ADDITIVE GAUSSIAN NOISE BASED DATA PERTURBATION IN MULTI-LEVEL TRUST PRIVACY PRESERVING DATA MINING

R.Kalaivani [#1],S.Chidambaram [#2]

[#]Department of Information Techology, National Engineering College, Kovilpatti, Tamilnadu, India

ABSTRACT

Data perturbation is one of the most popular models used in privacy preserving data mining. It is specially convenient for applications where the data owners need to export/publish the privacy-sensitive data. This work proposes that an Additive Perturbation based Privacy Preserving Data Mining (PPDM) to deal with the problem of increasing accurate models about all data without knowing exact details of individual values. To Preserve Privacy, the approach establishes Random Perturbation to individual values before data are published. In Proposed system the PPDM approach introduces Multilevel Trust (MLT) on data miners. Here different perturbed copies of the similar data are available to the data miner at different trust levels and may mingle these copies to jointly gather extra information about original data and release the data is called diversity attack. To prevent this attack MLT-PPDM approach is used along with the addition of random Gaussian noise and the noise is properly correlated to the original data, so the data miners cannot get diversity gain in their combined reconstruction.

KEYWORDS

Gaussian noise, Privacy preserving data mining, diversity attack, perturbed copies, multilevel trust.

1. INTRODUCTION

Data mining (knowledge discovery from data) is defined as the significant extraction of previously unknown, embedded and possible information from large data sets or databases. In order to store and record personal data about individuals, there are several better hardware technologies have been introduced. This has caused alarm that individual data may be used for a array of invasive or malicious purposes. Privacy preserving data mining help to achieve data mining goals without scarifying the privacy of the individuals and without learning underlying data values. Privacy-preserving data mining (PPDM) refers to the area of data mining that requests to safeguard sensitive information from unwanted or informal disclosure.

Data Perturbation is the important technique in MLT-PPDM. It is a kind of data modification approaches that save from harm, the sensitive data contained in a dataset by modifying a suspiciously selected portion of attribute-values pairs of its transactions. The employed modification makes the unconfined values inaccurate, thus protecting the sensitive data, but it also achieving protection of the statistical properties of the dataset.

The perturbation method used should be such that statistics computed on the perturbed dataset do not differ significantly from the statistics that would be obtained on the original dataset. Data perturbation approaches divide into two main categories namely probability distribution approach and the value alteration approach. The probability distribution approach changes the data with another sample from the same (estimated) distribution or by the distribution itself. On the other

DOI : 10.5121/ijdkp.2014.4303

hand, the value alteration approach perturbs the values of data elements or attributes directly by some additive or multiplicative noise before it is released to the data miner.

Data perturbation approach implicitly presumes single level trust on data miners. This approach established insecure about sensitive values of individuals, before data out to third parties [1][2][3][4][5][6][7]. This assumption limits some application where data owner trusts various levels of data miner, because data owner produces simply one perturbed copy of original data.

We present a three trust level scenario as an example.

- The banking application has three management levels each management has separate data, therefore perturb it more. The high management people can access less perturb copy and also access remain all the perturb copies. It would enviable this department does not have more power in reconstructing the original data by utilizing all perturb copies than when it has only the internal copy.
- In second scenario, if the internal copy is revealed to the low level management, then that management has equal power to the high management. It would enviable this department does not have more power in reconstructing the original data, because it uses only leaked internal copy.

This new part of Multilevel Trust creates some new problems for perturbation –based PPDM. The high trusted people can access less perturbed copy and all copies of low trust level. By using diversity across different copies the data miner can trying to reproduce the original data. The data miner able to reconstruct the data what is permitted by the data owner, this is called diversity attack.

In this paper deal with this challenge by additive perturbation approach where Gaussian noise is added to the original data, it can provides systematic solution. Our solution tolerate the generation of multiple perturbed copies of equivalent data based on the different trust levels.

2. RELATED WORK

Privacy Preserving Data Mining (PPDM) has two categories for protecting individual privacy. The first one is Secure Multiparty Computation (SMC), the basic idea of this approach is that a computation is safe if at the end of the computation, others cannot knows anything except its own input and the results. In this method enabled two parties can generate decision tree without learning anything about the other party's data.

The next category is Data Perturbation has multiple techniques 1) Additive Perturbation, 2) Multiplicative Perturbation, 3) K-anonymity, 4) Data Swapping, 5) Micro-aggregation, 6) Resampling, and 7) Data shuffling.

Additive Perturbation technique [1] [2] [4] [5][7] is masking the attribute values by adding noise to the original data. The noise added to the data is as large as possible from that the individual record cannot be recovered.

Multiplicative perturbations [3][6]can also be used to good result for privacy-preserving data mining. This technique conserve the inter record distances roughly, and therefore the altered records can be used in coincidence with various distance-intensive data mining applications. Difference between the additive and multiplicative perturbations is, the additive perturbation is reconstruct only combined distributions, but in the multiplicative perturbation preserve more sensitive information (e.g. distances).

The K-anonymity [14][15][16][17] contains two methods i.e, Generalization and suppression techniques. In generalization method the attributes values are generalized. For example, the date of birth can be generalized in the form of year of birth. In the Suppression technique the attribute

values are completely removed from that decrease the threat of recognition with use of public records, while decreasing the accuracy of applications on the changed data.

Data swapping method preserve the confidentiality in datasets that contain categorical variables and its can transform by replacing values of sensitive variables between individual records. In Micro-aggregation data is clustered into small group before publication. The common value of the group restores each value of the individual.

3. FUNDAMENTAL CONCEPTS

3.1 Multivariate Gaussian

In this paper, perturbed copy is generated by adding gaussian noise [1][2][4][5][7] to the original data that noises are multivariate gaussian noises.

Let X_1 to X_N are the N gaussian random variables then these variables are sometimes called multivariate gaussian and their joint probability density function is represented as follows:

$$F_X(x) = \frac{1}{\sqrt{(2\pi)^N \det(K_x)}} e^{-(x-\mu_x)^T K_x^{-1}(x-\mu_x)/2} \quad (1)$$

Where μ_x is mean and K_X covariance matrix of X

$$\mu_x = \begin{bmatrix} x_1 - \bar{X} \\ x_2 - \bar{X} \\ \vdots \\ x_N - \bar{X} \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_N \end{bmatrix}$$

$$K_X = \begin{bmatrix} K_{11} & K_{12} & \cdots & K_{1N} \\ K_{21} & K_{22} & \cdots & K_{2N} \\ \vdots & \vdots & & \vdots \\ K_{N1} & K_{N2} & \cdots & K_{NN} \end{bmatrix}$$

3.2 Additive Perturbation

The additive perturbation approach can generate the perturb data Y by adding the original data X with random noise Z. this can be represented as follows:

$$Y=X+Z \quad (2)$$

Here the original dataset follows the mean and covariance matrix. The covariance matrix is represented as follows

$$K_X = [(X - \mu_x)(X - \mu_x)^T] \quad (3)$$

The noise Z is a jointly Gaussian vector, mean and covariance matrix and the mean value is zero. The covariance matrix represented as follows

$$K_Z = [ZZ^T] \quad (4)$$

From that easy to obtain the mean value of perturbed copy Y is μ_x and the covariance matrix K_Z is denoted as

$$\mathbf{K_Y = K_X + K_Z}$$

Malicious data miner aim is to reconstruct the original data by removeing noise from perturbed copy. If the noise has same correlation then noise removing is difficult, or else the noise removed easily. To achieve this we have to choose $K_Z = \sigma_Z^2 K_X$. Here σ_Z^2 the perturbation magnitude.

3.3 Privacy Goal

The MLT-PPDM data owner can release multiple perturbed copies for various data miners based on the trust level. One of the main objective is to control the amount of the information about its adversaries may derive.

The data owner can control the amount of information of its data with single perturbed copy. Then the privacy of Y_i is $D(X, \hat{X}(Y_i))$ as follows

$$\mathbf{D\left(X, \hat{X}(Y_i)\right) = \frac{\sigma_{Z_i}^2}{\sigma_{Z_i}^2 + 1} \frac{1}{N} Tr(K_X)} \quad (5)$$

The data owner can easily control the privacy of individual copy through one-to-one mapping. Suppose the adversaries can access other perturbed copies merge with other data miner or revealed from any data miners, than adversaries can jointly gather from multiple perturbed copies should be more than that of the best effort using any individual copy.

The privacy goal achieved with multiple perturbed copies is represented by

$$D(X, \hat{X}(Y_c)) = \min_{Y_i \in Y_c} D(X, \hat{X}(Y_i)) \quad (6)$$

TABLE 1 -KEY TERMS

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
$\hat{X}(Y)$	LLSE Estimate of X given Y
K_X	Covariance matrix of X
K_Z	Covariance matrix of Z

Where Y_c are the group of perturbed copies. For achieving privacy goal, the smallest amount of privacy between any subset o M perturbed copies and the remaining copies in that subset contain no additional information about original data.

4. PROBLEM FORMULATION

Privacy is suitable an increasingly important issue in many data mining applications. A malicious data miner may have access to differently perturbed copies of the same data through a variety of means, and may combine these diverse copies to jointly infer additional information about the original data that the data owner does not intend to release. This is called as Diversity Attack. Every day users are transferring millions of electronic paths through various activities such as using swapping security cards, credit cards, and discussion over phones and using email. In

addition to this many organizations can sell the composed data to other organizations, which use these data for their personal purposes. For everyday activities organizations are exceedingly dependent on data mining. In the period of data mining process, sensitive information like financial and medical data can get exposed to collectors, miners. Exposure of such sensitive information can cause a violation of individual privacy. Personal information can also be revealed by linking multiple databases belonging to massive data warehouses and accessing web data. An intruder can find out sensitive attribute values such as Customer PIN number and other information's of a certain individual, through re-identification of the record from an uncovered data set. This has triggered the development of many privacy preserving data mining techniques that try to mine the data samples without directly accessing the original data and assurance that the mining process does not get plenty information to reconstruct the original data. Data Perturbation is a popular technique in PPDM and perturbation-based PPDM approach establishes Random Perturbation to individual values before data are published. The extent of perturbation-based PPDM is extended to Multi-Level Trust (MLT-PPDM). Even though MLT-PPDM is robust against diversity attacks, data perturbation is not supported by MLT-PPDM. Additionally MLT-PPDM considers only linear attacks but more authoritative adversaries apply nonlinear techniques to obtain original data and get more information.

5. PROBLEM SOLUTION

A data miner can access different perturbed copies based on their trust level. In this section present the challenge to reaching the privacy goal with two states. Consider the data owner can generate two perturbed copies for two trust level data miners. The data owner can generate perturbed copy of

$$Y_2=X+Z_2$$

The Gaussian noise Z_2~N $(0,\sigma_2^2\sigma_x^2)$ is independent of X. Then the data owner can generate another perturbed copy Y_1. Gaussian noise for this perturbed copy is Z_1~N $(0,\sigma_1^2\sigma_x^2)$, these noise is also independent of X. the data owner chooses $\sigma_2^2 > \sigma_1^2$ so that Y_2 more perturbed than Y_1.

The privacy goal for this two perturbed copy is

$$D(X,\hat{X}(Y_1,Y_2)) = D(X,\hat{X}(Y_1)). \qquad (7)$$

The two scenarios are 1) Batch generation, 2) On Demand generation.

5.1 Batch Generation

In this method, the data owner produce the M trust level for that M perturbed copies are generated. In this generation works as either parallel or sequentially.

5.1.1 Parallel Generation

In this method the mechanism of noise Z, i.e., Z_1 to Z_M, are generated concurrently based on the probability distribution function.

***Algorithm 1**: Parallel Generation*

1. Input: X, K_X, and $\sigma_{Z_1}^2$ to $\sigma_{Z_M}^2$
2. Output: Y
3. Construct K_Z with K_X and $\sigma_{Z_1}^2$ and $\sigma_{Z_M}^2$
4. Generate Z with K_Z
5. Generate Y = HX + Z
6. Output Y

Algorithm 1 serves as a basic algorithm for sequential and on demand generation algorithms.

Consider the data owner generates two perturbed copies. This perturbed copies are independent of original dataset X and independent of each other. Then the malicious data miner to perform joint LLSM and obtain privacy value for that combine perturbed copies.

$$D(X, \hat{X}(Y_1, Y_1)) = \frac{\sigma_X^2}{1 + 1/\sigma_1^2 + 1/\sigma_2^2} \quad (8)$$

This value definitely smaller than the error estimation of single perturbed copy Y_1 or Y_2

$$D(X, \hat{X}(Y_1)) = \frac{\sigma_X^2}{1 + 1/\sigma_1^2} \quad (9)$$

Thus,(9) is not satisfied. and it cannot be achieve desired privacy goal (7).

5.1.2 Sequential Generation

In this method, M independent noise Z_1 to Z_m Sequentially generated, and $(Z_i - Z_{i-1})$ for i from 2 to M. Algorithm 1 needs large memory so we cannot obtain memory efficient solution. Compare to parallel generation, sequentially generating noise Z, each of which a Gaussian vector of N dimension.

Algorithm 2*: Sequential Generation*

1. Input: X, K_X, and $\sigma_{Z_1}^2$ and $\sigma_{Z_M}^2$
2. Output: Y_1 to Y_M
3. Construct $Z_1 \sim N(0, \sigma_{Z_1}^2 K_X)$
4. Generate $Y_1 = X + Z_1$
5. Output Y_1
6. for i from 2 to M do
7. Construct noise $Z_2 \sim N(0, (\sigma_{Z_i}^2 - \sigma_{Z_{i-1}}^2)K_X)$
8. Generate $Y_i = Y_i - 1 + Z_2$
9. Output Y_i
10. end for

Consider this approach generates two copies then the malicious data miner using various trust level copies reconstructs X as perfect as follows

$$X = \frac{\sigma_2 Y_1 - \sigma_1 Y_2}{\sigma_2 - \sigma_1} \quad (10)$$

The estimation error value is zero, from that privacy goal (7)is not satisfied.

The main drawback of the batch generation is that it requires a data owner to predict all potential trust levels before producing perturbed copies. So On-demand generation scheme is proposed.

5.1.3 On Demand Generation

In contrasting to the batch generation, new perturbed copies are introduced on demand in this scenario.

Algorithm 3*: On Demand Generation*

1. Input: X, K_X, $\sigma^2_{Z_i}$ to $\sigma^2_{Z_M}$, and values of Z′:v_1
2. Output: New copies Z″
3. Construct K_Z with K_X and $\sigma^2_{Z_i}$ to $\sigma^2_{Z_M}$
4. Extract $K_{Z'}$,$K_{Z''Z'}$, and $K_{Z''}$ from K_Z
5. Generate Z″ as a Gaussian with mean and variance
6. for i from L + 1 to M do
7. Generate $Y_i = X + Z_i$
8. Output Y_i
9. end for

Imagine L (L < M) existing copies of Y_1 to Y_L, so that the data owner, ahead requests, generates additional M–L copies of Y_{L+1} to Y_M. Among three techniques on-demand generation offers data owner's maximum flexibility where data owners generate perturbed copies of the data at random trust levels

Basically privacy goal requires Y_1 is more perturbed than Y_2 does not improve efficiency. To satisfies the privacy goal can generate y2 as follows

$$Y_2 = Y_1 + (Z_2 - Z_1)$$

Y2 is perturbed from Y1. Then the joint error estimation value is same as the error estimation of single perturbed copies. The privacy goal (7) is achieved.

6. EXPERIMENTS

We run our experiments on a dataset like banking dataset, which is commonly used for privacy preserving for carrying their performance fully controlled manner. This data set holds multiple attributes: name, pin number, mobile number etc. We take several tuples and perform the experiments on the pin and mobile number. The statistics and distribution of the data are shown in Fig. 1, respectively.

Given input data X (Pin and Mobile number), to generate perturbed copies Yi at different trust levels i, we generate Gaussian noise Z_i according to N(0,$\sigma^2_{Z_i}$ K_X) and add Z_i with X. The constant $\sigma^2_{Z_i}$ stand for the perturbation magnitude determined by the data owner along with the trust level i.

In the Experiment have two performance metrics one is average normalized estimation error and next is distribution of estimation error. The normalized estimation error is obtained using LLSE estimation i.e.,$\hat{X}(Y)$. And the normalized estimation error is represented as

$$\frac{D\left(X, \hat{X}(Y)\right)}{Tr(K_X)} \qquad (11)$$

It gets values between 0 and 1. If the value is smaller then the LLSE estimation is more accurate. For more perturbed copies LLSE estimation normally decreases. The distribution of the estimation error is represented as

$$\sqrt{D\left(X, \hat{X}(Y)\right)} \qquad (12)$$

Figs. 2 show the normalized estimation errors of both on demand and batch generation.$\sigma^2_{Z_i}$as a perturbation magnitude of the number of perturbed copies.

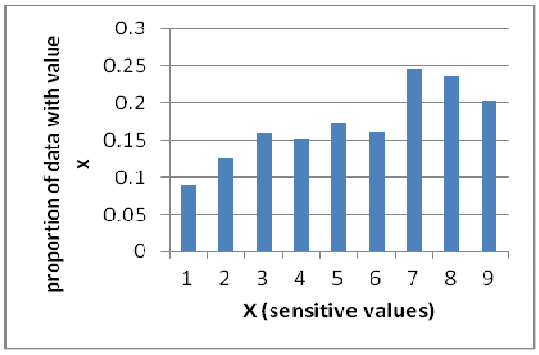

Figure 1 Distribution of sensitive value Pin and Mobile number.

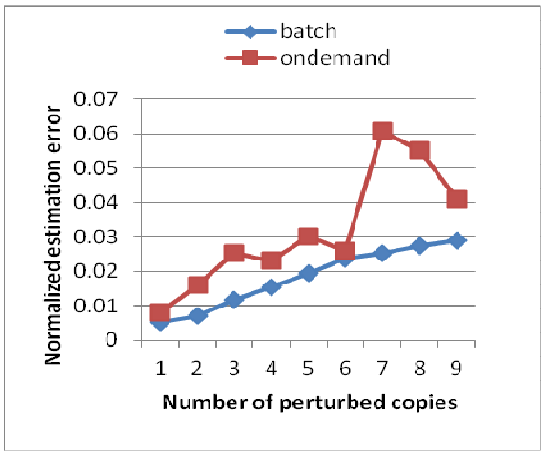

Figure 2. Normalized estimation error for batch and On demand generation

7. CONCLUSION

In this system, develop the scope of additive perturbation based PPDM to Multilevel trust. The MLT-PPDM produces various perturbed copies of the identical data for various trust levels. It prevents from diversity attacks i.e., data miners can jointly reproduce the original data more accurately by compared with the owner allowed data. Prevention of diversity attack can be done by appropriately correlating noise across at various trust levels and proved that the noise covariance matrix has corner wave property, and then the data miners have no diversity gain. Proposed solution allows admin to produce perturbed copies on demand. The proposed method provides maximum flexibility.

REFERENCES

[1] D. Agrawal and C.C. Aggarwal, "On the Design and Quantification of Privacy Preserving Data Mining Algorithms," Proc. 20th ACM SIGMOD-SIGACT-SIGART Symp. Principles of Database Systems (PODS '01), pp. 247-255, May 2001.

[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] 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.

[8] Y. Lindell and B. Pinkas, "Privacy Preserving Data Mining," Proc. Int'l Cryptology Conf. (CRYPTO), 2000.

[9] 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.

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

[11] 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.

[12] A.W.-C. Fu, R.C.-W. Wong, and K. Wang, "Privacy-Preserving Frequent Pattern Mining across Private Databases," Proc. IEEE Fifth Int'l Conf. Data Mining, 2005.

[13] B. Bhattacharjee, N. Abe, K. Goldman, B. Zadrozny, V.R. Chillakuru, M.del Carpio, and C. Apte, "Using Secure Coprocessors for Privacy Preserving Collaborative Data Mining and Analysis," Proc. Second Int'l Workshop Data Management on New Hardware (DaMoN '06), 2006.

[14] C.C. Aggarwal and P.S. Yu, "A Condensation Approach to Privacy Preserving Data Mining," Proc. Int'l Conf. Extending Database Technology (EDBT), 2004.

[15] E. Bertino, B.C. Ooi, Y. Yang, and R.H. Deng, "Privacy and Ownership Preserving of Outsourced Medical Data," Proc. 21st Int'l Conf. Data Eng. (ICDE), 2005.

[16] D. Kifer and J.E. Gehrke, "Injecting Utility Into Anonymized Datasets," Proc. ACM SIGMOD Int'l Conf. Management of Data, 2006.

AUTHORS

Kalaivani.R, I did my UG degree in P.S.R Rengasamy college of Engineering for Women, And also I am doing my PG degree under the Department of Information Technology in National Engineering College, Kovilpatti.

Mr.S.Chidambaram He did PG degree in National Engineering College, Kovilpatti, Persuing Ph.D and works as an Assistant Professor in the department of Information Technology in National Engineering College, Kovilpatti.