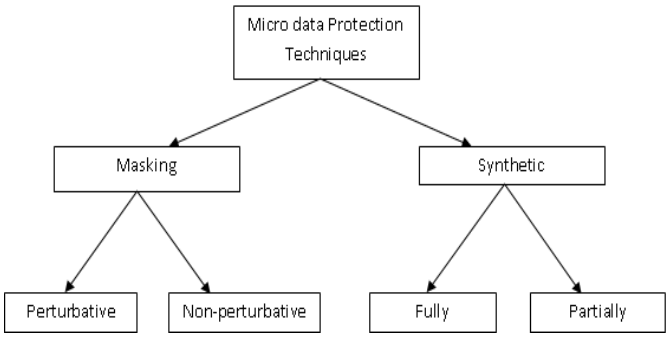


Fig.1. Classification of MPTs

1) 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:

   1.1 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.

   1.2 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.

2) 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).

# III. PROPOSED SYSTEM

## *A. Objective of the Problem*

The sensitive attribute can be selected from the micro data and it can be modified by a data 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,

1. Sensitive Attribute Selection
2. Data Transformation perturbative masking technique for modifying the sensitive attribute
3. Applying k-means algorithm for original and modified data
4. Compare the results

## *B. 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.

## *C. Data Transformation perturbative masking technique*

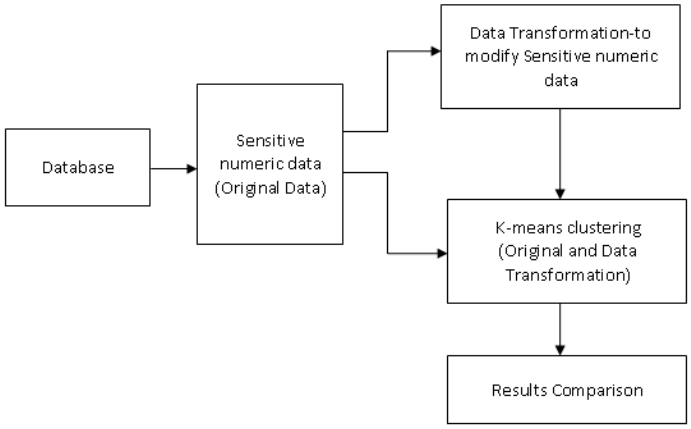


Fig.2 Proposed System Architecture

Algorithm for Data Transformation

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 $A_i \in T$ and $T_i \in D$
2. Identify the sensitive or confidential numeric attribute $A_R$

3. Consider the sensitive attribute $A_R$, find out the maximum number of digits, maxD $\in A_R$ and minimum number of digits, minD $\in A_R$
4. For grouping the same number of digits, find out the number of groups
5. Verify if (maxD=mind) then assign r as 1
   Arrange the entire sensitive data item $\in$ r group only
   Go to step 7
6. Compute r=(maxD-minD)+1
   Arrange all the sensitive data item $\in X_k$ (k=1,2,…r) groups, according to the number of digits. Check if (minD $\leq$ r $\leq$ maxD)
7. Find the number of items $K_l$(l=1,2,…m) in $\forall X_k$ group
   $\forall X_k$ group (k=1,2,…r)
   If number of items $k_l \in X_k$ =1 then no modification
8. Otherwise, if number of items $k_l \in X_k$ =2
   Then transform the value of $k_l$ to T, $k_m$ to $k_l$ ; and T to $k_m$
   else
9. Assign the value of $k_l$ to T
10. Repeat (for l=1 to m)
    begin
    Assign the value of $k_{l+1}$ to k
    end
    until (l>m)
11. Assign the value of T to $k_m$

### *D. 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

### *E. Comparing the results*

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

## IV. MICRO-AGGREGATION

Microaggregation 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 microaggregation.[3]

To obtain microaggregates 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]

## V. 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.

Data 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 chart shows that, in data transformation technique has produced the same statistical results after the modification also. 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.

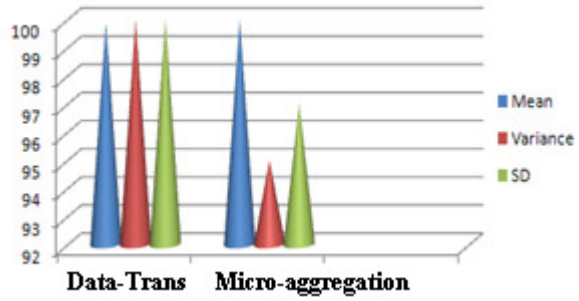


Fig.1. Statistical performances

The data-transformation technique will produce accurate statistical results compared to micro-aggregation

### *B. Privacy protection*

To verify the privacy protection, we test whether all the original data items are modified using the data 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.

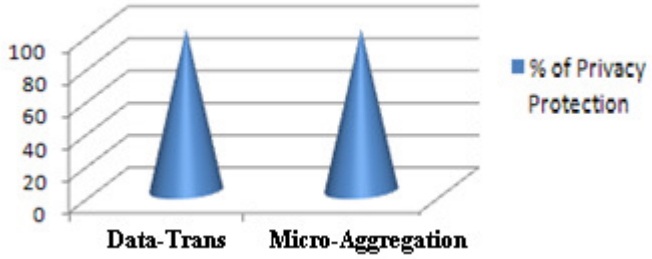


Fig.2. Privacy Protection

### *C. Accuracy of data mining algorithm*

The chart shows that the percentage of clustering accuracy is obtained from the data transformation and microaggregation. From the results, we come know that the data items found in the original clusters are same as the data transformation approach. Comparing the data-transformation and micro-aggregation the clustering accuracy is higher in data-transformation.

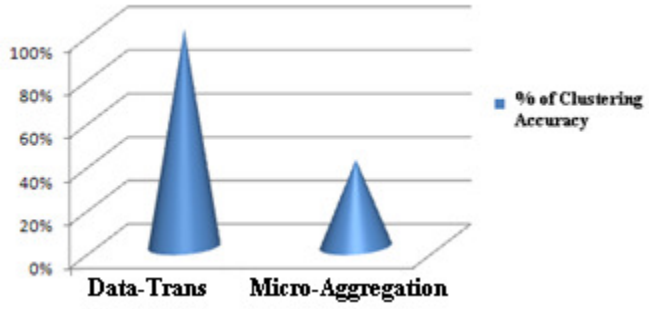


Fig.3. Accuracy of clustering algorithm

# VI. 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 data 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 data 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)