ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

Accumulative Privacy Preserving Data Mining Using Gaussian Noise Data Perturbation at Multi Level Trust

G.Mareeswari[1], V.Anusuya[2]

ME, Department of CSE, PSR Engineering College, Sivakasi, Tamilnadu, India[1]

Assistant Professor (Sl. Gr), Department of CSE, PSR Engineering College, Sivakasi, Tamilnadu, India[2]

ABSTRACT: Generally Data Mining develops the exact models about the collected data. Data perturbation, a widely employed and accepted Privacy Preserving Data Mining (PPDM) approach add random noise to original data , that prevent data miner to publish the accurate information about original data that is not allowed by data owner. Under the single level trust a data owner generate only one perturbed copy of its data with affixed amount of uncertainty. In this Project, the aim is to enlarge the scope of perturbation-based PPDM to Multilevel Trust (MLT-PPDM). In this system, different perturbed copies of same data are available to data miner at different trust level. If data miner is more trusted means, it can access the minor perturbed copy of the data. In case of malevolent data miner, may have access to differently perturbed copies of the same data and may combine these different copies to collaboratively induce more information about the original data that the data owner does not aim to release; this is the "DIVERSITY ATTACK". Inhibiting such diversity attacks is the major provocation of providing MLT-PPDM services. In this project, the scope is to provide the additive perturbation approach where random Gaussian noise is added to the original data with arbitrary distribution, so the data miner will have no diversity gain and provide a systematic solution. This solution allows a data owner to generate perturbed copies of its data on demand at arbitrary trust levels.

KEYWORDS: Privacy Preserving Data Mining, Multilevel Trust, Perturbation, Diversity attack

I. INTRODUCTION

Data mining, the extraction of interesting patterns or knowledge from huge amount of data stored either database, Data warehouse other information repositories. Data mining is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions.

1.1 PRIVACY PRESERVING DATA MINING

The main objective in privacy preserving data mining is to develop algorithms for modifying the original data in some way, so that the private data and private knowledge remain private even after the mining process. A number of techniques such as Trust Third Party, Data perturbation technique, Secure Multiparty Computation and game theoretic approach, have been suggested in recent years in order to perform privacy preserving data mining. The main consideration of PPDM is twofold. First, sensitive raw data like identifiers, names, addresses and so on, should be modified or trimmed out from the original database, in order for the recipient of the data not to be able to compromise another person's privacy. Second, sensitive knowledge which can be mined from a database by using data mining algorithms should also be excluded, because such knowledge can equally well compromise data privacy.

Data Perturbation is a widely employed and accepted Privacy Preserving Data Mining(PPDM) approach. It is a category of data modification approaches that protect the sensitive data contained in a dataset by modifying a carefully selected portion of attribute-values pairs of its transactions. Data perturbation includes a wide variety of techniques including (but not limited to): additive, multiplicative, matrix multiplicative, k-anonymization, micro-aggregation, categorical data perturbation, data swapping, re sampling, data shuffling. In this project additive perturbation is used for the purpose of Privacy Preserving Data Mining.

II. PROBLEM STATEMENT

To expand the scope of perturbation based PPDM to Multi-Level Trust, by relaxing the implicit assumption of single trust levels. To enable the MLT-PPDM (Multi Level trust-Privacy Preserving Data Mining) services to find whether the diversity attack is present or not.

III. PROPOSED SYSTEM

In proposed System describes new dimension of Multilevel Trust (MLT) poses new challenges for perturbation-based PPDM. In contrast to the single-level trust scenario where only one perturbed copy is released, now multiple differently perturbed copies of the same data are available to data miners at different trusted levels. The more trusted a data miner is, the less perturbed copy it can access; it may also have access to the perturbed copies available at lower trust levels. Moreover, a 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 reconstruction of the original data than what is allowed by the data owner.

The following Contributions are made in the Project

- The scope of this project is to expand the perturbation based PPDM to Multilevel trust PPDM, that provide flexibility for the data owners to generate differently perturbed copies of its data for different trust levels.
- In MLT PPDM, there is the possibility of Diversity attack, by combining the multiple perturbed copies data miner able to perform diversity attack to reconstruct the original data. Defending such attack is the major challenge of this project.
- This challenge is addressed by properly correlating perturbation across copies at different trust levels. In this paper, the work is to propose several algorithms to provide the solution that is robust against the diversity attacks.
- The solution allows data owners to generate perturbed copies of their data at arbitrary trust levels on-demand. This property offers data owners maximum flexibility.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

IV. SYSTEM DESIGN

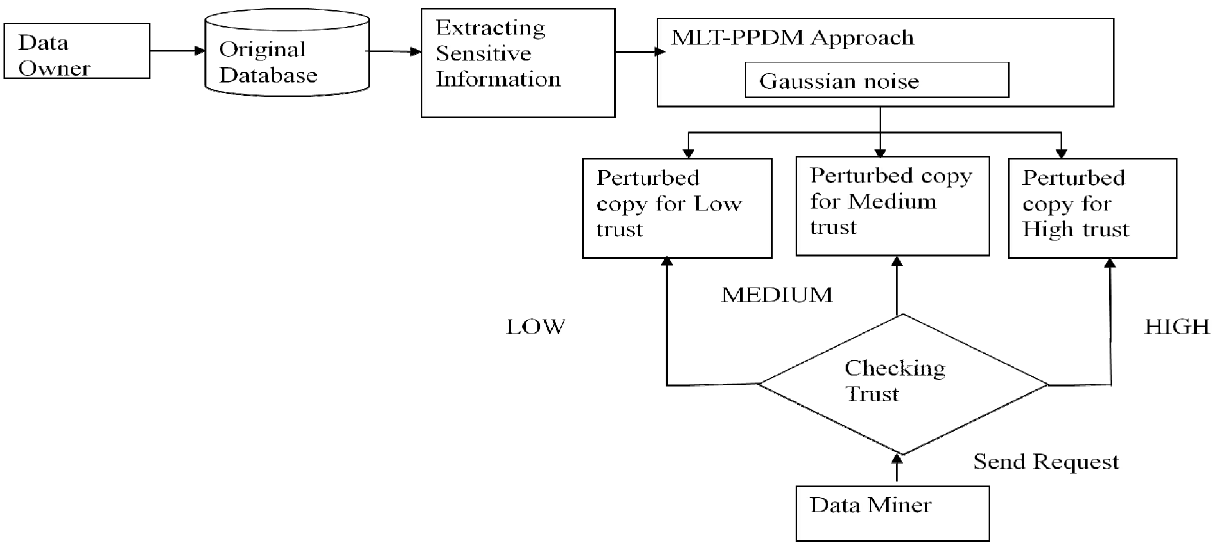

Fig 4.1 DATA FLOW DIAGRAM

V. SYSTEM IMPLEMENTATION

The proposed system consists of four main modules. They are

- Data Owner
- Admin
  - Assign Trust Level
- MLT PPDM Technique
  - Batch Generation.
  - On-Demand Generation.
- Performance Test.

a. DATA OWNERS

Data Owners are Users, whose personal or private information's are preserve. They provide their information to admin and they register the person details. In this application the data owner who provide their information is patients. The persons in the medical organization such as Doctors, Staff and Medical Representatives also provide their details to the admin. Admin Register their information in the separate database so the employees are also here referred as Data Owner.

b. ADMIN

Admin also can view the original data's. Admin is responsible for entering the patients and others detail. Doctors examine the patients only after the patients registration is done by the admin. Admin is also responsible for updating the patients details after the patient examine by the doctors.

Assigning Trust levels

In this Module having Data Miner Request and Trust level. Data Miner has specified trust level. After Getting Request from Data Miner, checking trust level. Based on the trust level perturbed copy is send.

c. MLT PPDM Technique

The original data's saved to the database are recollected and noise is add to the original data for Data Perturbation based on the trust level.

GAUSSIAN NOISE

Let G_1 through G_L be L Gaussian random variables. It is said to be jointly Gaussian if and only if each of them is a linear combination of multiple independent Gaussian random variables.

Its probability density function is as follows

$$F_G(g) = \frac{1}{\sqrt{(2\pi)^L \det(K_Z)}} e^{-(g-\mu_x)^T K_Z^{-1}(g-\mu_x)/2}$$

Using this Probability density function noise Z_1 to Z_M generated.

BATCH GENERATION

In the first scenario, the data owner determines the M trust levels a priori, and generates M perturbed copies of the data in one batch. In this case, all trust levels are predefined and when generating the noise. Refer this scenario as the batch generation .Admin determines the M trust level a priori. Generate M perturbed copies of data in the batch.

$$Y_1 = X + Z_1$$
$$Y_2 = X + Z_2$$

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

ON DEMAND GENERATION

In the second scenario as opposed to the batch generation, new perturbed copies are introduced on demand. Since the requests may be arbitrary, the trust levels corresponding to the new copies would be arbitrary as well. The new copies can be either lower or higher than the existing trust levels. Refer this scenario as on-demand generation. Achieving the privacy goal in this scenario will give data owners the maximum flexibility in providing MLT-PPDM services.

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

d. PERFORMANCE TEST

In case of the malicious data miners can access all the M perturbed copies. This represents the most severe attack scenario where data miners jointly estimate original value using all the available M perturbed copies.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) **Vol.2, Special Issue 1, March 2014**

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

Since the perturbed copies are released one by one, the number of the available perturbed copies also increases one by one. The performance test is done by the family of linear reconstruction methods, where estimates only be the linear function of perturbed copy. Linear Least Squares Error (LLSE) estimation has the minimum square errors between the estimated values and the original values.

VI. RESULT AND DISCUSSION

In the batch generation approach the attempt is generate the perturbed copies independently. The added noise is not only independent of the original data, but also independent of each other. In the on-demand generation, the perturbed copy generated for second trust level is depends upon the perturbed copy of the first trust level. The Linear Least Square Error estimation shows that the difference between the estimated value and the original value is maximum for on-demand generation when compared to the Batch Generation. So the diversity attack is prevented in the On-Demand Generation.

Fig 6.1 Home Page Of Medical Organization

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

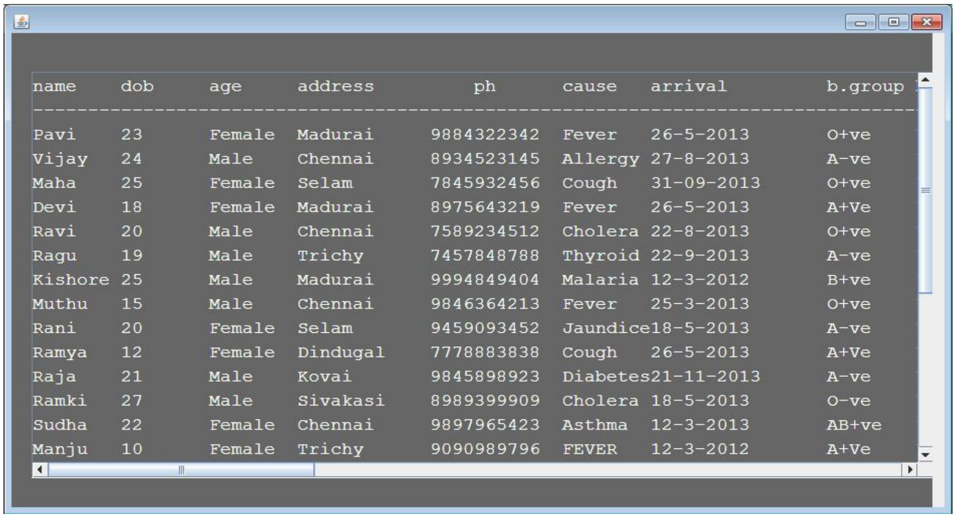

Fig 6.2 Original Data

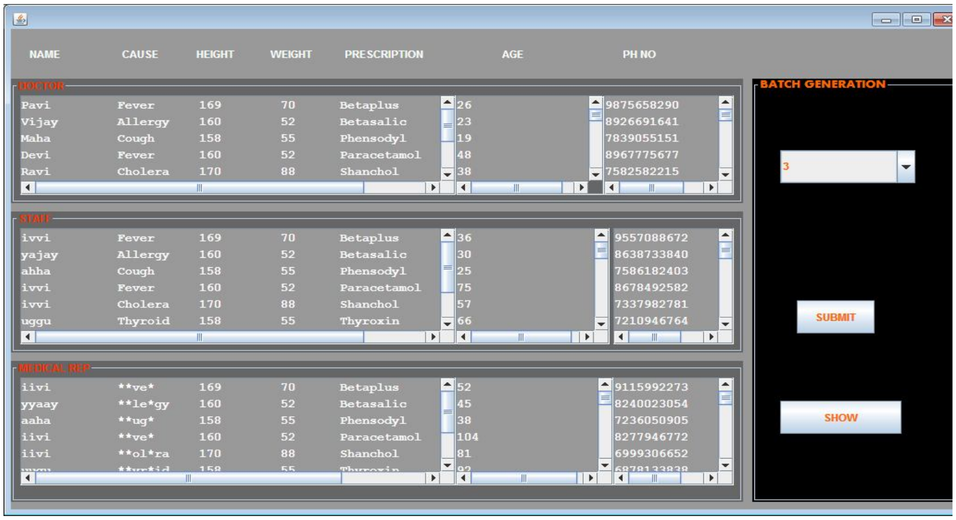

Fig 6.3 Perturbed Copy for Batch Generation

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

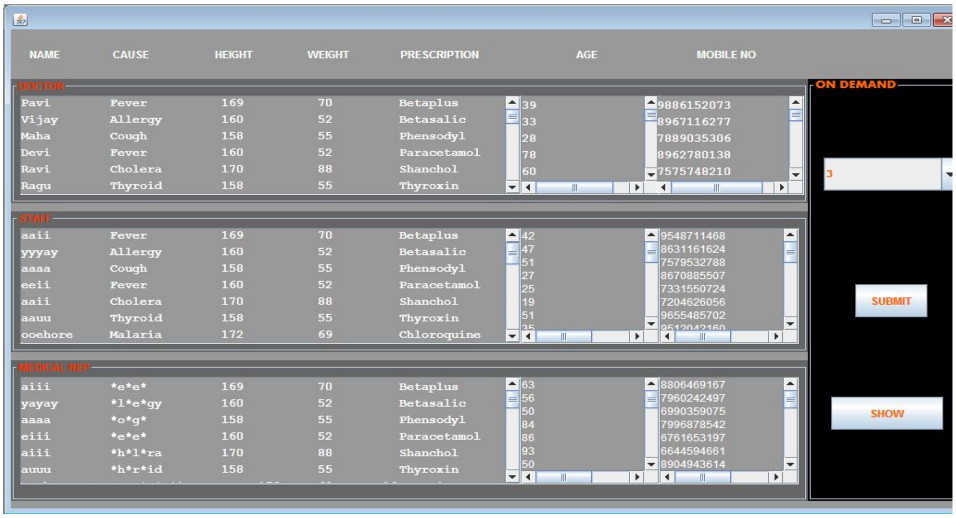

Fig 6.4 Perturbed Copy For On-Demand Generation

Fig 6.5 Performance Measure

VII. CONCLUSION

In this work, MLT-PPDM allows data owners to generate differently perturbed copies of its data for different trust levels. The major challenge is to prevent the diversity attack that is done by the proposed On-demand generation approach. But the approach is defending only against the linear attack. More powerful adversaries may apply nonlinear techniques to derive original data and recover more information. The future work is to study the MLT-PPDM problem under the adversarial model.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

REFERENCES

[1] X. Xiao, Y. Tao, and M. Chen, "Optimal Random Perturbation at Multiple Privacy Levels," Proc. Int'l Conf. Very Large Data Bases, 2009.
[2] G. Wang, Z. Zhu, W. Du, and Z. Teng, "Inference Analysis in Privacy-Preserving Data Re-Publishing," Proc. Int'l Conf. Data Mining, 2008.
[3] 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.
[4] J. Byun, Y. Sohn, E. Bertino, and N. Li, "Secure Anonymization for Incremental Datasets," Proc. Third VLDB Workshop Secure Data Management, 2006.
[5] Bertino.E, Ooi .B.C, Y. Yang, and R.H. Deng, "Privacy and Ownership Preserving of Outsourced Medical Data," Proc. 21st Int'l Conf. Data Eng. (ICDE), 2005.
[6] K. Chen and L. Liu, "Privacy Preserving Data Classification with Rotation Perturbation," Proc. IEEE Fifth Int'l Conf. Data Mining, 2005.
[7] Z. Huang, W. Du, and B. Chen, "Deriving Private Information From Randomized Data," Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD), 2005.
[8] Aggarwal .C.C and Yu .P.S, (2004)"A Condensation Approach to Privacy Preserving Data Mining," Proc. Int'l Conf. Extending Database Technology (EDBT).
[9] H. Kargupta, S. Datta, Q. Wang, and K. Sivakumar, "On the Privacy Preserving Properties of Random Data Perturbation Techniques," Proc. IEEE Third Int'l Conf. Data Mining, 2003.
[10] A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke, "Privacy Preserving Mining of Association Rules," Proc. ACM SIGKDD Int'lcConf. Knowledge Discovery and Data Mining, 2002.
[11] Yaping Li, Minghua Chen, Qiwei Li, and Wei Zhang "Enabling Multilevel Trust in Privacy Preserving Data Mining" IEEE Transaction on knowledge and data Engineering vol.24,No.9, Sep 2012.