International Journal of Computational Intelligence and Informatics, Vol. 1 : No. 2, July - September 2011

Confidential Numeric Data Protection in Privacy Preserving Data Mining

S. Vijayarani
Department of Computer Science
Bharathiar University, Coimbatore-641 046, India
vijimohan_2000@yahoo.com

A. Tamilarasi
Department of Computer Applications
Kongu Engineering College Erode-638 052, India
drtamil@kongu.ac.in

Abstract

Data mining is used to extract hidden knowledge from the large data repositories. This knowledge is very essential and useful for solving the complicated problems and tackles the difficult situations in a simple way. In many circumstances, the knowledge extracted from data mining can be misused for variety of purposes. This condition raises the concerns of performing the data mining tasks in a secured manner. Privacy preserving data mining is a novel research area in the field of data mining and it mainly concentrates on the side effects which are generated during the data mining process. To perform data mining tasks, most of the times, the data may be shared among people for various reasons. To maintain data privacy, first we have to modify the original confidential data and then the modified data may be shared by others. In the literature, many protection techniques are proposed for modifying the confidential data items. In this research work, we have proposed two new techniques namely Bit++ and Bit-- for protecting the confidential numeric attribute. The performances of the proposed techniques are compared with the existing techniques additive noise and micro aggregation.

Keywords: privacy, confidential data, bit++, bit--, additive noise, micro aggregation

1. Introduction

Privacy preserving data mining is very important to perform the data mining tasks in a safe way because the knowledge extracted from data mining can be mishandled and it creates a privacy issue. To handle this problem in an effective way many techniques and algorithms have been proposed. Most of these techniques apply some form of alteration of the data in order to execute the privacy preservation. The main objective of privacy preserving data mining is to modify the confidential data by using new techniques and algorithms so that the confidential data and knowledge continued to be confidential even after the mining process also [11]. The need for preserving the confidential data from revelation has played an important role in recent years. Various business concerns, government offices, commercial organizations, statistical offices, health sectors, science and technology, etc. are always been concentrated in this problem because they release and share significant data. The speedy increases in the organization's ability to gather, observe, accumulate and distribute data, there has also been a growing constraint for organizations to give protection to the confidential data from inappropriate disclosure.

The massive development of progression in the information technology field allowed the organizations like census agencies, hospitals and other industries to collect and record the huge quantity of individual data which also holds some confidential data items like salary, health issues, etc. Moreover, these data items are also handled by the researchers for various data analysis purposes which also produce a threat to the individual's privacy. Usually, anonymization can be used for solving this type of problems. Anonymization is nothing but modifying the confidential data before makes it as public. The anonymization process can be taken in a very vigilant manner, so that the published data not only prevents a rival from getting confidential information, but also remains useful for analyzing data [3].

In order to protect the confidential data items, many techniques, methods and algorithms are used in privacy preserving data mining. Some of the important research problems in the literature of privacy preserving data mining are statistical disclosure control; k-Anonymity, query auditing, cryptographic techniques and association rule hiding. In this research work, we have proposed two techniques namely bit++ and bit-- for protecting the confidential numerical data items. Some of the existing techniques used for protecting confidential numerical data items are additive noise, rounding, perturbation and micro aggregation.

The rest of this paper is organized as follows. In Section 2, we present an overview of the related works. In Section 3 we discuss about the problem definition and the proposed solution. In Section 4 the proposed protection techniques bit++ and bit—are discussed. Section 5 gives the experimental results and the performance analysis of the proposed techniques with the existing techniques. Conclusions are given in Section 6.

ISSN : 2349 - 6363

2. Related Works

The bigger power and interconnectivity of computer systems available today give the facility of accumulating and dealing out huge quantities of data, resulting in networked information accessible from anywhere at any time. The information sharing and broadcasting procedure is obviously careful. In reality, there is a necessity to expose various data items to public. At the same time, data protection is very essential for many reasons and the confidential data should not be disclosed. Consider an example, in private business concerns and organizations using and sharing the business data (sales, marketing, products, etc.). The confidential business data should be protected, for example, customer identities, plans, future products and new marketing strategies. The historical data may be released by the government offices and agencies require a modification process to modify the information considered as confidential. Sharing and dissemination of information can be taken place effectively only if the data owner has given some guarantee that whereas releasing information, revelation of responsive information is not a threat [5].

Statistical databases contain sensitive information about individuals or companies. The objective is to provide access to statistics about groups of individuals, while restricting access to the information about any particular individual. For example, Census bureau's, are responsible for collecting information about all citizens and reporting this information in a way that it does not expose an individual's privacy. The problem is that the statistics contains vestiges of the original information. By correlating different statistics, a clever user may be able to infer confidential information about some individual. For example, by comparing the total salaries of two groups differing only by a single record, the user can infer the salary of the individual, whose record is in one group but not in the other. The intention of inference controls is to guarantee that the statistics released by the database do not lead to the discovery of confidential data.

Every database D can be viewed as a file with n records, where each record contains m attributes on an individual respondent [6]. The attributes are classified as identifiers, quasi-identifiers, confidential attributes and non-confidential attributes. Identifier attributes explicitly identify the respondent. Examples are the passport number, social security number, namesurname, etc. Quasi-identifiers or key attributes are attributes which identify the respondent with some degree of uncertainty. Examples are an address, gender, age, telephone number, etc. Confidential attributes contain sensitive information on the respondent. Examples are salary, religion, political affiliation, health condition, etc. and Non-confidential attributes are attributes which do not fall in any of the categories above.

In [2], the author uses additive noise technique for protecting confidential information in the database. Although several algorithms were developed with different characteristics, adding white noise to the data was the simplest one. More complicated methods use composite transformations of the data and more difficult error-matrices to improve the results. The paper provides an impression of different algorithms and discusses their properties.

Cox L.H et al, [4] describes how the statistical data may be rounded to integer values for statistical disclosure limitation. The principal issues in evaluating a disclosure limitation method are whether the method is effective for limiting disclosure and whether the effects of the method on data quality are acceptable or not. The authors examined the first question in terms of the posterior probability distribution of original data, given rounded data and the second by computing expected increase in total mean square error and expected difference between pre-rounding and post-rounding distributions, as measured by a conditional chi-square statistic, for four rounding methods.

Anton Flossman et al, [1], proposed to combine two separate disclosure limitation techniques, blanking and addition of independent noise, in order to protect the original data. The proposed approach gives a decrease in the probability of disclosing the individual information, and can be applied to linear as well as nonlinear regression models. The authors explained how to combine the blanking method and the measurement error method, and how to estimate the model by the combination of the Simulation-Extrapolation (SIMEX) approach and the Inverse Probability Weighting (IPW) approach.

Krish Muralidharan et al, [10], discussed about the rank based proximity swap as a data masking mechanism for continuous data. Recently, more complicated measures for masking continuous data that are based on the idea of shuffling the data have been proposed. Evaluation of the performance of the swapping and shuffling procedures is carried out where the shuffling procedures have been performed better than the swapping.

Domingo Ferrer J et al, [7, 8 and 9], discussed micro aggregation technique in statistical disclosure control technique. Raw micro data, i.e., individual records or data vectors are grouped into small aggregates prior to publication. Every aggregate should contain at least k data vectors to prevent disclosure of individual information, where k is a constant value preset by the data protector. Today, no precise polynomial algorithms are available to perform optimal micro aggregation i.e. with minimal variability loss. Various methods discussed in the literature are, ranking of data items is partitioned into groups of fixed size. In the multivariate case, ranking is performed by projecting data vectors onto a single axis. The authors have characterized candidate optimal solutions to the multivariate and univariate micro aggregation problems. In the univariate case, two heuristics based on hierarchical clustering and genetic algorithms are introduced

which are data oriented in that they try to protect natural data aggregates. In the multivariate case, fixed size and hierarchical clustering microaggregation algorithms are presented which do not require data to be projected onto a single dimension; such methods clearly reduce variability loss as compared to conventional multivariate microaggregation on projected data.

3. Proposed Algorithm

The main goal of this research work is to protect the confidential numeric data items in the database. Suppose, if the data owner wishes to share (or) outsource his data to some third party to perform the data mining tasks. But he is not interested to provide the original database as it is, because it contains the confidential data items. So he decided to modify the confidential data items by using some protection technique and given to the third party. We have to ensure that the modification process will not affect the data mining results which can be produced using the original database i.e. the modified database also provide the same result as the original database.

3.1 Proposed Solution

In this research work, first the confidential data items are selected from the original database (D). Then the confidential data items are modified by the proposed protection techniques bit ++ and bit-- and the existing techniques additive noise and micro aggregation. This modified database (D´) can be given to the organizations and data mining researchers. The data mining techniques, for example clustering, classification, and association rule are applied to D´ it should produce the same result as we get through D. In this research work, we have analyzed the k-means clustering performance of D and D´. The system architecture of the proposed work is represented in figure 1.

i. Identify the confidential data items
ii. Modification
    a. Proposed Techniques
    b. Bit++
    c. Bit—
    d. Existing Techniques
    e. Additive Noise
iii. Micro Aggregation
iv. Performance Analysis

Consider an example; an employee database is given in table 1 which consists of three categorical attributes namely employee name, qualification, and designation and one numerical attribute as income. The numerical attribute income is considered to be the confidential attribute.

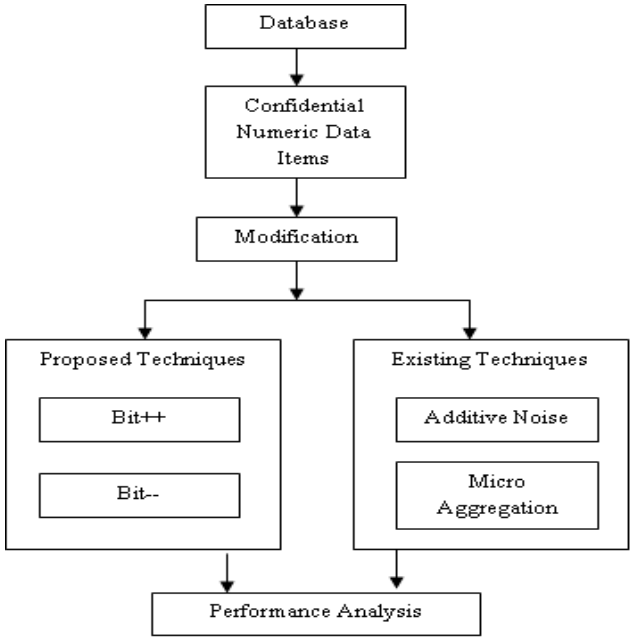

Fig. 1 System Architecture

Table 1: Employee Database

Emp. Name	Qualification	Designation	Income in Rs.
Raja	MCA	Software Engg.	65982
Priya	M.Tech	System Analyst	75675
Rama	B.E	Programmer	56030
Arun	B.Sc	Assistant	9657
Ragul	B.Com	Accountant	9954
Ramya	MCA	Software Engg.	86791
Abhi	M.Tech	Software Engg.	96786
Babu	M.E	System Analyst	54359
Shankar	B.Sc	Assistant	7650
Sita	B.Com	Accountant	8763

3.2 Additive Noise

Additive noise is one of the existing protection technique used for modifying the numerical data items. The main concept of this technique is that a noise is added to perturb the confidential data items of the original database. The data items of the confidential numeric attribute are modified by adding noise value to the original data item or subtracting noise value from the original data item. The algorithm for additive noise technique is presented below.

i. Consider a database *D* which consists of *T* tuples, where $D=\{T_1,T_2,\ldots T_n\}$.
ii. Each tuple *T* in *D* consists of set of attributes $T=\{A_1,A_2,\ldots A_p\}$, where $A_j \in T$, $T \in D$ and $j=1,2\ldots p$

iii. Identify the confidential numeric attribute CA_R

iv. Calculate the mean value($\overline{X}$) of the data items CA_R

$$\overline{X} = \frac{\sum_{i=1}^{n} d_i}{n}$$

// n represents number of data items

v. Initialize countgre=0 and countmin=0

// find the data items which are greater than or equal to mean

vi. If (d_i >= mean) then {

// form a group which contains the data items which are greater than mean value

6.1 Store d_i into group1

6.2 countgre=countgre+1 }

// find the data items which are less than mean value

vii. else {

// form a group which contains the data items which are less than mean value

7.1 Store d_i into group2

7.2 countmin=countmin+1 }

viii. Calculate the value of oise1=(2* $\overline{X}$)/countgre

ix. Calculate the value of

noise2 = (2* $\overline{X}$)/countmin

x. // subtract the noise1 value from the data items of the group 1

10.1 { for i=1 to n)

10.2 $d_i = d_i$ - noise1}

xi. // add the noise2 value to the data items of the group

11.1 {for i=1 to n)

11.2 $d_i = d_i$ + noise2}

xii. Verify the mean value of *D'* which is same as *D*

xiii Verify the sum of noise1 and noise2 which is 0.

xiv. Repeat the same process for all the sensitive attributes

xv. Get the new modified data set *D'*

xvi. Stop

3.3 Microaggregation

Microaggregation technique is also a protection technique used for modifying the data items of the confidential attribute of the original database. In this technique, the first step is to group the confidential data items into several clusters according to some conditions. In this work, Euclidean distance measure is used for finding the similarity between the data items and then the data items are clustered. The next step is to calculate the average value for the sensitive data items which are found in every cluster. In the third step, the modification is made by replacing the confidential data items found in each cluster with its average value of the cluster. This value is considered as modified value of confidential data items. The same process is repeated for all the clusters.

i. Consider a database *D* which consists of *T* tuples, where $D=\{T_1,T_2,\ldots T_n\}$.

ii. Each tuple *T* in *D* consists of set of attributes $T=\{A_1,A_2,\ldots A_p\}$, where $A_j \in T$, $T \in D$ and $j=1,2..p$

iii. Identify the sensitive numeric attribute(s) SA_R

iv. // Partition the sensitive data items using Euclidean distance

a. Apply the clustering condition to partition the sensitive data items d_i into C_j clusters where $i = 1,2\ldots n$ and $j = 1,2,\ldots k$

b. The data items in the cluster are similar to each other.

c. Find out the number of data items n in every cluster C_j

v. // Mean calculation for all the clusters

5.1 Calculate the $\overline{X}$ value of C_j $\overline{X} = \frac{\sum_{i=1}^{n} d_i}{n}$

vi. // Modification

6.1 Mean value $\overline{X}$ for each cluster is used to replace the values of the data items of the sensitive attribute for every cluster.

6.2 Repeat the same process for all the clusters.

vii. Repeat the same process for all the sensitive attributes

viii. Get the modified data set *D'*

x. End process

3.4 Bit++ Technique

Bit++ technique is a new technique proposed for modifying the confidential numeric data items of the micro data. The first step of this technique is to retain the *MSB* of the sensitive data item. This technique verifies whether the bit positions *MSB-1*, *MSB-2*,…*LSB* of data item is equal to *9* and modifies that bit position to *0*. The remaining bits of the data item are modified by incrementing one with that digit. The same process is repeated for all the data items.

i. Consider a database *D* which consists of *T* tuples. $D=\{T_1,T_2,\ldots T_n\}$

ii. Each tuple *T* in *D* consists of set of attributes $T=\{A_1,A_2,\ldots A_p\}$ where $A_j \in T$, $T \in D$ and $j=1,2..p$

iii. Identify the sensitive numeric attribute(s) SA_R

iv. For every d_i, calculate the length and assign it as *L*.

v. In the sensitive data item d_i retain the value of MSB as it is

vi. For (k=1 to L-1)

vii. Check (if the value of (MSB-k) = 9) then

a replace the value of (MSB-k) as 0

viii. else replace the value of (MSB-k) is incremented

by 1

ix. Get the new sensitive data for the current sensitive data

x. Repeat the steps 3 to 8 for all the data items

xi. Get the modified data set *D'*

xii. End process

3.5 Bit - - Technique

Another protection technique used for modifying the sensitive data items is Bit-- technique. This technique is same as Bit++ technique, but with one difference. In Bit++ technique, the digits are incremented by one but in Bit-- technique the digits are decremented by one. In Bit-- technique the *MSB* of the sensitive data item is not modified and it is retained as it is but the remaining bits of the sensitive data item are modified by decrementing one from the digit. Before performing modification, this technique verifies whether the *MSB-1, MSB-2,…, LSB* is equal to *0*. If any of these bit position value is *0* then this value is modified as *9* and other bit position values are decremented by one. The same process is repeated for all the data items.

i. Consider a database *D* which consists of *T* tuples. $D=\{T_1,T_2,\ldots T_n\}$.

ii. Each tuple *T* in *D* consists of set of attributes $T=\{A_1,A_2,\ldots A_p\}$ where $A_j \in T$, $T \in D$ and $j=1,2..p$

iii. Identify the sensitive numeric attribute(s) SA_R

iv. For every d_i, calculate the length and assign it as *L*.

v. In the sensitive data, retain the value of MSB as it is

vi. For (k=1 to L-1)

vii. Check (if the value of (MSB-k) = 0) then replace the value of (MSB-k) as 9

viii. else replace the value of (MSB-k) is decremented by 1

ix. Get the new sensitive data for the current sensitive data item

x. Repeat the steps 3 to 8 for all the data items

xi. Get the modified data set *D'*

xii. End process

The following table shows how the confidential data item income is modified by applying the proposed and the existing protection techniques.

Table 2. D and D'

Original Database (D)	Modified Database (D´)			
	Proposed Techniques		Existing Techniques	
	Bit++	Bit--	AN	MA
65982	66093	64871	50260	58790
75675	76786	74564	59953	86417
56030	57141	55929	40308	58790
9657	9768	9546	33240	36024
9954	9065	9843	33537	36024
86791	87802	85680	71069	86417
96786	97897	95675	81064	86417
54359	55460	53248	38637	58790
7650	7761	7549	31233	36024

4. Performance Analysis

The performance factors used for measuring the efficiency of the existing and the proposed protection techniques are statistical accuracy, privacy protection and clustering accuracy. The proposed and the existing protection techniques are implemented using Visual Basic and MATLAB. The synthetic employee-income dataset is created with 20K records. This data set consists of 15 attributes of which 5 are numerical and 10 are categorical. From these, three numerical attributes namely salary, income tax and weeks worked in a year are considered as confidential attributes. Different sizes of data set with 3K, 5K, 10K, 15K and 20K instances are used for implementing the proposed and the existing protection techniques.

The experiment was conducted in Intel Core I3 processor with a CPU clock rate of 2.4 GHz, 500 GB Hard Disk and 4 GB RAM running windows operating system. The average percentages of these three confidential attributes are given in the following performance factors.

To find the statistical accuracy, the mean value of the data items of the confidential attribute(s) is considered. First, the mean value is calculated for original database D and then the mean value is calculated for the modified database D′ i.e. after performing modification using the existing and the proposed protection techniques. The mean values of D and D′ are compared.

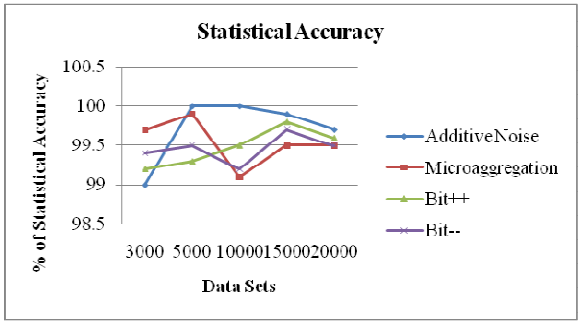

Fig. 2. Statistical Accuracy

Figure 2 shows the statistical mean accuracy of the existing and the proposed protection techniques. By analyzing the results, it is found that for all the data sets, the additive noise technique has produced better results than the other techniques.

The privacy protection accuracy ensures that all the data items in the confidential attributes of the original database (D) are modified by the existing and the proposed masking techniques. This performance factor is mainly used for finding out whether the protection techniques have properly modified the confidential data items or not. The original confidential data items are compared with the modified data items to verify whether both have the same value. If both the data

items have different values, then privacy protection is good. Based on the results represented in figure 3, it is found that the bit++ technique privacy protection ensures more accuracy than the other techniques.

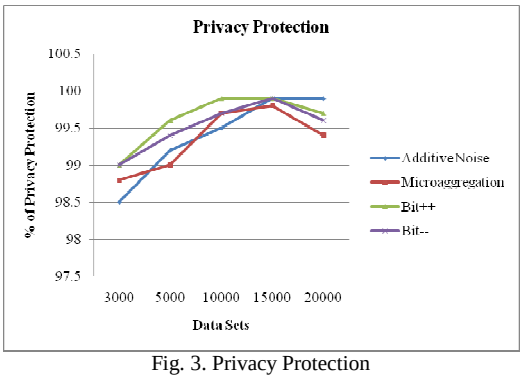

Fig. 3. Privacy Protection

The popular clustering algorithm, K-means clustering algorithm, is used for measuring the clustering accuracy performance. The k-means algorithm is applied to both the original database (D) as well as the modified database (D′). The data items in the clusters of D and D′ are verified. If both D and D′ have the same number of data items in every cluster, then the clustering accuracy is very high.

K-means Clustering Algorithm

The *K*-means algorithm is used for partitioning, where each cluster's center is represented by the mean value of the objects in the cluster [10]

Input

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

Output: Set of k clusters

Method

I. arbitrarily choose k objects from D as the initial cluster centers;
Repeat
II. (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.
III. Update the cluster means, i.e. calculate the mean value of the objects for each cluster Until no change.

Figure 4 depicts the K-means clustering accuracy for the existing and the proposed protection techniques. Different clusters and data sets are used for this analysis. From the results, it is observed that the Bit++ clustering accuracy is better than other techniques.

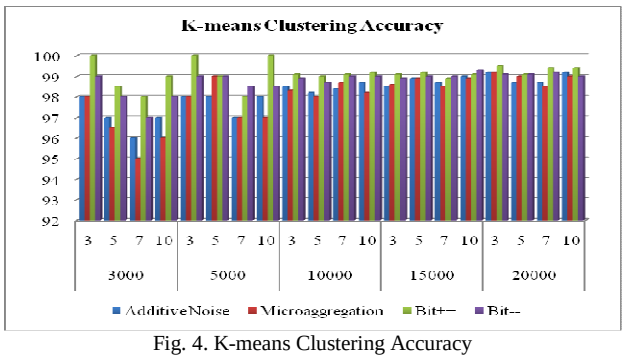

Fig. 4. K-means Clustering Accuracy

5. Conclusion

Confidential data protection and knowledge extraction are very complicated tasks in the data mining domain. This research work has discussed about confidential numeric data protection. Two new protection techniques bit++ and bit-- are proposed. The performances of the proposed techniques are compared with the existing techniques additive noise and micro aggregation. By analyzing the experimental results, we come to know that the proposed bit++ protection technique has produced better results compared with the other techniques. In future, we would develop new protection techniques for protecting the categorical attributes.

References

[1] Agrawal. R., Imielimski. T and Swami. A.N., "Mining Association Rules between sets of items in large database", Proceedings of ACM SIGMOD International Conference Management of Data, ACM Press, pp. 207-216, 1993.

[2] Anton Flossmann, Sandra Lechner, "Combining Blanking and Noise Addition as a Data Disclosure Limitation Method", LNCS 4302, Springer-Verlag Berlin Heidelberg 2006, pp. 152-163.

[3] Brand R, "Micro data protection through noise addition", Inference Control in Statistical Databases, Vol. 2316 of LNCS, pp. 97-116. Springer, Berlin Heidelberg, 2002.

[4] Charu C Aggarwal, Philip S, " Privacy Preserving Data Mining : Models and Algorithms" Yu University of Illinois at Chicago, USA, 2008.

[5] Cox L H and Kim J J, [2006], "Effects of Rounding on the Quality and Confidentiality of Statistical Data".Privacy in Statistical Databases-PSD 2006, Volume 4302 of Lecture Notes in Computer Science, pages 48–56, Berlin Heidelberg.

[6] Ciriani, S.De Capitani di Vimercati, S.Foresti S.Foresti, and P.Samarati Universitua degli Studi di Milano, "Microdata Protection" 26013 Crema, ltalia, Springer US, Advances in Information Security (2007).

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

[8] Domingo-Ferrer J & Mateo-Sanz J.M, "Practical Data Oriented Micro Aggregation for Statistical Disclosure Control", IEEE Transactions on Knowledge and Data Engineering, Vol. 14, no. 1, pp. 189-201, (2002b).

[9] Domingo-Ferrer & Torra V, "Ordinal, Continuous and heterogenerous k-anonymity through microaggregation", Data Mining and Knowledge Discovery, vol. 11, no. 2, pp. 195-212, 2005.

[10] Krish Muralidhar, Rathindra Sarathy, Ramesh Dandekar, [2006], "Why Swap When You Can Shuffle? A Comparison of the Proximity Swap and Data Shuffle for Numeric Data", PSD 2006, LNCS 4302, pp. 164 – 176, Springer-Verlag Berlin Heidelberg.

[11] Vassilios S. Verykios, 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.

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. Her fields of research interest are data mining, privacy and security issues in data mining, Data streams and Medical data mining.

Dr. A. Tamilarasi is a Professor and Head in the Department of Computer Applications, Kongu Engineering College, Perundurai.