Privacy Preserving through Data Perturbation using Random Rotation Based Technique in Data Mining

Mr. Swapnil Kadam, Prof. Navnath Pokale

***Abstract*— Data perturbation technique is, a widely employed and accepted Data Mining (PPDM) approach, used to single level trust on data miners. Privacy Preserving Data Mining deals with the problem of developing accurate models about aggregated data without access to precise information or original records in individual data record. Perturbation-Based PPDM approach deals random perturbation to the individual values for preserving the privacy of data before data are published. Earlier work of this approach are not suitable of single-level trust on data miners. In this work, we considering this assumption for expand the scope of Perturbation-Based PPDM to Multilevel Trust (MLT-PPDM). the less perturbed copy can access permissions to the more trusted a data miner. Under this, a malicious data miner have access to different copies of the same data through various forms, and it combine these copies to jointly infer additional metadata about the original data that the data owner does not intend to release. Preventing diversity attacks is the challenge of providing MLT-PPDM services. here trying to resolve this challenge by properly assignment perturbation across copies at different trust levels. We prove that our solution is good against diversity attacks with respect to our privacy policy. That is, arbitrary collection of the perturbed copies access an data miners, our technique prevent them from jointly reconstructing the original data more accurately than the best effort using any individual copy in the collection. Our technique is useful to a data owner to produce perturbed copies of its data for as per trust levels on demand. This technique offers data owners maximum flexibility.**

***Index Terms*— Multilevel Trust, Privacy Preserving Data Mining, random perturbation.**

.

I. INTRODUCTION

Preserving privacy in Data Mining (PPDM) technique introduces uncertainty about individual values before data are published or released to third parties for data mining purposes [1], [2], [3], [4], [5], [6], [7],[16],[17]. In the single level trust assumption, the data owner or admin can create only single perturbed copy of its data with a fixed amount of uncertainty. So this type of assumption will limited in different applications where a data owner or admin trusts a data miners at different levels. We are going to present a two trust level scenario as a motivating example as described below. The business or a government have to do useful internal data mining which should be most trusted , but they may also want to release the data to the public, and might perturb it more. The less perturbed internal copy is received by a mining department. This mining department also has access to the more perturbed public copy. It could be desirable that mining department do not have any more authority for recreating the original data by utilizing both copies than when it has only the internal copy. Similarly, when an internal copy is discharged to the public, then obviously the public has all the power of the mining department. Although, it could be advisable if the public can't recreate the original data more accurately when it uses both copies than when it uses only the leaked internal copy. These new dimensions of Multi-level Trust (M.L.T.) poses challenges for PPDM which is based on perturbation. In contrast to the single-level trust scenario where only one perturbed copy is released, now more number of differently perturbed copies of the actual data are accessible to data miners at various trusted levels. It means the less perturbed copy can be accessed by the more trusted data miner ; it may also have access to the perturbed copies available at lower trust levels. Generally, the data miner could access multiple perturbed copies through various other means, e.g., accidental leakage or colluding with others. By utilizing diversity across differently perturbed copies, the data miner may be able to produce a more accurate data from original data than what is allowed by the data owner. attack is called as diversity attack. It includes the colluding attack scenario where adversaries combine their copies to mount an attack; it also includes the scheme where an utilizes public information to perform the attack on its own. Preventing diversity attacks is the key challenge in solving the MLT-PPDM problem.In this paper, we address this challenge in enabling MLT-PPDM services. In particular, watching on the additive perturbation approach where random Gaussian noise is added to the original data with random distribution, and provide a systematic solution. Through a one-to-one mapping, our solution allows a data owner to generate various perturbed copies of its data according to different trust levels. Defining trust levels and determining such mappings are beyond the scope of this paper.

***Manuscript receivedJan, 2016*.**

***Mr Swapnil Kadam** pursuing Master's degree in Computer Engineering at Department of Computer Engg.,BSCOR, Narhe, Pune University.9423817389*

***Prof. Navnath B. Pokale** Curently working as a Assistant Professor at Department of Computer Engg.,BSCOR, Narhe, Pune University.*

ISSN: 2278 – 1323

A. Contributions

Following are the some contributions.

- Design and implement perturbation-based PPDM to multilevel trust, to generate different perturbed copies of its data based on trust levels and introducing random rotation based algorithm.
- design an mechanism to identify goals in enabling MLT-PPDM services. In MLT-PPDM, data miners have access to multiple perturbed copies. By joining different duplicates, information excavators might have the capacity to perform assorted qualities assaults to recreate the first information more precisely than what is permitted by the data owner.
- Implement method to properly assigning perturbation across copies at per trust levels. here introducing is robust solution against diversity attacks.
- Design a solution that allows data owners to generate perturbed copies of their data as per trust levels on-demand.

owner to generate distinctly perturbed copies of its data according to different trust levels. Defining trust levels and determining such mappings are beyond the scope of this paper.

II. LITERATURE SURVEY

Data Mining (PPDM) was initially proposed in [2] and [8] all the while. To address this issue ,scientists have subsequent to proposed different arrangements that fall into two general classifications in view of the level of security assurance they give. The first class of the Secure Multiparty Computation (SMC) approach gives the most grounded level of protection; it empowers commonly doubtful substances to mine their aggregate information without uncovering anything aside from what can be induced from an element's own data and the yield of the mining operation alone [8],[9]. On a fundamental level, any information using so as to mine calculation can be actualized nonexclusive calculations of SMC [10].However, these calculations are uncommonly costly by and by, and unfeasible for genuine use. To maintain a strategic distance from the high computational cost, different arrangements that are more effective than nonexclusive SMC calculations have been proposed for particular mining assignments. Answers for assemble choice trees over the on a level plane parceled information were proposed in [8] . For vertically apportioned information, calculations have been proposed to address the affiliation principle mining [9], k-implies bunching [11], and continuous example mining issues [12]. The work of [13] uses a protected coprocessor for security saving synergistic information mining and investigation.

The second classification of the incomplete data stowing away approach exchanges protection with enhanced execution as in malevolent information excavators might derive certain properties of the first information from the hidden information. diffrent arrangements in this classification permit an information proprietor to change its information in diverse approaches to shroud the genuine estimations of the first information while in the meantime still allow helpful mining operations over the altered information. This methodology can be further partitioned into three classifications: 1) k- anonymity [14], [15], 2) maintenance substitution (which holds a component with likelihood p or replaces it with a component chose from a likelihood appropriation capacity on the area of the components) , and 3) information irritation (which presents instability about individual qualities before information are distributed) [1], [2], [3], [4], [5], [6], [7].The information annoyance approach incorporates two principle classes of strategies: added substance [1], [2], [4], [5], [7] and framework multiplicative [3], [6] plans. These strategies apply for the most part to persistent information. In this paper, we concentrate singularly on the added substance annoyance approach where clamor is added to information values Little trusted gadgets were utilized for secure capacity assessment as a part of [16].

.

III. PAPER LAYOUT

The rest of the paper is organized as follows: we go over methodologies in Section 4. We formulate the problem, and define our privacy goals in Section 5.1. It highlights the key challenge in achieving our privacy goals, and presents the ENABLING MULTILEVEL TRUST IN PRIVACY PRESERVING DATA MINING intuition that leads to our solution. In Section 6, we formally present our solution and algorithms , and prove that it achieves our privacy goal.

IV. METHODOLOGIES

A. Jointly Gaussian

In this work, we concentrate on perturbing data by added noises in Gaussian clamor [1],[16],[17]. i.e., the included commotions are together Gaussian. 1. Let G1 through GL be L Gaussian variables. They are said to be together Gaussian if and just if each of them is a direct combination of various free Gaussian random variables.

2. Proportionally, G1 through GL are together Gaussian if and just if any straight combination of them is likewise a Gaussian random variable. A vector framed by mutually Gaussian arbitrary variables is known as a together Gaussian vector.

B. Jointly Gaussian

Given an irritated duplicate of the information, a vindictive information mineworker might endeavor to remake the first information as precisely as would be prudent. Among the group of straight reproduction strategies, where assessments must be direct elements of the bothered duplicate, Linear Least Squares Error (LLSE) estimation has the base square blunders between the evaluated values and the first values[30],[31].estimation is that it all the while minimizes all these estimation mistakes.
$Y = X + Z$

C. Linear Least Squares Error Estimation

ISSN: 2278 – 1323

Given an irritated duplicate of the information, a vindictive information mineworker might endeavor to remake the first information as precisely as would be prudent. Among the group of straight reproduction strategies, where assessments must be direct elements of the bothered duplicate, Linear Least Squares Error (LLSE) estimation has the base square blunders between the evaluated values and the first values[30],[31].estimation is that it all the while minimizes all these estimation mistakes.

V. PROBLEM FORMULATION

Privacy Preserving Data Mining (PPDM) addresses the problem of developing accurate models about aggregated data without access to precise information in individual data record. A previously studied perturbation-based PPDM approach introduces random perturbation to individual values to preserve privacy before data are published. Available previous solutions of this approach are limited in their tacit assumption of single-level trust on data miners. In this work, by using random rotation based data perturbation, as perour assumption ,trying to expand the scope of perturbation-based PPDM to Multilevel Trust (MLT-PPDM). In our work, the less perturbed copy of the data can only access the more trusted a data miner.

Table 1: Key Notations

Notation	**Definition**
X	Input Original data Sheet
Yi	Perturbed copy of X of trust level i
Ri	Noise added to X to generate Yi
M	Number of trust levels
N	Number of attributes in X
Y’	A vector of all M perturbed copies
R’	A vector of noise R1 to RM
KX	Covariance matrix of X
KR’	Covariance matrix of R’

A. Problem Formulation

In the MLT-PPDM problem, we consider in this work, a data owner trusts data miners at different levels generates a series of perturbed copies of its data for different trust levels. This is done by adding varying amount of noisy data to the original data.

Under the multilevel trust setting, data miners at higher trust levels can access less perturbed copies. Such less perturbed copies are not accessible by data miners at lower trust levels. the motivating example here give at the beginning , data miners at higher trust levels may also have access to the perturbed copies at more than one trust levels. Data miners at different trust levels may also collude to share the perturbed copies among them. As latter such, it is common that data miners can have access to more than one perturbed copies. Specifically, we assume that the data owner wants to release M perturbed copies of its data X, which is an N x 1 vector with mean and covariance Kx as defined . These M copies can be generated in various fashions. They can be jointly generated all at one time .Alternatively, they can be generated at different times upon receiving new requests from data miners, in an on-demand fashion. The latter case gives data owners maximum flexibility. It is true that the data owner may consider to release only the mean and covariance of the original data. We remark that simply releasing the mean and covariance does not provide the same utility as the perturbed data. in many real time applications, knowing only the mean and covariance may not sufficient to apply data mining methods, such as clustering classificationsn, PCA [6]. By using random perturbation technique to release the data set, the data owner allows the data miner to exploit more statistical information without releasing the exact values of sensitive attributes.

VI. SOLUTION TO GENERAL CASES

We now show that the solutions to the general cases of arbitrarily fine trust levels follow naturally from that to the trust levels.

A. Shaping the Noise

1. *Independent Noise Revisited*

In Section 4, shows adding independent noise to generate '2' differently perturbed copies, although convenient, fails down to achieve our privacy . The increase in the number of differently generated copies aggravates the situation; the error actually goes to 0, as this number increases indefinitely. In this case, the attackers can perfectly reconstruct the original data tuple. We define this observation in the following methodologies.

2. *Properly Correlated Noise*

We show by the case study that the key to achieving the desired privacy goal is to have noise $Z_i, 1 \leq i \leq M$ properly correlated. To this end, we further develop the pattern found in the 2 * 2 noise covariance matrix in (13) into a corner-wave property for a multidimensional noise covariance matrix. This property becomes the cornerstone of Theorem 4 which is a generalization of Theorems 1 and 2. Corner-wave Property. here states that for M perturbed copies, the privacy goal is achieved if the noise covariance matrix $K_{\mathbb{Z}}$ has the corner-wave pattern. Specifically, we say that an M * M square matrix has the corner-wave property if, for every i from 1 to M, the following entries have the same value as the (I,i) th entry:

- all entries to the right of the (i,i) th entry in row i, and
- all entries below the (i,i) th entry in column i.

ISSN: 2278 – 1323

The distribution of the entries in such a matrix looks like corner-waves originated from the lower right corner.

B. Batch Generation

In the first scenario, the data owner determines the M trust levels priori, and generates M perturbed copies of the data in one batch. In this case, all trust levels are predefined and $\sigma_{Z_1}^2$ to $\sigma_{Z_M}^2$ are given when generating the noise. We refer to this scenario as the batch generation. We propose two batch algorithms. Algorithm 1 generates noise Z1 to ZM in parallel while Algorithm 2 sequentially.

1. *Algorithm 1: Parallel Generation*

Without loss of generality, we assume $\sigma_{Z_1}^2$ to $\sigma_{Z_M}^2$ where $1 \le i \le M-1.$. Algorithm 1 generates the components of noise $\mathbb{Z}$, i.e., Z1 to ZM, simultaneously based on the following probability distribution function, for any real (N.M)- dimension vector v,

$$f_{\mathbf{Z}}(v) = \frac{1}{\sqrt{(2\pi)^M \det(K_{\mathbf{Z}})}} e^{-\frac{1}{2}v^T K_{\mathbf{Z}}^{-1} v},$$

Where $K_{\mathbb{Z}}$ is given by Algorithm 1 then constructs $\mathbb{W}$ as HX -$\mathbb{Z}$ and outputs it. We refer to Algorithm 1 as parallel generation. Algorithm 1 serves as a baseline algorithm for the next two algorithms[30],[31].

Input: X, KX, σ_{R1}^2 to σ_{RM}^2
Output: Y'
Construct KR' with KX and σ_{R1}^2 to σ_{RM}^2
Generate vector of noise R1 to RM with KR'
i.e. R' with R_1 to R_M
Generate Y'=HX+R'
Where H = $[I_N . . . I_N]^T$
Output Y'

2. *Algorithm 2: Sequential Generation*

The large memory requirement of Algorithm 1 motivates us to seek for a memory efficient solution. Instead of parallel generation, sequentially generating noise Z1 to ZM, each of which a Gaussian vector of N dimension. The validity of the alternative procedure is based on the insight in the following theorem[16],[17].

Input: X, K_X, σ_{R1}^2 to σ_{RM}^2
Output: Y_1 to Y_M
Construct R_1~N (0, $\sigma_{R1}^2 K_X$)
Generate Y_1=X+R_1
Output Y_1
For i from 2 to M do
Construct noise ξ ~N (0, (σ_{Ri}^2 -σ_{Ri-1}^2) K_X)
Generate Y_i=Y_{i-1}+ξ
Output Y_i
End for

3. *Disadvantages*

The main disadvantage of the batch generation is that it requires a data owner to foresee all possible trust levels a prior. This requirement is not flexible and sometimes impossible to meet criteria. One such scenario in our case study is the data owner already released a perturbed copy Y2, a new request for a less distorted copy Y1 arrives. The sequential generation algorithm cannot handle such requests since the trust level of the new request is lower than the existing one. In today' sever-changing world, it is desirable to have technologies that adapt to the dynamics of the society. In our problem generating new perturbed copies on-demand would be a vital feature.

C. On-Demand Generation

Instead of the clump era, new bothered duplicates are presented on interest in this situation. Following the solicitations might be subjective, the trust levels comparing to the new duplicates would be discretionary too. The new duplicates can be either lower or higher than the current trust levels. We consider this situation as on-interest era technique. to Achieving the protection objective in this situation will give information proprietors the most extreme adaptability to give MLT-PPDM administrations.

Algorithm 3: Random rotation based data perturbation algorithm[16].

Input: X, K_X, σ_{R1}^2 to σ_{RM}^2
Output: Y
Generate Noise K_X
Random(range);
Output Y=X+Z

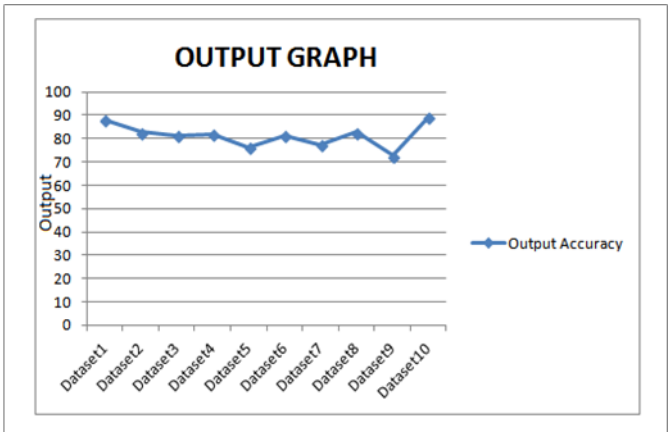
Random n:m
For i from n to M do
Output n
End for

VII. RESULT AND DISCUSSION

Here the various perturbed copies are generated with different trust level os trust. And intruders cannot completely reconstruct the original copy of data by knowing the additional perturbed copies. as per the user the number of perturbed copies can generated. Here data owner have the maximum flexibility, so the data owner can release what he intended to release. following graph shows how the privacy is preserved in terms of percentage.

ISSN: 2278 – 1323

Sr.No	Expected Output	System Output	Percentage Output
Dataset1	100	88	88
Dataset2	93	89	82
Dataset3	98	83	81
Dataset4	91	90	81
Dataset5	92	83	76
Dataset6	97	84	81
Dataset7	85	91	77
Dataset8	92	90	82
Dataset9	99	73	72
Dataset10	99	90	89



GRAPH1: OUTPUT GRAPH

Graph 1 shows the output accuracy of the perturbed data sheet. result shows conclusion i.e. higher accuracy of the data shows complexity of the data is higher .

VIII. CONCLUSION

In this work, we expand the scope of additive perturbation based PPDM 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. here we can solution on the problem addressed by properly correlating noise across copies at different trust levels. This property offers the data owner maximum flexibility. We believe that multilevel trust privacy preserving data mining can find many applications. enable MLT-PPDM services we introduced one step . Many interesting and important directions are worth exploring. For instance, it is not clear how to extend the extent of different methodologies in the zone of incomplete data stowing away, for example, arbitrary revolution based information bother, k anonymity, and maintenance substitution, to multilevel trust. It is likewise of incredible enthusiasm to extend our way to deal with handle advancing information streams. Likewise with most existing work on bother based PPDM, our work is constrained as in it considers just direct assaults. All the more intense enemies might apply nonlinear systems to infer unique information and recuperate more data. Concentrating on the MLT-PPDM issue under this ill-disposed model is a fascinating future scope.

IX. 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'lConf. Knowledge Discovery and Data Mining, 2002.*

[10] O. Goldreichi, "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.A. Huberman, M. Franklin, and T. Hogg, "Enhancing Privacy and Trust in Electronic Communities," *Proc. First ACM Conf.Electronic Commerce, pp. 78-86, Nov. 1999.*

[14] M. Freedman, K. Nissim, and B. Pinkas, "Efficient Private Matching and Set Intersection," *Advances in Cryptology—EUROCRYPT, vol. 3027, pp. 1-19, 2004.*

ISSN: 2278 – 1323

[15] L. Kissner and D. Song, "Privacy-Preserving Set Operations," *Proc. Int'l Cryptology Conf. (CRYPTO), 2005.*

[16] P. Valake Tejashri and S. Patil Sachin A Enabling Multilevel Trust Privacy Preserving Data Mining using Random Rotation Based Data Perturbation © *Elsevier Publications ERCICA-2014.*

[17] Enabling Multilevel Trust in Privacy Preserving Data Mining *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 24, NO. 9, SEPTEMBER 2012*

Mr. Swapnil Kadam received the Bachelor degree (B.E.) in Computer Engineering .Currently, He is pursuing Master's degree in M.E. Computer Engineering at Department of Computer Engg.,BSCOR, Narhe, Pune University. His current research interests include Data mining .

Prof. Navnath B. Pokale obtained M.E. computer. Curently working as a Assistant Professor at Department of Computer Engg.,BSCOR, Narhe, Pune University. He has 14 yrs of teaching experience. His research interests include Networking, Image processing and Data Mining.

ISSN: 2278 – 1323