ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2015

Protection of Sensitive Data for Multi-Level Trust Privacy Preserving Data Mining

Rajashree R.Joshi [1], Prof. U.A.Nuli [2]

P.G. Student, Department of Computer Science& Engineering, DYPCET, Kolhapur, Maharashtra, India[1]

Assistant Professor, Department of Computer Science & Engineering, Dkte's TEI, Ichalkaranji, Maharashtra, India[2]

ABSTRACT: The study of perturbation based privacy preserving data mining (PPDM) approaches introduces random perturbation that is number of changes made in the original data. The limitation of previous solution is single level trust on data miners but new work is perturbation based PPDM to multilevel trust. When data owner sends number of perturbed copy to the trusted third party that time adversary cannot find the original copy from the perturbed copy means the adversary diverse from original Copy this is known as the diversity attack. To prevent diversity attack is main goal of MLT-PPDM services. The malicious data miner has different perturbed copy by applying different MLT-PPDM algorithms to add the noise into original data.

KEYWORDS: Privacy preserving data mining, multilevel trust, additive perturbation, Diversity attack

I. INTRODUCTION

Now a day's privacy preservation topic is issues in various organizations which depend on data mining technology. Data mining refers to extracting or mining knowledge from large amount of data. it support for user decision making process, by using data mining techniques and algorithms it prevent leakage of privacy data. At the same time, it preserves the privacy also. The problems challenge the traditional privacy-preserving data mining (PPDM) has become one of the newest trends in privacy and security and data mining research. The main goal of privacy-preserving data mining to develop such type of algorithm that original data can easily modify so that the private data remain private even after mining the process.

The Perturbed copy means number of changes is made in the original data, means adding the noise into original data. The new approach is multilevel trust in privacy-preserving data mining (MLT-PPDM) extended features for PPDM in previous approach only one perturbed copy is send to the trusted third party. But now there is multiple numbers of perturbed copies of the same data are send to the different trust level to data miners. If there is large number of trusted stages then the less number of perturbed copies can access. The main goal of MLT-PPDM is to prevent the diversity attack.

Multi-level trust privacy preserving data mining can find many applications. This work takes the initial step to enable MLT-PPDM services. Proposed system will use data perturbation techniques to protect sensitive data. It will preserve privacy for complete databases.

II. RELATED WORK

Privacy Preserving Data Mining (PPDM) was first proposed in [2] and [3] simultaneously. To address this problem, researchers have since proposed various solutions that fall into two broad categories based on the level of privacy protection they provide. The first category of the Secure Multiparty Computation (SMC) approach provides the strongest level of privacy; it enables mutually distrustful entities to mine their collective data without revealing anything except for what can be inferred from an entity's own input and the output of the mining operation alone [3]. In principle, any data mining algorithm can be implemented by using generic algorithms of SMC.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2015

Algorithms are extraordinarily expensive in practice, and impractical for real use. To avoid the high computational cost, various solutions those are more efficient than generic SMC algorithms have been proposed for specific mining tasks. Solutions to build decision trees over the horizontally partitioned data were proposed in [3]. The second category of the partial information hiding approach trades privacy with improved performance in the sense that malicious data miners may infer certain properties of the original data from the disguised data. Various solutions in this category allow a data owner to transform its data in different ways to hide the true values of the original data while at the same time still permit useful mining operations over the modified data. This approach can be further divided into data perturbation (which introduces uncertainty about individual values before data is published) [1],[2],[3],[4]. The data perturbation approach includes two main classes of methods: additive [1],[2],[3] and matrix multiplicative schemes. These methods apply mainly to continuous data. Here, additive perturbation approach is used where noise is added to data value.

Xiao et al. proposed an algorithm of multi-level uniform perturbation. Multi-level privacy preserving is proposed for additive Gaussian noise perturbation and use a measure based on how closely the original values can be reconstructed from the perturbed data.

III. PROPOSED SYSTEM

In the proposed work, Data owner want to provide some information to data miner. But it also wants to protect sensitive data such as name, age, salary. Hence to protect this information following steps are necessary:-

1) First, We need to generate the noise. In order to generate noise, we consider two algorithms such as parallel generation and sequential generation.

2) For obtaining perturbed copy, these generated noise needs to add to the original data.

3) Ten number of trust levels are taken.

4) Then, multiple differently perturbed copies will available for data miners at different trust level.

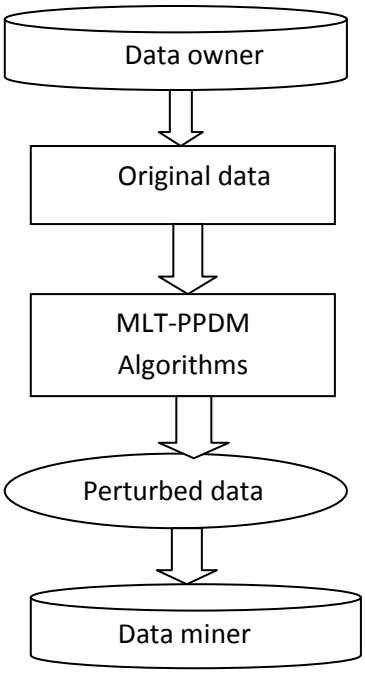

Fig 1: Mechanism of advanced MLT-PPDM

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2015

IV. EXPERIMENTAL DATASET

Stock market data which is used in 'Irrational Exuberance' is taken as original data. This data set contains one million tuples with some attributes such as data, price, dividend, earnings, index, and interest. Among these, first 10^2 tuples are taken and conducted the experiments on price, dividend, earnings, index and data.

Minimum	1871.01
Maximum	1871.1
Mean μ_χ	1871.055
Standard deviation	0.03

Table 1: Statistics of the data

V. IMPLEMENTATION

Noise is generated using batch generation. There are two batch generation algorithms i.e. parallel and sequential generation algorithm.

5.1 Parallel generation algorithm:-

Input- X , Kx

Output- Y

Construct Kz with Kx

Generate Z with Kz

Generate Y=HX + Z

Output Y

Covariance matrix of noise Z is given by the matrix,

$$K_{\mathbf{Z}} = \begin{bmatrix} \sigma^2_{Z_1} K_X & \sigma^2_{Z_1} K_X & \cdots & \sigma^2_{Z_1} K_X \\ \sigma^2_{Z_1} K_X & \sigma^2_{Z_2} K_X & \cdots & \sigma^2_{Z_2} K_X \\ \vdots & \vdots & \ddots & \vdots \\ \sigma^2_{Z_1} K_X & \sigma^2_{Z_2} K_X & \cdots & \sigma^2_{Z_M} K_X \end{bmatrix}.$$

Value for $\sigma^2 z1$ to $\sigma^2 zm$ is chosen by data owner.

5.2 Sequential generation algorithm:-

Input: X, Kx and $\sigma^2 z_1$ to $\sigma^2 zm$

Output: Y_1 to Ym

Construct Z_1

Generate $Y_1=X + Z_1$

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2015

Output Y_1

for i from 2 to m do

Construct noise ξ

Generate $Y_i = Y_{i-1} + \xi$

Output Y_i

end for

VI. EXPERIMENTAL RESULTS

From original data, we have taken only 50 rows and 10 columns for noise generation. Among these, following are some results obtained from Parallel generation algorithm.

Sample Results obtained from parallel generation algorithm:-

1870.41	0.78966	8.26	6.08655	12.3207
1868.81	3.40082	8.26	-7.6	12.9021
1870.60	4.84308	-0.359	0.9011	11.5984
1870.02	4.215281	-1.344	-1.7168	12.93227
1870.09	2.761424	-2.809	-0.2599	9.310325

Table 2: Sample results for Parallel generation

From original data, we have taken only 10 rows and 10 columns for noise generation. And number of trust levels are 10. So they released 10 perturbed copies. These copies are available to the data miners. By combining all these copies, following are some results obtained from sequential generation algorithm.

Sample Results obtained from sequential generation algorithm:-

1870.987	4.4736	0.26	0.4	13.0778
1871.004	4.07976	0.26	0.4	14.12896
1871.04	4.74	0.26	0.4	12.56
1871.05	4.86	0.26	0.4	12.27
1871.298	5.91896	0.26	0.4	7.53116

Table 3: Sample results for Sequential generation

VII. SYSTEM EVALUATION

Performance of the system is evaluated in terms of time and space complexity. Algorithm 1 has a time complexity of $O(N^3 + MN^2)$. The space complexity is $O(M + N^2)$, as Algorithm 1 has to store the noise levels $\sigma^2 1 \ldots\ldots \sigma^2 M$ and the covariance matrix KX. When generating one perturbed copy of X, then algorithm 2 of KX has a time complexity of $O(N^3)$, and the rest part costs $O(N^2)$. For generating M perturbed copies of X, Algorithm 2 has a time complexity $O(N^3+MN^2)$. Algorithm 2 only requires a memory of size $O(N^2)$ for the covariance matrix KX, as the noise levels $\sigma^2 i$ (i = 1... M) are input sequentially.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2015

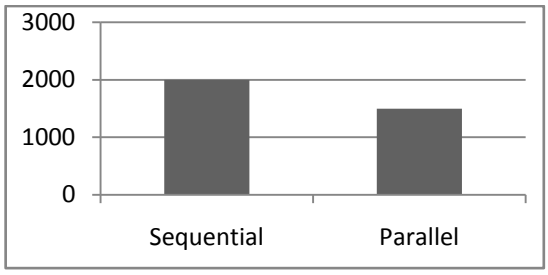

Fig 2: Time Complexity

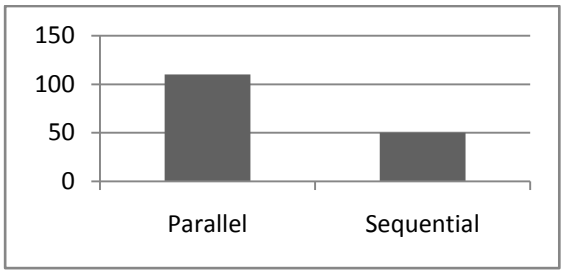

Fig 3:-Space Complexity

X-axis represents algorithms and Y-axis represents data proportion.

Sequential generation algorithms require larger time as compare to Parallel generation algorithms. Minimum space is required for sequential generation algorithm as compare to Parallel algorithm.
System also uses data mining task such as clustering for obtaining data mining results. Weka is a data mining software in java. Weka simple K-means algorithm handles mixture of categorical and numerical attribute.

Simple K means clustering is used for discovering groups and structures in the data. This algorithm randomly selects k number of objects, each of which initially represents a cluster mean or center. For remaining object, an object is assigned to cluster to which it is most similar based on the distance between object and cluster mean. Then it computes new mean for each cluster. Process iterates until criterion function converges.

Simple K means algorithm is performed on original data as well as perturbed data as shown in Fig 4 and Fig 5 respectively. Results obtained are as follows:

```
Cluster centroids:
                        Cluster#
Attribute    Full Data          0          1
                  (10)        (4)        (6)
=========================================
MYCT          1871.055   1871.025   1871.075
MMIN             4.692     4.5725     4.7717
MMAX              0.26       0.26       0.26
CACH               0.4        0.4        0.4
CHMIN           12.376    12.7225     12.145
ABC           1871.418 1871.1675   1871.585
DEF              5.335      5.325     5.3417
GHI             88.314     83.595      91.46
PQR              4.889      4.755     4.9783
STU              7.521     7.3125       7.66


Clustered Instances

0       4 ( 40%)
1       6 ( 60%)
```

Fig 4

```
Cluster centroids:
                        Cluster#
Attribute    Full Data          0          1
                  (10)        (3)        (7)
=========================================
MYCT          1871.0665   1871.03 1871.0821
MMIN             4.7227     4.6177     4.7677
MMAX              0.26       0.26       0.26
CACH               0.4        0.4        0.4
CHMIN           12.3345    12.8074    12.1318
ABC           1871.514   1871.211 1871.6439
DEF              5.3385     5.3267     5.3435
GHI             89.1434    83.8818    91.3983
PQR              4.9056     4.7243     4.9832
STU              7.5465     7.2649     7.6672


Clustered Instances

0       3 ( 30%)
1       7 ( 70%)
```

Fig 5

Simple K-means algorithm implies that results obtained from original data and perturbed data are similar.

VIII. CONCLUSION

Scope of additive perturbation based PPDM has been expanded to multilevel trust (MLT), by relaxing an implicit assumption of single-level trust in exiting work. MLT-PPDM allows data owners to generate differently perturbed copies of its data for different trust levels. The key challenge lies in preventing the data miners from combining copies at different trust levels to jointly reconstruct the original data more accurate than what is allowed by the data owner. This challenge is addressed by properly correlating noise across copies at different trust levels. Performance is evaluated in terms of Time and Space complexity. Minimum space and larger time is required for Sequential generation

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2015

algorithm. Maximum space and less time is required for Parallel generation algorithm. After implementation of clustering technique, result obtained from original data and perturbed data are similar.

REFERENCES

1] Y. Li and M. Chen, "Enabling Multi-Level Trust in Privacy Preserving Data Mining," IEEE transactions on knowledge and data engineering, sept.2012, pp 1598-1612
2] R. Agrawal and R. Srikant, "Privacy Preserving Data Mining," Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD '00), 2000,pp.439-450
3]Y. Lindell and B. Pinkas, "Privacy preserving data mining," in *Proc. Int'l Cryptology* Conference (CRYPTO), 2000, pp. 36-53
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] S. Papadimitriou, F. Li, G. Kollios, and P.S. Yu, "Time Series Compressibility and Privacy," Proc. 33rd Int'l Conf. Very Large Data Bases (VLDB '07), 2007.