Implementation of Multi-Level Trust in Privacy Preserving Data Mining against Non-Linear Attack

Vaishali Bhorde
Computer Engineering Department, Savitribai Phule
Pune University
JSPM's Imperial College Engineering & Research, Wagholi, Pune

R.N.Phursule
Computer Engineering Department, Savitribai Phule
Pune University
JSPM's Imperial College Engineering & Research, Wagholi, Pune

ABSTRACT

The study of perturbation based Privacy Preserving Data Mining (PPDM) [1] [2] approaches introduces random perturbation that is number of changes made in the original data. The limitation of existing work is single level trust on data miners but proposed work is focus on perturbation based PPDM to multilevel trust. [1] When data owner sends number of perturbated copy to the trusted third party, adversary cannot find the original copy from the perturbated copy means the adversary diverse from original copy this is known as the diversity attack. To prevent diversity attack is main goal of Multilevel Trust in Privacy Preserving Data Mining (MLT-PPDM) services. [1]The different MLT-PPDM algorithms are used to produce noise into original data. In existing system by applying nonlinear collusion attack on MLT-PPDM approach, it is possible to reconstruct original data. In proposed system by applying masking noise linear transformation algorithm which produce noise into original data. When same nonlinear collusion attack is applied on proposed approach it cannot reconstruct original data means it preserve the privacy. That means existing system is limited only for linear attack. [1] But proposed system is working on the non-linear attack also. Linear attack is calculating average between all perturbated copies. Nonlinear attack is calculating minimum, maximum, median function estimation.

Keywords

Diversity Attack, Multi-Level Trust, Non-Linear error estimation, Parallel Generation. Sequence Generation, On Demand Generation, LLSEE.

1. INTRODUCTION

Currently privacy preservation issue produced in several organizations which depend on data mining knowledge. To knowledge valuable data without leakage of sensitive data. This is main approach behind privacy preserving; in additional approach we can say use non-confidential data to deduce confidential data. There are several exploration and branches in this domain. Most of them examine and improve the tools and procedures of privacy preserving data mining. Privacy Preserving Data Mining (PPDM) consist of two types of privacy first is Single level trust privacy preserving data mining (SLT-PPDM) and second is multilevel trust privacy preserving data mining (MLT-PPDM). Data owner produce only one perturbated copy of its data with uncertainty about different values before data is out to trusted thirty party. In existing system only one perturbated copy is send to the trusted third party; hence it is called as single level trust. But in proposed system multiple copies send to the reliable third party; hence it is called as multilevel trust. A data owner having various changeable copies of the similar data. Data owner have the authority to send which copy to which user. When data owner sends various perturbated copies to trusted third party for preservation purpose. That stage opponent cannot discover original copy from the huge number of changeable copies, means the opponent distract from original copy. In previous technique combine all multiple perturbated copies. When nonlinear attack i.e. collusion attack is applied on MLTPPDM method. It is possible to recreate original data, which concept not acceptable through the data owner. In current method used masking noise linear transformation to generate noise into original record. When applying same nonlinear attack i.e. collusion attack on algorithm. It is not possible to recreate original data, hence it preserve the privacy this is core objective of proposed system. MLT-PPDM preserves the privacy for sensitive data against linear attack. It is a necessity to expand the system for nonlinear attack. The existing system is limited only for linear attack. But proposed system is working on the non-linear attack. Linear attack is calculating average between all perturbated copies. Nonlinear attack is calculating minimum, maximum, median function estimation. Finally compare proposed system with existing system in terms reliability and performance.

2. RELATED WORK

In several organizations used set of records are collected for various mean for their own purpose. The confidential records can breach through third person and it cannot access by publically so confidentiality is core an approach. Data Perturbation is a widespread procedure in PPDM. The PPDM constructed perturbation presents random perturbation to distinct values. To preserve confidentiality previously data is distributed among many organizations. The several organizations consist of huge of data are collected for various mean for their own purpose. Data perturbation contains of dual categories first one is probability distribution method and second is value distortion method. The probability distribution method swaps the data with alternative data from the similar distribution or itself also. The value distortion method consist modification of feature by addition nearly additive as well as multiplicative noise before records are out. To ignore the attack many anonymization methods are used. In generality and bucketization there are no strongly split up among sensitive as well as quasi identifier features. In previous approach used slicing method in that data can be partition into vertically as well as horizontally. .Data partition similarly involves three categories.

1) Those attributes are identifiers but that should be uniquely identified for example name as well as social safety number.

2) Those attribute are already recognize by the opponent that attribute are called as quasi identifiers (QI) for example age, gender and zip code.

3) Those attribute are unidentified by the opponent that attribute are called as sensitive identifiers (SA) for example disease, salary.

The methods of the initial type provide preceding method k-Anonymization techniques there is one possibility consist damage of information. The core knowledge is to suppress or generalize approximate public data so that every record turns out to be indistinguishable since at minimum k -1 other records, when projected on the subset of public attributes. Therefore, the confidential records might be connected to groups of records of range at minimum k. To avoid the linking attack is main approach of k anonymization. One approach to prevent such linking attack is masking the sensitive information of these attributes as following:

1) If there is a hierarchical description for a definite attribute (for example, Birthplace), it can generalize a specific value description into a less specific but semantically consistent information.

2) If there is no hierarchical description for a definite attribute, it can suppress a value description to a “null value” denoted ⊥

3) If the attribute is a continuous attribute it can discretize the range of the attribute into a small number of intervals.

- The second methods consist of secure Mutual calculation (SMC) provides robust level of confidentiality. It distributes protected records without revealing inside data of specific individual. The SMC procedure is very luxurious in practice, and unfeasible for actual practice. To ignore the high computation cost it use the another solution. It constructs a decision tree by considering horizontal split up data as well as vertically split up data procedure for association rule and frequent types of mining problems. These two rules calculate how much data is repeated into particular tuple.

- The method of third type again it divide into number of category like k-anonymity, retention replacement, data perturbation approach again data perturbation consist of two types first probability dissemination methodology and second value alteration methodology. Preceding result is limited singly for linear attack. In previous scheme anonymization procedures are used for column generality. In previous scheme losses data easily. Generality as well as bucketization method does not consist strongly split up between sensitive as well as quasi identifier features. Existing system cannot handle large amount of data. Question and answering procedure is used sliced data for analysis purpose. The slicing method is used to increase the performance of present state of data. Slicing algorithm perform partition of data by column wise, after partition applies generality on column. Then split up tuples into number of buckets. Those attributes are strongly correlated are consider into same column. Existing results show that slicing preserves much better data usefulness than generality. Slicing gives better performance than generality.

3. PROPOSED WORK

A data owner having original data by adding noise into that by guassion noise then partition of data with four attribute and two attribute. A data owner having authority to send different perturbated copies to different user. The hiding takes place only the integer value for example age and salary, this sensitive data cannot reconstruct by nonlinear attack. It calculate covariance matrix by using number of users and value of pertubated copy by adding guassion noise. Covariance matrix used to calculate how much two random variables change together, random is nothing but the set of different possible values. MLT-PPDM algorithms consist of parallel generation, sequential generation & on demand generation. This algorithms usage to add noise into original data.

Parallel Generation [1]

```
1. Input: X, $K_X$ and  $б^2_{Z1}$, $б^2_{ZM}$
2. Output: Y
3. Construct $K_Z$ with $K_X$ and $б^2_{Z1}$ $б^2_{ZM,}$ according to (1)
4. Generate Z with $K_{Z,}$ according to (2)
5. Generate Y = X + Z.
6. Output Y
```

$$K_{\mathbb{Z}} = \begin{bmatrix} \sigma^2_{Z_1} K_X & \sigma^2_{Z_1} K_X & \cdots & \sigma^2_{Z_1} K_X \\ \sigma^2_{Z_1} K_X & \sigma^2_{Z_2} K_X & \cdots & \sigma^2_{Z_2} K_X \\ \vdots & \vdots & \ddots & \vdots \\ \sigma^2_{Z_1} K_X & \sigma^2_{Z_2} K_X & \cdots & \sigma^2_{Z_M} K_X \end{bmatrix}. \quad (1)$$

$$Fz(v)= \frac{1/}{\sqrt{(2\pi)^M \det(K_z)}}\; e^{-1/2\; v^T K_z^{-1}\, v} \quad (2)$$

Some notations are using in generation of perturbated in various algorithms.

Table 1: Key Notation

NOTATION	DEFINITION
X	ORIGINAL DATA
Y_i	PERTURBED COPY OF X OF TRUST LEVEL i
Z_i	NOISE ADDED TO X TO GENERATE Y_i
M	NUMBER OF TRUST LEVELS
N	NUMBER OF ATTRIBUTES IN X
Y	A VECTOR OF ALL M PERTURBED COPY
K_X	COVARIANCE MATRIX OF X
K_Z	COVARIANCE MATRIX OF Z

Sequential Generation [1]

```
1. Input: X, $K_X$ and  $б^2_{Z1}$  $б^2_{ZM}$
2. Output: $Y_1$ to $Y_M$
3. Construct $Z_1$ ~ N (0, $б^2_{Z1,}$ $K_X$)
4. Generate $Y_1$=X+$Z_1$
5. Output $Y_1$
6. for i from 2 to M do
7. Construct noise £ ~ N (0, ($б^2_{Zi}$ - $б^2_{Zi-1}$) $K_X$)
```

8. Output $Y_i = Y_{i-1} + £$

9. Output Y_i

10. end for

On Demand Generation [1]

1. Input: X, K_X and б^2_{Z1} б^2_{ZM} and values of Z´: v_1

2. Output: New copies Z´´

3. Construct K_Z with K_X and б^2_{Z1} $\text{б}^2_{ZM,}$ according to (15)

4. Construct $K_{Z',}$ $K_{Z''Z'}$ and $K_{Z'}$ from K_Z

5. Generate Z' as a Gaussian with mean and variance in (3) and (18), respectively

6. for i from L + 1 to M do

7. Generate $Y_{i\,=}$ X + Z_i.

8. Output Y_i

9. End for Gaussian with mean

$$K_{Z''Z''}K^{-1}_{Z'}v_1 \quad (3)$$

And covariance

$$K_{Z''\text{-}}K_{Z''\,Z'}K^{-1}_{Z'}\,K^{-T}_{Z'\,Z'} \quad (4)$$

4. COLLUSION ATTACK

Collusion attack means several colluders get together, combine information from different perturbated copies of the same data and generate a new copy where the original sensitive data are removed.

A. Collusion model.

Collusion attack is classified into linear collusion and nonlinear collusion according to the performance of collusion model.

B. Linear collusion attacks

Linear collusion is given several differently pertubated copies of the same content; the colluders linearly combine all the copies to generate a colluded copy.

C. Nonlinear collusion attacks

Nonlinear collusion attacks is based upon such operations as taking the maximum, minimum, and median of corresponding components of the colluders pertubated copies. The colluders can output any value between the minimum and maximum corresponding values. A representative collusion for linear collusion is average attack. There is some typical nonlinear collusion named minimum attack, maximum attack, median attack, min-max attack.When nonlinear collusion attacks is applied on MLTPPDM, it reconstruct original data, it didn’t preserve the privacy but in proposed system by appling masking noise linear transformation it produced pertubated copy and apply nonlinear attack on same. It cannot reconstruct original means it preserve privacy this is main goal of proposed system. Assume that k colluders Sc containing set containing the directories of the colluders. The collusion attack is With *K* different copies $\{\mathbf{X}^k\}_{k\in S_C}$ the colluders generate the *j*th (*j* =1……*N*) component of the attacked copy V = [V_1 *V*2 ..... V_N]^T using one of the following collusion functions:

Average: $V_j^{\,avg} = \sum X^k_j/K,$

$k\epsilon S_c$

Minimum: $V_j^{\,min} = \min\{X^K_j\},$

$k\epsilon S_c$

Maximum: $V_j^{\,max} = \max\{X^K_j\},$

$k\epsilon S_c$

Median: $V_j^{\,med} = \text{median } k\epsilon S_c \{X^K_j\},$

MinMax: $V_j^{\,minmax} = (V^{min}_{j\,+}V^{max}_{j})/2$

D. Steps to do attack

1. If data owner have 2 copies of data, then it add noise as z_1, z_2 respectively,
2. So perturbated copies means final data is
   a. Y_1=X+Z_1;
   b. Y_2=X+Z_2;
3. Then send it to user

We make the same assumption that adversaries have the knowledge of the statistics of the original data X and the noise Z, i.e., mean μ_X, and covariance matrices K_X and K_Z.

4. Then user retrieved it by using Linear least square error estimation
5. This equation gives us error rate
6. If estimation error is minimum then it hard to get original data.
7. If error rate is maximum then it is added to each copy of the pertubated data from n series of data
8. Then calculate mean of each copy.
9. Then mean is actually the original data.

E. Masking noise for linear transformation

To apply empirical distribution function on original data for producing noise,smoothing means to do more secure data because no gurantee to recover data hence it convert into random variable.Whatever value gnerated after conversion that noise add into original data.

1. First calculate empirical distribution function for every variable;

2. Smoothing the empirical distribution function;

3. converting the smoothed empirical distribution function into a uniform random variable and this into a standard normal random variable;

4. Adding noise to the standard normal variable;

5. Back-transforming to values of the distribution function;

6. Back-transforming to the original scale.

5. SYSTEM IMPLEMENTATION

Fig.1 illustrate architecture of proposed system.The entire system is based upon comparative analysis between exisiting and proposed system.The propoesd syetm used real dataset called as census use only two sensitive attributes which preserve privacy.The attributes are partition into two and four attributes,because data owner having authority to send different pertubated copy to different user.By perfroming different MLTPPDM algorithm added one more noise into original data for privacy purpose.When nonlinear collusion attack is applied on MLTPPDM it is possible to recnstruct original data,so exisiting system can not preserve privacy.In proposed system used masking noise linear transformation to produced noise into data. When applying nonlinear attack it is not possible to reconstruct original data, hence proposed system preserve privacy. The result shows comparative analysis between existing and proposed system.

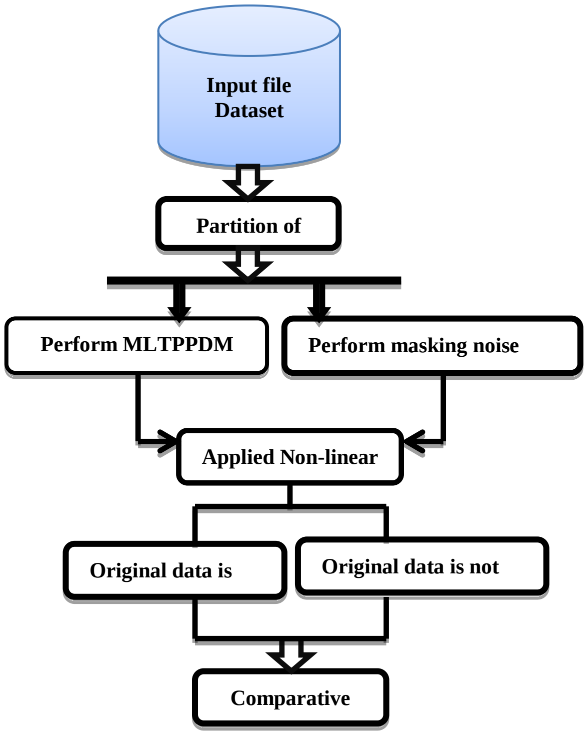
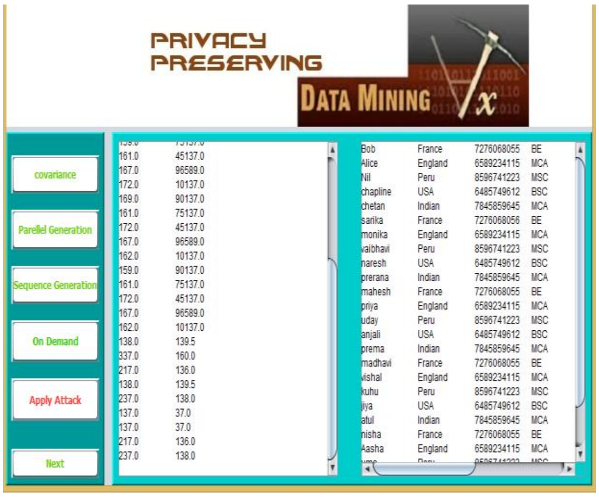


Fig.1 System Architecture

6. RESULT AND DATASET

It uses real dataset CENSUS which is commonly used in the preservation such as, for carrying out the experiments and evaluating their performance. This dataset contains one million tuples with four attributes: Age, Education, Occupation, and Income It takes the first 105 tuples and conducts the experiments on the Age and salary attributes. In proposed approach α is unique, until α did not detect nor adversary can access original copy from perturbed copies.

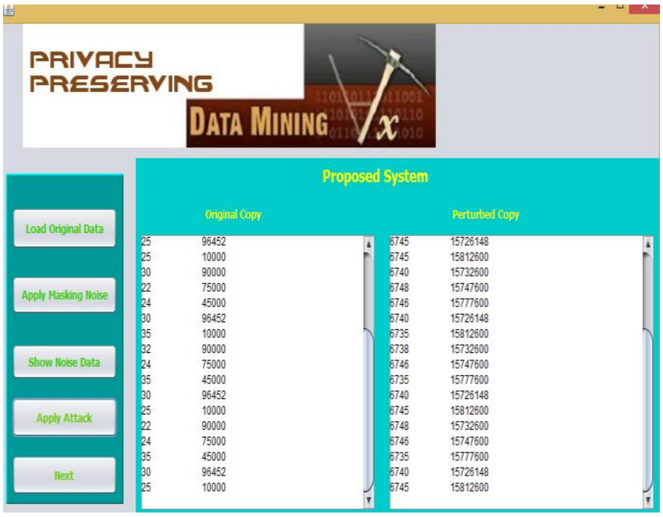


```
161.0     45137.0
167.0     96589.0
172.0     10137.0
169.0     90137.0
161.0     75137.0
172.0     45137.0
167.0     96589.0
162.0     10137.0
159.0     90137.0
161.0     75137.0
172.0     45137.0
167.0     96589.0
162.0     10137.0
138.0     139.5
337.0     160.0
217.0     136.0
138.0     139.5
237.0     138.0
137.0     37.0
137.0     37.0
217.0     136.0
237.0     138.0
```

```
Bob        France     7276068055   BE
Alice      England    6589234115   MCA
Nil        Peru       8596741223   MSC
chapline   USA        6485749612   BSC
chetan     Indian     7845859645   MCA
sarika     France     7276068056   BE
monika     England    6589234115   MCA
vaibhavi   Peru       8596741223   MSC
naresh     USA        6485749612   BSC
prerana    Indian     7845859645   MCA
mahesh     France     7276068055   BE
priya      England    6589234115   MCA
uday       Peru       8596741223   MSC
anjali     USA        6485749612   BSC
prema      Indian     7845859645   MCA
madhavi    France     7276068055   BE
vishal     England    6589234115   MCA
kuhu       Peru       8596741223   MSC
jiya       USA        6485749612   BSC
atul       Indian     7845859645   MCA
nisha      France     7276068055   BE
Aasha      England    6589234115   MCA
```

Fig.2 Applying nonlinear attack reconstructs original data

Using different MLTPPDM algorithms it produced different pertubated copy when collusion attack applied on it easily reconstruct original data.



Proposed System

```
Original Copy            Perturbed Copy
25    96452              6745    15726148
25    10000              6745    15812600
30    90000              6740    15732600
22    75000              6748    15747600
24    45000              6746    15777600
30    96452              6740    15726148
35    10000              6735    15812600
32    90000              6738    15732600
24    75000              6746    15747600
35    45000              6735    15777600
30    96452              6740    15726148
25    10000              6745    15812600
22    90000              6748    15732600
24    75000              6746    15747600
35    45000              6735    15777600
30    96452              6740    15726148
25    10000              6745    15812600
```

Fig.3 Applying nonlinear attack original data is not reconstructs

In proposed system applying masking noise transformation algorithm produced noise in original data by adding extra field α which is unique, so when collusion attack is applied on same it cannot reconstruct original data and it preserve the privacy. To compare between existing algorithms and proposed algorithm in terms of space complexity. Existing algorithms require more space than proposed algorithm, hence proposed algorithm increases reliability and performance.

Table 2: Result

Original **Data**		**MLT** **PPDM**		**Masking noise linear transformation**	
Age	Income	Age	Income	Age	Income
22	9000	22	9000	1329	3068068
24	75000	24	75000	1324	3154520
35	45000	35	45000	1332	3074520
30	96452	30	96452	1319	3089520
25	1000	25	1000	1330	3119520

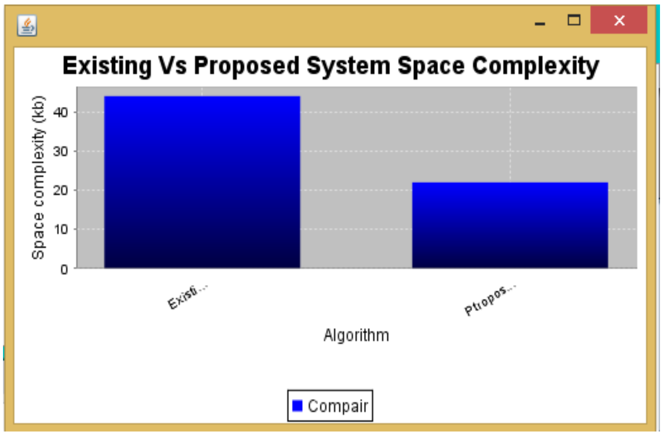


Fig.4 Space complexity

7. CONCLUSIONS

Privacy preserving data mining expand the scope of data perturbation to multilevel trust means data owner produced different perturbated copies at different trust level. The main aim of proposed approach is to preserve privacy for sensitive data. The existing system is limited only for linear attack, but proposed system working on nonlinear attack. The existing system is limited to preserve privacy for sensitive data against linear attacks. In existing approach data owner sends number of perturbated copies to the trusted third party, adversary combining all perturbated copies at different trust levels. When nonlinear collusion attack is applied on MLTPPDM algorithms it is possible to reconstruct the original data. Data owners to generate different perturbed copies of its data at arbitrary trust levels. This offers the data owner maximum flexibility. Proposed approach usage masking noise linear transformation algorithm is used to add noise into original data. When applying nonlinear collusion attack, it is not possible to reconstruct the original data, hence proposed system preserves the privacy for sensitive data. The nonlinear collusion attack is applying on different perturbated copies. The data owner has authority to produced different perturbated copies at different trust levels. The perturbated copies send to user. Then user retrieved it by using linear least square estimation. It gives an error rate, if error rate is minimum then it difficult to get original data. If it is maximum then added to each copy of the perturbated data from n series of data. Then calculate mean of each perturbated copies. This mean is actually original data called as nonlinear collusion attack. The proposed algorithm masking noise linear transformation is used to add noise into original data. It calculates empirical distribution function for every original data.Converting the empirical distribution function into a uniform random variable and standard normal random variable. Then adding noise to the standard normal variable. The quality of proposed approach and algorithm is measured in terms of reliability and performance. The existing algorithms require more space than proposed algorithm. The existing algorithms are divided into three parts i.e. parallel, sequence, on demand generation. It is used to add noise into original data. The proposed algorithm usage masking noise linear transformation algorithm which used empirical distribution function and calculate noise. This algorithm consists only one part, hence it require less space. The existing algorithms require 44kb space in the system and proposed algorithm requires 22kb space in the system. The existing system requires more time to detect collusion attack as compared proposed system. Existing system requires 20ms to detect collusion attack but it is possible to reconstruct original data and proposed system requires 12ms to detect collusion attack. The proposed approach is compared with existing approach in terms of space complexity and time complexity. The proposed approach increases reliability and performance as compared existing approach. Existing system used different MLTPPDM algorithms to add noise into original data. Existing system is limited to preserve privacy for sensitive data against linear attack, but proposed system preserve privacy for sensitive data. The system uses census dataset contains one million tuples with four attributes: Age, Education, Occupation, and Income. The age and income attributes are sensitive attributes. Proposed system preserves privacy for that sensitive attributes. Proposed system used masking noise linear transformation to produce noise into original data. When applying nonlinear attack, it is not possible to reconstruct original data, hence proposed system preserve privacy for sensitive data against nonlinear attack. It proves that proposed system is more secure than existing system. PPDM consist of MLTPPDM which produced different perturbated copies at different trust level. The existing system is limited only for linear attack, but proposed system working on nonlinear attack. Linear attack is calculating average between all perturbated copies. Nonlinear attack is calculating minimum, maximum, median function estimation. Proposed system is strengthening by using one of the algorithm called as masking noise linear transformation which gives better solution of data security for MLTPPDM against Non-linear attack. When data owner sends various perturbated copies to trusted third party for preservation purpose. The adversary cannot discover original copy from the huge number of perturbated copies, means the adversary diverse from original copy. Previous technique combines all multiple perturbated copies. When nonlinear attack i.e. collusion attack is applied on MLTPPDM approach. It is possible to reconstruct original data, which is not acceptable through the data owner. Proposed system used masking noise linear transformation to generate noise into original data. When applying same nonlinear attack i.e. collusion attack on algorithm. It is not possible to reconstruct original data, hence it preserve the privacy for sensitive data against nonlinear attack. The future scope is proposed approach is applicable only for offline dataset. It is not tested for online dataset. As users/attacker may attack online, same approach can be applicable for that with some modification.

8. REFERENCES

[1] Yaping Li, Minghua Chen, Qiwei Li, and Wei Zhang, “Enabling Multilevel Trust in Privacy Preserving Data Mining”, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 24, NO. 9, SEPTEMBER 2012**.**

[2] R. Agrawal and R Srikant, “Privacy Preserving Data Mining,”Proc. ACM SIGMOD Int’l Conf. Management of Data (SIGMOD ’00), 2000.

[3] K. Chen and L. Liu, “Privacy Preserving Data Classification with Rotation Perturbation,” Proc. IEEE Fifth Int’l Conf. Data Mining, 2005.

[4] Z. Huang, W. Du, and B. Chen, “Deriving Private Information From Randomized Data,” Proc. ACM SIGMOD Int’l Conf. Management of Data (SIGMOD), 2005.

[5] F. Li, J. Sun, S. Papadimitriou, G. Mihaila, and I. Stanoi, “Hiding in the Crowd: Privacy Preservation on Evolving Streams Through Correlation Tracking,” Proc. IEEE 23rd Int’l Conf. Data Eng. (ICDE), 2007.

[6] K. Liu, H. Kargupta, and J. Ryan, “Random Projection-Based Multiplicative Data Perturbation for Privacy Preserving Distributed Data Mining,” IEEE Trans. Knowledge and Data Eng., vol. 18,no. 1, pp. 92-106, Jan. 2006.

[7] J. Vaidya and C. Clifton,“Privacy-Preserving K-Means Clustering over Vertically Partitioned Data,” Proc. ACM SIGKDD Int’l Conf. Knowledge Discovery and Data Mining, 2003.

[8] Kanishka Bhaduri, Mark D. Stefanski and Ashok N. Srivastava, “Privacy-Preserving Outlier Detection Through Random Nonlinear Data Distortion” IEEE TRANSACTIONS ON SYSTEMS VOL. 41, NO. 1, 2011

[9] Benjamin C.M. Fung, Ke Wang, Philip S. Yu "Anonymizing Classification Data forPrivacy Preservation" IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 19, NO. 5, MAY 2007

[10] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu. Privacy

[11] Xiao-Bai Li, Sumit Sarkar "A Tree-Based Data Perturbation Approach for Privacy Preserving Data Mining" 19 July 2006.

[12] J. Vaidya and C.W. Clifton, "Privacy Preserving Association Rule Mining in Vertically Partitioned Data," Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, 2002.

[13] O. Goldreich, "Secure Multi-Party Computation," Final (incomplete) draft, version 1.4, 2002.