IJETCR Open Access

ISSN: 2348 - 2117

International Journal of Engineering Technology and Computer Research (IJETCR)
Available Online at www.ijetcr.org
Volume 2, Issue 4, Page No. 15-18

A SURVEY PAPER ON IMPROVED METHOD FOR PRIVACY PRESERVING DATA MINING CONSIDERING LINEAR AND NON- LINEAR ATTACK

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

[1]*Computer Engineering Department, Pune University, JSPM's Imperial College Engineering & Research Wagholi Pune, India*
bhorde.vaishali1@gmail.com

[2]*Computer Engineering Department, Pune University, JSPM's Imperial College Engineering & Research Wagholi Pune, India*
Sonalitidke11@gmail.com

ABSTRACT
Privacy Preserving Data Mining (PPDM) is used to extract knowledge from dataset and a preserve the privacy before data can be release. The study of perturbation based PPDM approaches introduces random perturbation is the 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. Malicious data miners have access the different pertubated Copy of the same data through the various mean to combine this all the diverse copy to get original data very accurately this goal of with respect to privacy. In this work a user produce large number of pertubated copies of its data for random trust level on demand. Hence the user having maximum flexibility. The previous work is limitated only for linear attack. But proposed result is work on the non-linear attack also. In previous anonymization techniques generalization and bucketization related to the privacy of individual information to the adversary. The generalization involves loss of information and bucketization approach does not protection from membership disclosure, but in proposed approach slicing with tuple grouping algorithm partitioned from membership disclosure and also can handle large amount of data.

Key words: Diversity Attack, K-Anonymity, Multi-Level Trust, Non-Linear Attack, Parallel Generation. Introduction

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. To compare between the estimated value and original value to create data very accurately. 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 apply 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 slicing method data can be partition both vertically and horizontally. 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.

- In existing system losses data easily. In existing system generalization and bucketization there is no clear separation between sensitive and quasi identifier attributes. In existing system cannot handle large amount data
- In this paper introduce a novel data anonymization technique called slicing to improve the current state of the art. In this paper there is clear separation between sensitive and quasi identifier attributes. In this paper it used to attribute protection prevent membership disclosure. In this paper conduct extensive workload experiments. The results confirm that slicing preserves much better data utility than generalization.
- In previous approach k-Anonymization techniques there is possibility of loss of information The main idea is to suppress or generalize some of the public data so that each of the records becomes indistinguishable from at least k -1 other records, when projected on the subset of public attributes. Consequently, 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.
- Another category is partial information hiding approach use to improve the performance to hide the data for preserving data. Again it divide into number of category like k-anonymity, retention replacement, data perturbation approach

III. SYSTEM IMPLEMENTATION

Implementation is the stage of the project when the theoretical design is turned out into a working system. Thus it can be considered to be the most critical stage in achieving a successful new system and in giving the user, confidence that the new system will work and be effective. Investigation of the existing system and it's constraints on implementation, designing of methods to achieve changeover and evaluation of changeover methods. A data owner having own data. They could provide authentication with their respective data. They have to register into database. User can view their original data's. Whatever they are stored into database In this module a data having the multilevel trust and random rotation perturbation. A data owner having original data and now the add the noise by using sequence by the random rotation perpetuation in order to the sequenced Algorithm. Admin also can view the original data's. Whatever stored in the whole admin can

login and view the original data's. The module 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. Create the noise adding the null values.

System Architecture

A data owner having original data by using random perturbation & 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.

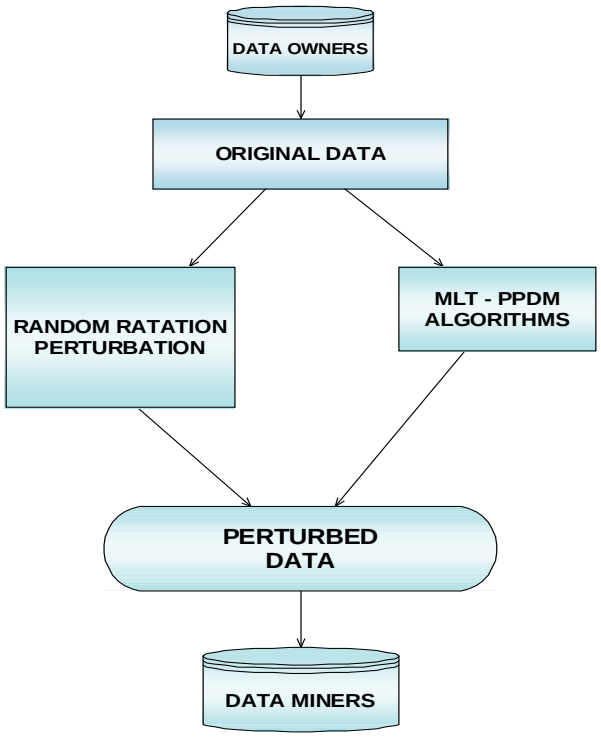

Figure 1: Mechanism of Advanced MLT-PPDM

The data perturbation means number of changes is made in the original data. 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. In additive noise technique is used to add the random noise into original data. Suppose X is original data and random noise is Z then it obtain pertubated copy as Y=X+Z. The data is divided into number of columns generate number of pertubated copy by adding noise into original data.

IV. PROPOSED WORK

To compare between estimated copy and original copy to reconstruct data very accurately, hence LLSEE method is used. Linear attack is considered as the linear function is used into algorithms and nonlinear attack consider as nonlinear function is used into algorithm. In implementation part first data is divided into number of column then for preserving the privacy create the number of trust levels by adding noise into that original data or by using different MLTPPDM algorithms it create multiple trust level. Then we calculate the computation cost means how much reconstruct original data from pertubated copy for linear attack and same calculate computation cost for nonlinear using another type of algorithm. Compare between these two algorithms and decide which is efficient. If the noise level is within the minimum threshold value, then adversary cannot find original data, after the addition of noise it again compared continue this process up to the noise level goes within the minimum threshold value. Proposed system overcome problem of both the linear and non-linear attacks from the hackers. In the result analysis it perform graph like non-linear technique gives a higher security level compared to the other existing techniques, additive and multiplicative perturbation. Privacy can be preserved by making use of Nonlinear techniques in proposed system is

- We have to first to load the data on server.
- Extract the sensitive data from database.
- To add the noise into the original data based on different trust levels.
- This different trust level given to the data miners for preserving privacy before published data.
- By applying nonlinear estimation technique compare the estimated copy and original copy.
- Depend on that comparison decide that which algorithm gives maximum computation cost.
- In non-linear techniques gives maximum computation cost because in that input values are constant but output values goes varies.
- To analysis the result linear technique gives straight graph, and nonlinear technique gives curve graph, this is main difference linear and nonlinear attack.

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. The paper approach is allows data owner to produced pertubated copies at different trust levels on their demand. This function offers the data owner having maximum flexibility.

VI. ACKNOWLEDGMENT

I wish to express my sincere thanks and deep gratitude towards Dr. S. V. Admane [Principal ICOER] and my guide Prof. S.M.Tidke for her guidance, valuable suggestions and constant encouragement in all phases. I am highly for indebted to her help in solving my difficulties which came across whole Paper work. Finally I extend my sincere thanks to respected Head of the department Dr. S.B.Patil [P.G.Co-ordinator] and all the staff members for their kind support and encouragement for this paper. Last but not the least, I wish to thank my Mother for her unconditional love and support.

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.