Joint Perturbed Copies Verification Using Data Mining Techniques (Correlation)

[1]Samparthi V.S.Kumar, [2]S.Sateesh Kumar, [3]Nagaram Phani Kumar
[1,2]Dept.of CSE, Sir CRR College of Engineering, Eluru, AP, India
[3]Dept.of CSE, Raja Mahendhra College of Engineering & Technology, India

Abstract

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 widely studied perturbation-based PPDM approach introduces random perturbation to individual values to preserve privacy before data are published. Previous solutions of this approach are limited in their tacit assumption of single-level trust on data miners. In this work, we relax this assumption and expand the scope of perturbation-based PPDM to Multilevel Trust (MLT-PPDM). In our setting, the more trusted a data miner is the less perturbed copy of the data it can access. Under this setting, a malicious data miner may have access to differently perturbed copies of the same data through various means, and may combine these diverse copies to jointly infer additional information about the original data that the data owner does not intend to release. Preventing such diversity attacks is the key challenge of providing MLT-PPDM services. We address this challenge by properly correlating perturbation across copies at different trust levels. We prove that our solution is robust against diversity attacks with respect to our privacy goal. That is, for data miners who have access to an arbitrary collection of the perturbed copies, our solution prevent them from jointly reconstructing the original data more accurately than the best effort using any individual copy in the collection. Our solution allows a data owner to generate perturbed copies of its data for arbitrary trust levels on- demand. This feature offers data owners' maximum flexibility.

Keywords

Privacy Preserving Data Mining, Secure Multi Party Computation, Multilevel Trust, Random Perturbation

I. Introduction

DATA perturbation, a widely employed and accepted Privacy Preserving Data Mining (PPDM) approach, tacitly assumes single-level trust on data miners. This approach introduces uncertainty about individual values before data are published or released to third parties for data mining purposes. Under the single trust level assumption, a data owner generates only one perturbed copy of its data with a fixed amount of uncertainty. This assumption is limited in various applications where a data owner trusts the data miners at different levels.

We present below a two trust level scenario as a motivating example.

- The government or a business might do internal (most trusted) data mining, but they may also want to release the data to the public, and might perturb it more. The mining department which receives the less perturbed internal copy also has access to the more perturbed public copy. It would be desirable that this department does not have more power in reconstructing the original data by utilizing both copies than when it has only the internal copy.
- Conversely, if the internal copy is leaked to the public, then obviously the public has all the power of the mining department. However, it would be desirable if the public cannot reconstruct the original data more accurately when it uses both copies than when it uses only the leaked internal copy.

II. Background

A. 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.

B. 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.

C. 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.

D. 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.

E. Kronecker Product

In the MLT-PPDM problem, the covariance matrix of noises can be written as the Kronecker product of two matrices. In this paper, we explore the properties of the Kronecker product for efficient computation.

The Kronecker product is a binary matrix operator that maps two matrices of arbitrary dimensions into a larger matrix with a special block structure.

III. Literature Survey

A. 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".

B. Secure Multiparty Computation [SMC]

Secure Multiparty Computation (SMC) approach provides the strongest level of privacy; it enables mutually distrustful entities to mine their collective data without revealing anything except for what can be inferred from an entity's own input and the output of the mining operation alone,. In principle, any data mining algorithm can be implemented by using generic algorithms of SMC.

However, these algorithms are extraordinarily expensive in practice, and impractical for real use. To avoid the high computational cost, various solutions those are more efficient than generic SMC algorithms have been proposed for specific mining tasks. Solutions to build decision trees over the horizontally partitioned data were proposed in. For vertically partitioned data, algorithms have been proposed to address the association rule mining, k-means clustering, and frequent pattern mining problems. The work of uses a secure coprocessor for privacy preserving collaborative data mining and analysis.

C. 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 Kx 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 releasing 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 … Z_n] be the vector of noise. Let H be an identity matrix.

IV. Proposed System

This 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. We refer to this attack as a diversity attack. It includes the colluding attack scenario where adversaries combine their copies to mount an attack; it also includes the scenario where an adversary 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, we focus on the additive perturbation approach where random Gaussian noise is added to the original data with arbitrary distribution, and provide a systematic solution.

Through a one-to-one mapping, our solution allows a data owner to generate distinctly perturbed copies of its data according to different trust levels.

A. Advantages

1. MLTPPDM introduces another dimension of flexibility which allows data owners to generate differently perturbed copies of its data for different trust levels.
2. In MLT-PPDM, data miners may have access to multiple perturbed copies. By combining multiple perturbed copies, data miners may be able to perform diversity attacks to reconstruct the original data more accurately than what is allowed by the data owner.
3. We address this challenge by properly correlating perturbation across copies at different trust levels .We prove that our solution is robust against diversity attacks.

V. Results and 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.

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.

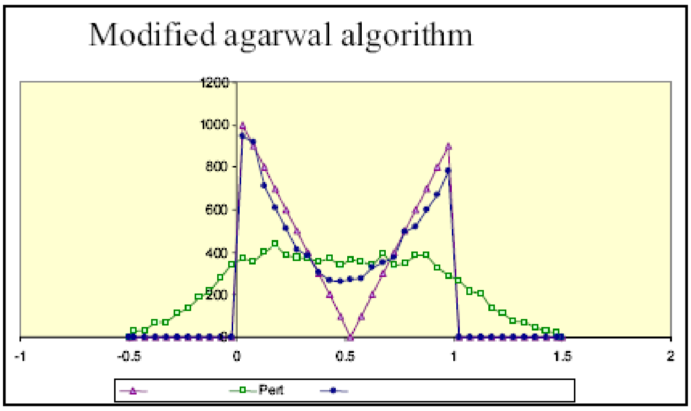

Fig. 1:

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.

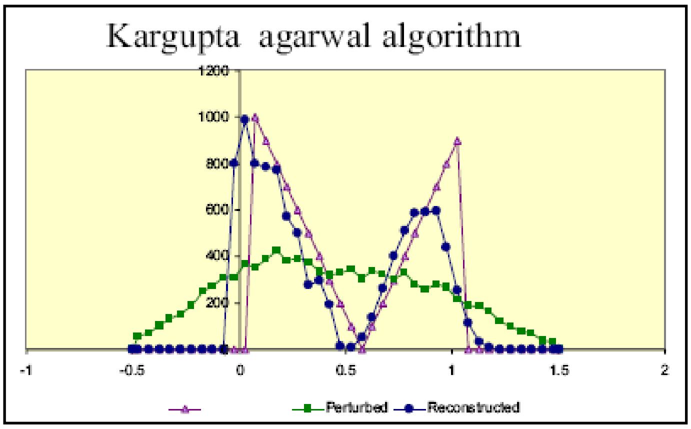

Fig. 2:

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 define the results and the approaches.

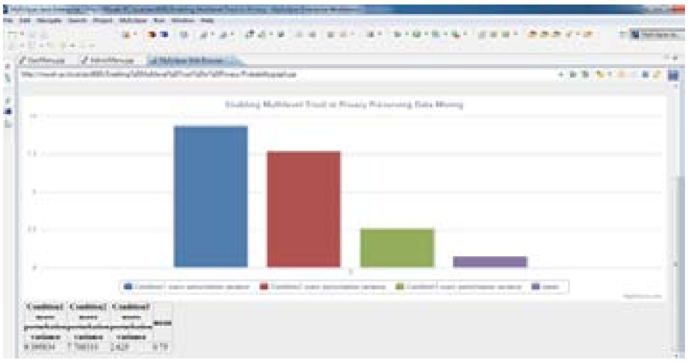

Fig. 3:

From above graph by using multilevel trust. If condition1 is satisfied then it checks for the condition2 then after if condition 2 is satisfied it checks for condistion3.data will not be transferred. If all conditions are satisfied then it calculates by the mean value. Compare to mean and co variance values.

VII. Conclusion and Future Work

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] K. Chen, L. Liu,"Privacy Preserving Data Classification with Rotation Perturbation", Proc. IEEE Fifth Int'l Conf. Data Mining.

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

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

[6] O. Goldreich,"Secure Multi-Party Computation", Final (Incomplete) Draft, Version 1.4, 2002.

[7] K. Liu, H. Kargupta, 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.

[8] D. Agrawal, 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.

[9] R. Agrawal, R. Srikant,"Privacy Preserving Data Mining.

[10] F. Li, J. Sun, S. Papadimitriou, G. Mihaila, I. Stanoi, "Hiding in the Crowd: Privacy Preservation on Evolving Streams Through Correlation Tracking", Proc. IEEE 23rd Int'l Conf. Data Eng. (ICDE), 2007.

Samparthi V S Kumar completed his M.Tech in CSE, NIT JALANDHAR. He is working as Asst. Professor in SIR C R REDDY College of Engineering, Eluru. He is having 3 years of Experience in Teaching.

S.SATEESHKUMAR pursing his M.Tech in CST in Sir C R Reddy College of Engineering, Eluru.

NAGARAM PHANI KUMAR completed his M.Tech in IT in Vignan University. He is working as Asst.Professor in CSE Department in Raja Mahendhra College of Engineering & Technology. Having 2 Years of Teaching Experience.