Multilevel Data Aggregated Using Privacy Preserving Data mining

V.Nirupa
Department of Computer Science and Engineering
Madanapalle, Andhra Pradesh, India

M.V.Jaganadha Reddy
Department of Computer Science and Engineering
Associate Professor, Department of Computer Science and Engineering, Madanapalle,
Andhra Pradesh, India

R.Ushasree
Department of Computer Science and Engineering,
Andhra Pradesh, India

Abstract - We propose a novel reconstruction procedure to accurately estimate the distribution of original data values. By using these reconstructed distributions, we are able to build classifiers whose accuracy is comparable to the accuracy of classifiers built with the original data. The primary task in data mining is the development of models about aggregated data we studied perturbation-based PPDM approach which introduces Random perturbation to individual values to preserve privacy before data is published. Previous solutions of this approach are limited in their tacit assumption of single-level trust on data miners. With this work, we can relax this assumption and expand the scope of Perturbation based PPDM to Multilevel Trust (MLT-PPDM). MLT-PPDM allows data owners to generate differently perturbed copies of its data for different trust levels.We verify our claim and demonstrate the effectiveness of our solution through numerical evaluation. Last but not the least, our solution allows data owners to generate perturbed copies of its data at arbitrary trust levels on-demand.

Keywords: Privacy preserving data mining, Secure Multi Party Computation, Multilevel trust, random perturbation.

I. INTRODUCTION

The 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. We address this challenge by properly correlating noise across copies at different trust levels. We prove that if we design the noise covariance matrix to have corner-wave property, then data miners will have no diversity gain in their joint reconstruction of the original data. This property offers the data owner maximum flexibility. We believe that multilevel trust privacy preserving data mining can find many applications. Our work takes the initial step to enable MLT-PPDM services. Many interesting and important directions are worth exploring. For example, it is not clear how to expand the scope of other approaches in the area of partial information hiding, such as random rotation based data perturbation, and retention replacement, to multilevel trust. As with most existing work on perturbation based PPDM, our work is limited in the sense that it considers only linear attacks. More powerful adversaries may apply on linear techniques to derive original data and recover more information. Studying the MLT-PPDM problem is a very interesting task that has been provided.

II. BACKGROUND

2.1. Original Data:

We conduct extensive workload experiments. Our results confirm that slicing preserves much better data utility than generalization. In workloads involving the sensitive attribute, slicing is also more effective than bucketization. In some classification experiments, slicing shows better performance than using the original data.

2.2. Jointly Gaussian:

Generalized Data, in order to perform data analysis or data mining tasks on the generalized table, the data analyst has to make the uniform distribution assumption that every value in a generalized interval/set is equally possible, as no other distribution assumption can be justified. This significantly reduces the data utility of the generalized data.

2.3. Additive Perturbation:

We show the effectiveness of slicing in membership disclosure protection. For this purpose, we count the number of fake tuples in the sliced data. We also compare the number of matching buckets for original tuples and that for fake tuples. Our experiment results show that bucketization does not prevent membership disclosure as almost every tuple is uniquely identifiable in the bucketized data.

2.4. Linear Least Squares Error Estimation:

We observe that this multi set based generalization is equivalent to a trivial slicing scheme where each column contains exactly one attribute, because both approaches preserve the exact values in each attribute but break the association between them within one bucket.

III. LITERATURE SURVEY

3.1. PRIVACY PRESERVING

There has been some research considering how much information can be inferred, calculated or revealed from the data made available through data mining process, and how to minimize the leakage of information In [1], data perturbation techniques are used to protect individual privacy for classification, by adding random values from a normal/Gaussian distribution of mean 0 to the actual data values. One problem with this approach is the existing tradeoff between the privacy and the accuracy of the results. More recently, data perturbation has been applied to Boolean association rules An interesting feature of this work is a flexible definition of privacy; e.g., the ability to correctly guess a value of `1' from the perturbed data can be considered a greater threat to privacy than correctly learning a `0'.

Three possible definitions of privacy[2]:

- Privacy as the right of a person to determine which personal information about Himself/herself may be communicated to others.
- Privacy as the control over access to information about oneself.
- Privacy as limited access to a person and to all the features related to the person.

From the above three definitions we know that "The right of an individual to be secure from unauthorized disclosure of information about oneself that is contained in an electronic repository". Performing a final tuning of the definition, we consider privacy as "The right of an entity to be secure from unauthorized disclosure of sensible information that are contained in an electronic repository or that can be derived as aggregate and complex information from data stored in an electronic repository".

3.2.SECURE MULTI-PARTY COMPUTATION [SMC]:

Secure Multi-Party Computation [SMC][3] considers the problem of evaluating a function of two or more parties' secret inputs, such that no party learns anything but the designated output of the function. Concretely, we assume we have inputs x1, …,xn, where party i owns xii, and we want to compute function f[x1,…,xn] = [y1,…,yn] such that party I gets yi and nothing more than that. Example:- As an example we may consider Yao' smillionaire's problem: two millionaires meet in the street and want to find out who is richer without having to reveal their actual fortune to each other. The function computed in this case is a simple comparison between two numbers. If the result is that the first millionaire is richer, then he knows that, but this should be all information he learns about the other guy.

3.3 Multilevel trust, random perturbation:

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. In some scenarios, such as the motivating example we give at the beginning of Section 1, 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 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 vector with mean X and covariance K_x as defined. These M copies can be generated in various fashions. They can be jointly generated all at once. 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 owner's 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. For many real applications, knowing only the mean and covariance may not be sufficient to apply data mining techniques, such as clustering, principal component

analysis, and classification [4]. By using random perturbation 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 .Let Y=[Y_1, Y_2, Y_3, Y_4, . . . Y_n] be the vector of all perturbed copies. Let Z= [Z_1, Z_2, Z_3, Z_4, … Zn] be the vector of noise. Let H be an identity matrix.

IV. RESULTS DISCUSSION

Data reduction techniques suffer from homogeneity attack [5]. That is specially to mention that sensitive data lacks diversities in values. Also if adversary has additional background knowledge then he can infer sensitive data pertaining to individuals. During the application of anonymization techniques two important assumptions hold true. First, it may be very hard for the owner of a database to determine which attributes are available in external tables. Second, a specific type of attack is assumed, but in real scenarios there is no reason why an attacker would not try other methods of attacks.

Perturbation techniques do apply independent treatment of different attributes, but the main disadvantage being reconstructing original data values back from the published data. Perturbation techniques also become vulnerable in Known Input-Output Attack.In this case, the attacker knows some linearly independent collection of records, and their Corresponding perturbed version. In such cases, linear algebra techniques can be used to reverse-engineer the nature of the privacy preserving transformation. Also for **Known Sample Attack,** perturbation techniques are not found to be satisfactory. Here, the attacker has a collection of independent data samples from the same distribution from which the original data was drawn. In such cases, principal component analysis techniques can be used in order to reconstruct the behavior of the original data. Data swapping technique does not follow the general principle in randomization which allows the value of a record be perturbed independently of the other records. Therefore, this technique must be used with other techniques.

Let a q*-block be a set of tuples such that its non-sensitive values generalize to q*-. A q*-block is l-diverse if it contains l “well represented” values for the sensitive attribute S. A table is l-diverse, if every q*- block in it is l-diverse[6]. Finally, we introduce a universal measure of data privacy level, proposed by Bertino et al. in [2]. The basic concept used by this measure is information entropy, which is defined by Shannon Let X be a random variable which takes on a finite set of values according to a probability distribution $p(x)$. Then, the entropy of this probability distribution is defined as follows:

$$h(X) = -\text{a}p(x)\ \log 2(p(x)) \quad \text{or, in the continuous case:}$$

$$h(X) = -\text{Z}f(x)\ \log 2(f(x))dx\ .$$

In this paper, we have conducted experiments on three real-world data sets. These data sets may have their limitations, but still our experimental results indicate that perturbation methods do not work well when applied to some real-world data sets. This implies that for some data distributions, we should not try to solve the hard reconstruction problem as an intermediate step. Although, at this point we can not simply say that the reconstruction method is not applicable for real-world data sets. Clearly, caution needs to be exercised before applying reconstruction phase in practice.

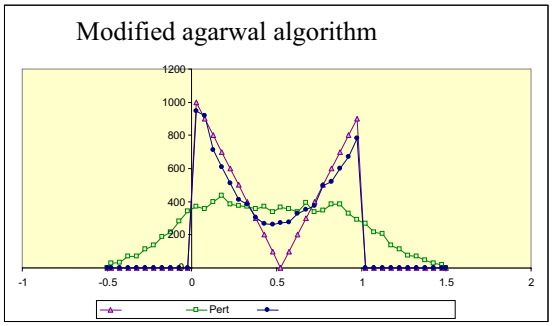
V. IMPLEMENTATION

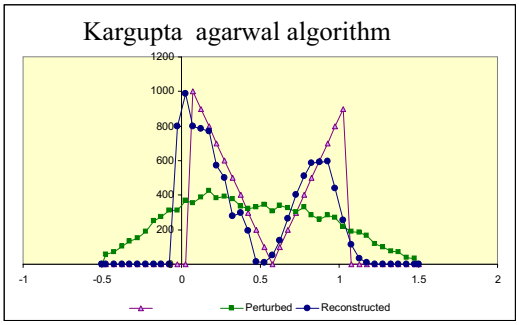
We design three sets of experiments. The first set is used to show that the discussed classifiers are invariant to rotations. The second set shows privacy quality of the good rotation perturbation. The third one compares the privacy quality between the condensation approach and the random rotation approach. Due to the space limitation, we report some results of the later two sets of experiments. The datasets are all from UCI machine learning database. Algorithm combining the local optimization and the test for ICA attacks, [8] we develop a random iterative algorithm to find a better rotation in terms of privacy quality. The algorithm runs in a given number of iterations. In each iteration, it randomly generates a rotation matrix. Local optimization through row-swapping rows is applied to find a better rotation matrix, which is then tested by the ICA reconstruction.. The rotation matrix is accepted as the best perturbation yet if it provides highest P among the previous perturbations.

VI. RESULTS

Here the differently perturbed copies are generated with different trust levels. And intruders cannot reconstruct the original copy of data by knowing the perturbed copies.According to the user the number perturbed copies can generated. Here data owner have the maximum flexibility, so the data owner can release what he intended to release. Here privacy is preserved.

We can obtain the relationship between the estimation error and three parameters, namely the privacy assurance metric, the dimension size N of transition matrix, and the total number n of data records. Randomization and perturbation are two very important techniques in privacy preserving data mining. Loss of information versus preservation of privacy is always a trade off. Furthermore, an approach that uses random matrix properties has recently posed a challenge to the perturbation-based techniques. The question is, can perturbation based techniques still protect privacy? In order to find the answer to this question, we scrutinize two different approaches; one proposed by Agarwal et. al using Bayes density functions and the other proposed by Kargupta et. al using random matrix. We set up simulation experiments to study these two approaches. The question is, besides the properties of random noise what else do we know about reconstructing the original distribution? First we compared the assumptions and preconditions of the two approaches. Then, by using different conditions, we have obtained some interesting results and have made some observations. We propose a modified version of Agarwal et. al's algorithm, which reconstructs the original distribution from the perturbed distribution rather than using the perturbed data. Furthermore, under the same conditions, and by using the random matrix filter approach we failed to obtain the original distribution. We give a hypothesis to explain this observation. Based on this hypothesis, we propose an adaptable perturbation model, which accounts for the diversity of information sensitivity. The adaptable perturbation model presented here has a parameter to adjust the perturbation level to best fit the different privacy concerns.

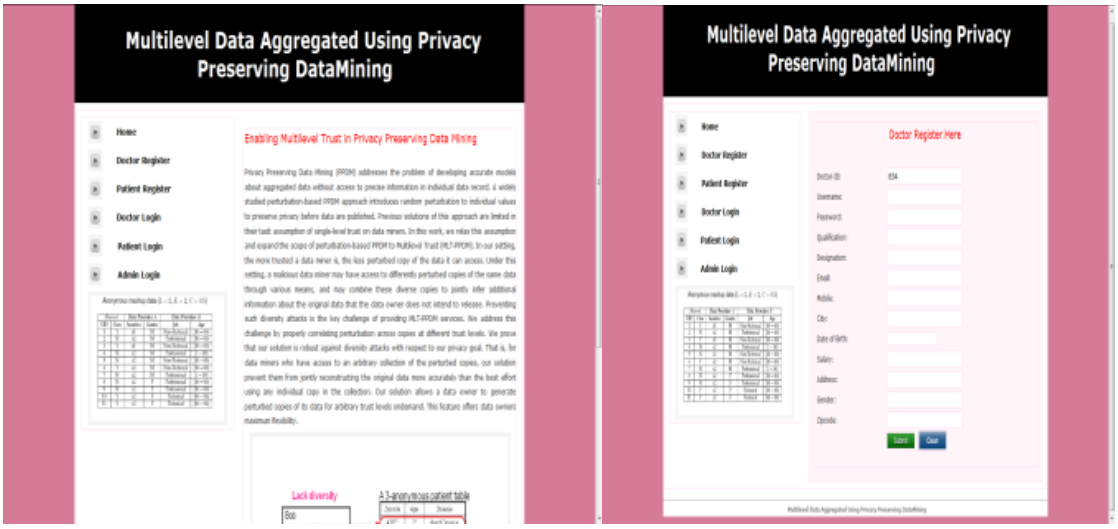




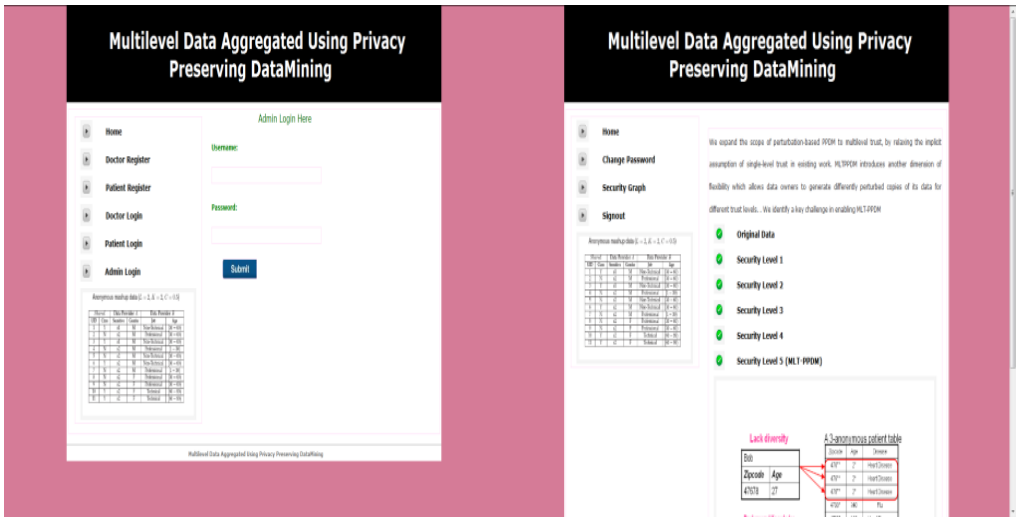
The above defined results are from the Agarwal and Kargupta random matrix results defines the results and the approaches.

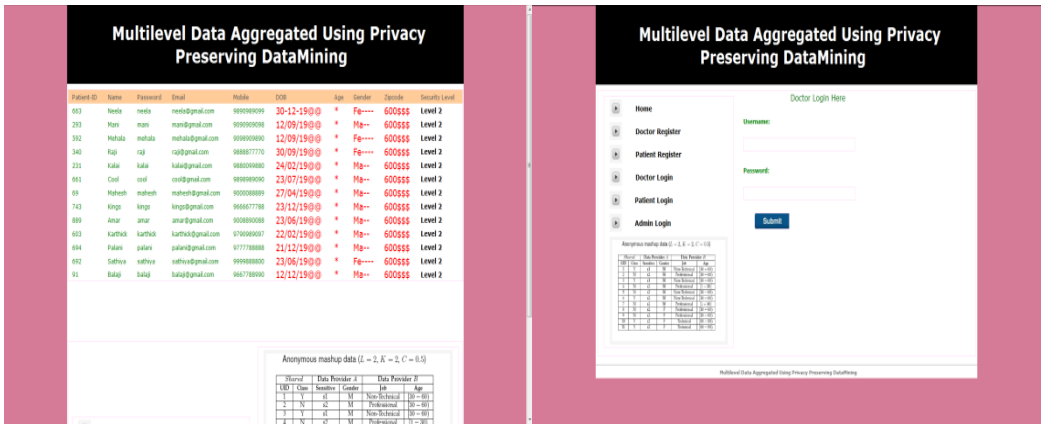
VII. EXPERIMENTAL RESULTS

The superiority of our scheme over the scheme that simply adds independent noise, and how data miner's knowledge affects the power of LLSE-based diversity attacks.



We assume that data miners can access all the M perturbed copies. This setting represents the most severe attack scenario where data miners jointly estimate X using all the available M perturbed copies. Since the perturbed copies are released one by one, the number of the available perturbed copies also increases one by one.





Here the differently perturbed copies are generated with different trust levels. And intruders cannot reconstruct the original copy of data by knowing the perturbed copies. Different group of people will access different copies. They will get data what the data owner intended to release.

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.

We address this challenge by properly correlating noise across copies at different trust levels. We prove that if we design the noise covariance matrix to have corner-wave property, then data miners will have no diversity gain in their joint reconstruction of the original data. We verify our claim and demonstrate the effectiveness of our solution through numerical evaluation.

Last but not the least, our solution allows data owners to generate perturbed copies of its data at arbitrary trust levels on-demand. This property offers the data owner maximum flexibility.

We believe that multilevel trust privacy preserving data mining can find many applications. Our work takes the initial step to enable MLT-PPDM services.

Many interesting and important directions are worth exploring. For example, it is not clear how to expand the scope of other approaches in the area of partial information hiding, such as random rotation-based data perturbation, k anonymity, and retention replacement, to multilevel trust. It is also of great interest to extend our approach to handle evolving data streams.

As with most existing work on perturbation-based PPDM, our work is limited in the sense that it considers only linear attacks. More powerful adversaries may apply nonlinear techniques to derive original data and recover more information. Studying the MLT-PPDM problem under this adversarial model is an interesting future direction.

REFERENCES

[1] Privacy preserving Data Mining Algorithms without the use of Secure Computation or Perturbation Alex Gurevich Ehud Gudes Department of Computer Science Department of Computer Science Ben-Gurion University Ben-Gurion University.

[2] A Survey of Quantification of Privacy Preserving Data Mining Algorithms Elisa Bertino, Dan Lin, and Wei Jiang

[3] An Overview on Privacy Preserving Data Mining Methodologies Umesh Kumar Singh, Bhupendra Kumar Pandya , Keerti Dixit Institute of Computer Science Vikram University Ujjain, India.

[4] Enabling Multilevel Trust in Privacy Preserving Data Mining Yaping Li, MinghuaChen, Qiwei Li, and Wei Zhang.

[5] An Efficient Approach for Statistical Anonymization Techniques for Privacy preserving Data Mining K.Anbazhagan1,R.Sugumar2,M.Mahendran3, CMJ University, Shillong, Meghalaya, India.

[6] A General Survey of Privacy-Preserving Data Mining Models and Algorithms Charu C. Aggarwal *IBM T. J. Watson Research Center Hawthorne, NY 10532* Philip S. Yu *University of Illinois at Chicago Chicago, IL 60607.*

[7] The applicability of the perturbation based privacy preserving data mining for real-world data Li Liu *, Murat Kantarcioglu, Bhavani Thuraisingham Computer Science Department, University of Texas at Dallas, Richardson, TX 75080, USA

[8] Privacy Preserving Data Classification with Rotation Perturbation Keke Chen Ling LiuCollege of Computing, Georgia Institute of Technology *f*kekechen, @cc.gatech.edu