Implementation of Multi-Level Trust in Privacy Preserving Data Mining against Non-linear attack

Vaishali Bhorde[#1], Prof. S.M. Tidke [*2]

[#]*Computer Engineering Department, Pune University*
JSPM's Imperial college Engg. & Research Wagholi Pune
bhorde.vaishali1@gmail.com

[#]*Computer Engineering Department, Pune University*
JSPM's Imperial college Engg. & Research Wagholi Pune
Sonalitidke11@gmail.com

***Abstract*--**The study of perturbation based 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 pertubated copy to the trusted third party that time adversary cannot find the original copy from the pertubated 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 pertubated copy by applying different MLT-PPDM algorithms to add the noise into original data. By applying LLSEE and Nonlinear error estimation algorithm to calculate how much noise present into original data, do prediction that how much get original data very accurately from this diverse copy. The comparative study between LLSEE and Nonlinear error estimation to decide that nonlinear error estimation gives maximum accuracy. The previous work is limitated only for linear attack means linear function. But proposed result is work on the non-linear attack also means nonlinear function estimation.

Keywords— *Diversity Attack, K-Anonymity, Multi-Level Trust, Non-Linear error estimation, Parallel Generation. Sequence Generation, On Demand Generation, LLSEE.*

*****

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 the private data remain private after mining the process in another way we can say getting valid data mining result learning the underlying data values. There are many research and branches in this area. Most of them analyse and optimize the technologies and algorithms of privacy preserving data mining. Privacy Preserving Data Mining approach limited only single level trust on data miners in this work the data owner generate only one perturbated copy of its data with Uncertainty about individual values before data is released to trusted thirty party. The Perturbated 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 perturbated copy is send to the trusted third party. But now there is multiple numbers of perturbated copies of the same data are send to the different trust level to data miners. If there are large number of trusted stages then the less number of perturbated copies can access. The main goal of MLT-PPDM is to prevent the diversity attack. When data owner sends number of perturbated copy to the trusted third party that time adversary unable to find the original copy from the large number of perturbated copy means the adversary varied from original Copy this is known as the diversity attack. To combine this all perturbated copy and create the original data more accurately which is given by user this is main goal of data miners. The MLT-PPDM works consider the liner attack and non-liner attack. The previous work is limited only for liner attack but current work is applied on non-liner attack to recreate the original data.

II. RELATED WORK

In various organizations the set of data are collected for various mean for their own purpose. The sensitive data can breach through third person and it cannot access by publically so privacy is main an approach. Data Perturbation is a popular technique in PPDM and perturbation-based PPDM approach introduces random perturbation to individual values to preserve privacy before data is published. Data Perturbation consists of two types first one is probability distribution approach and

second is value distortion approach. The probability distribution approach replaces the data with another data from the same distribution or itself also. The value distortion approach change of attribute by adding some additive and multiplicative noise before data is released. To avoid the attack various anonymization techniques are used, in generalization and bucketization there is no clear separation between sensitive and quasi identifier attributes. The most basic form of PPDP consists of four sets:D (Explicit Identifier, Quasi_Identifier, Sensitive_Attributes, Non-Sensitive_Attributes). Data partition also consists of three types 1) some attributes are identifiers that can be uniquely identified for e.g.name, social security number. 2) Some attributes are quasi identifiers (QI) which adversary knows for e.g. Birth date, sex, zip code. 3) Some attributes are sensitive attribute (SA) which are unknown to the adversary for e.g. Disease, salary. Previous solution is limitated only for linear attack the scope of perturbation-based PPDM to data owner sends only of perturbated copy to single-level trust. In existing system anonymization algorithms can be used for column generalization.

- Data swapping is transfer database related attribute between selected record pairs so lower order frequency are preserved and data should be private. These techniques come under data perturbation type.
- K-Anonymity is the to share a person specific data collection without releasing personal information about individual person this is the main goal of ,to use generalization of data and suppression technique are used to protect private sensitive data. the private data may be linked to sets of records of size at least k.
- In previous approach Secure Multiparty Computation (SMC) provides strongest level of privacy. It publish secure data without revealing internal data of particular entity, but this SMC algorithm is very expensive in practice, and impractical for real use. To avoid the high computation cost it use the another solution for avoid SMC.It build a decision tree over horizontal partitioned data & vertically partitioned data algorithm for association rule & frequent pattern mining problems.
- Rule hiding is to transform the database so that all the underlying patterns can be discovered and the sensitive rules are masked. the association rule hiding technique which is perturbation-based is implemented by changing a selected set of 1-values to 0-values
- The distributed data mining (DDM) approach considers data mining models computations and "patterns" extraction at a given (chosen) node by providing only the minimal required information among the other participating nodes

III PROPOSED WORK

A data owner having original data by using MLT-PPDM algorithm likes parallel generation, sequential generation & on demand generation. Parallel generation, sequential generation is also known batch generation algorithm by using this it produced pertubated Copy. Following algorithm is applied on pertubated copy to find out error present in pertubated copy and made predication on to get original data accurately. Original dataset having weight for attributes in privacy evaluation having multiple iteration take place. That time also calculate rotation matrix, random translation, noise level for each pertubated copy and covariance matrix. It starts with randomly generating rotation matrix then swapping the rows. Then Independent Component Analysis (ICA) technique to estimate the independent components (the row vectors definition) of the original dataset *X* from the perturbed data. Then determine the privacy guarantee to the distance-inference attack with the perturbation, to calculate the noise level for each

Algorithm for finding a Nonlinear error estimation perturbation
(Xd×N, w, m)

Input:

- Xd×N: the original dataset,
- **w**: weights for attributes in privacy evaluation,
- m: the number of iterations.

Output:

- Rt: the selected rotation matrix,
- Ψ: the random translation,
- σ2: the noise level,
- p: privacy quality
- calculate the covariance matrix C of X;
- p = 0, and randomly generate the translation Ψ;

for Each iteration **do**

Step I: randomly generate a rotation matrix R;

Step II: swapping the rows of R to get R1which maximizes Min1≤i≤d{ 1/wi (Cov (R1X − X) (i, i)};

Step III: P0 = the privacy guarantee of R1, P1 = 0;

Step IV: if P0 > P **then** generate ˆX with ICA;{(1), (2), . . . , (d)} = argmin{(1),(2),...,(d)} Σ i=1 ΔPDF(Xi, O(i))

Step V: p1 = min1≤k≤d1 1/wk VoD (X k, O(k))

end if

Step VI:if p < min(p0, p1) **then** p = min(p0, p1),Rt = R1

end if

end for

Step VII: p2 =the privacy guarantee to the distance-inference attack with the perturbation $G(X)=Rt\ X+\Psi+\Delta$.

Step VIII: Tune the noise level σ2, so that $P2 \geq P$ if $P < \varphi$ or $P2 > \varphi$ if $P < \varphi$.

IV SYSTEM IMPLEMENTATION

A data owner having original data to preserve privacy by adding the noise into original data. It use real dataset CENSUS which is commonly used in the privacy 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 10^5 tuples and conducts the experiments on the Age and Income attributes. After selecting data, partition it in to number of column. Then applying different types of MLT-PPDM algorithm like parallel generation, sequence generation, on demand generation to produced different pertubated copy. Pertubated copy means number of changes made in the original data, to Produced copy in the form batch and sequence also, that produced pertubated copy on M different trust level it requires a data owner to foresee all possible trust levels a priori. In on demand produced copy according to user demand hence it gives maximum flexibility. Then by applying linear least square estimation technique on that pertubated copy it calculate how much noise added into the original data and make prediction about how much original data get from above technique, it is linear estimation.

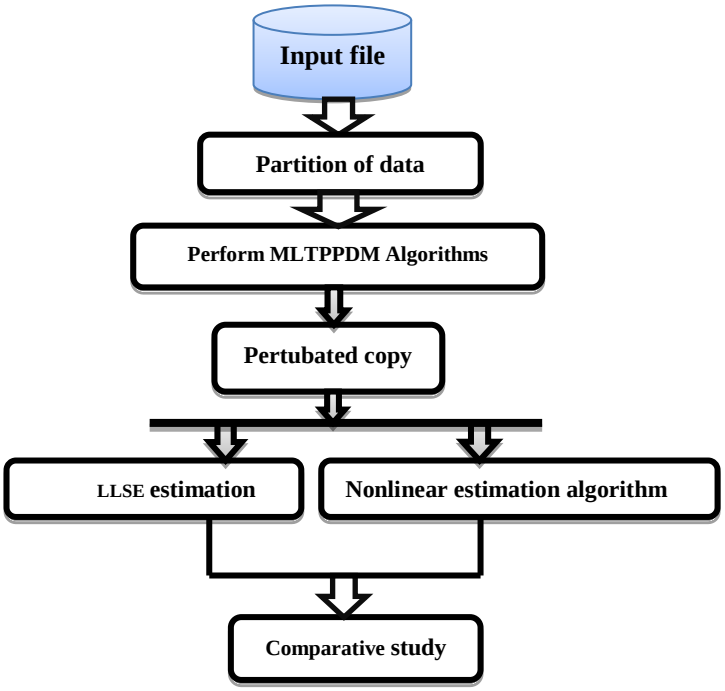

Fig.1.System Architecture

In fig.1. Shows system architecture having produced pertubated copy by applying number of MLTPPDM algorithms that time LLSEE estimation first find out error between this pertubated coy and by using nonlinear error estimation algorithm is applied on pertubated copy to calculate also error into pertubated copy then make comparison between these two algorithms. The value of LLSE estimation is larger than value of nonlinear estimation algorithm means minimums error found in nonlinear estimation algorithm ,hence it generate large original data accurately this is the main goal of nonlinear estimation algorithm.

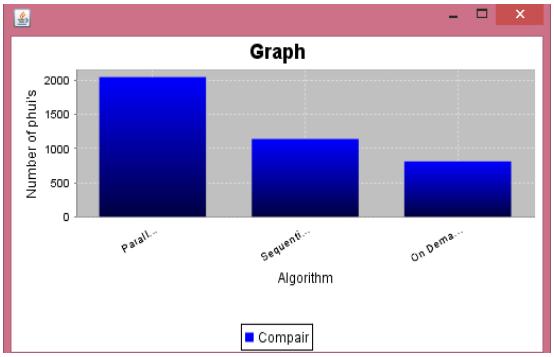

Fig.2.Time complexity

Fig.2.shows the time complexity between MLTPPDM algorithms like parallel generation, sequence generation and on demand generation. parallel generation require lager time s compare to other algorithms.

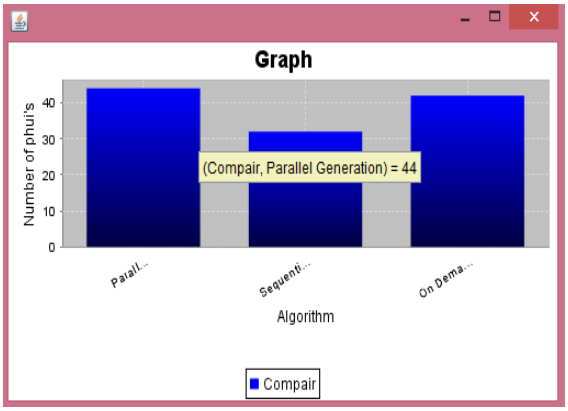

Fig.3.Space complexity

Fig.3.shows space complexity between MLTPPDM algorithms having minimum space is required sequence generation as compared to other algorithm.

Comparative analysis

Parameter	LLSE estimation	Nonlinear error estimation
Noise Level	5.242	1.076

Table 1.Noise level analysis

Result analysis

But in proposed work we use nonlinear estimation algorithm to calculate how much noise added into original data and make prediction about how much original data get from proposed

 ISSN: 2321-8169
1875 – 1878

technique. Again comparative study between LLSEE and nonlinear estimation algorithm to calculate which technique gives minimum noise and to reconstruct maximum original data accurately, hence nonlinear estimation algorithm gives maximum accuracy to calculating noise into that.

Fig.4.shows comparative analysis between LLSE and nonlinear estimation algorithm which gives error value that is LLSE gives lager error value larger than nonlinear error estimation; hence if error is less it gives large accurate data.

Nonlinear estimation algorithm calculate minimum error and if it is below than threshold value that consider as normal value. In that nonlinear function is used to reconstruct original data very accurately. In above table shows comparison between LLSE and nonlinear estimation depend on the noise level.

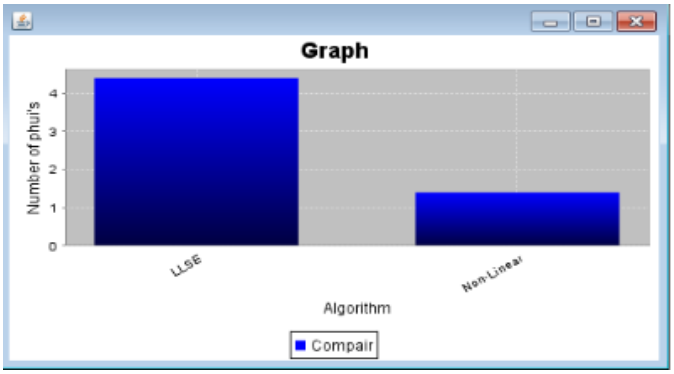

Fig.4.Comparative study

V CONCLUSIONS

It expand the scope of data perturbation of PPDM to multilevel trust means it produced copy at different trust level. The main goal of MLT-PPDM is to combining of all pertubated copy at different trust levels to reconstruct the original data very accurately. Using LLSE & nonlinear estimation calculate error between pertubated copies, hence nonlinear estimation calculate minimum error to reconstruct original data very accurately this is main challenges of MLTPPDM

VI 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] 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.