Comparative Study of Perturbation and Cryptography in PPDM

Shrikant Zade
PhD Scholars, Mewar University,Rajasthan
cdzshrikant@gmail.com

Prof.Pradeep Chouksey
Vice-Principal & Asso. Prof., TIT,Bhopal, Madya Pradesh
dr.pradeep.chouksey@gmail.com

Prof. R.S.Thakur
Asso. Prof., MANIT,Bhopal, Madhya Pradesh
ramthakur2000@yahoo.com

Abstract :In this paper, we studied technique for the perturbation and cryptography method. In developing this new theoretical basis we measure data utility and disclosure risk. Basically we compare random rotation and cryptographic technique for the privacy preserving. Perturbation is better than the cryptography in single party while cryptographic generally used in multi – party privacy preserving in data mining.

Key Words: Privacy, Data Utility, Disclosure Risk, Perturbation, Cryptographic

1. Introduction:

Data mining and facts finding in databases are two new research areas that investigate the automatic extraction of previously unidentified patterns from large amounts of dataset.

Privacy preserving data mining is a novel research course in data mining and statistical databases.

The common definition of privacy in the cryptographic community limits the information that is leak by the distributed calculation to be the information that can be learned from the designated output of the computation. There is also another technique called data perturbation for privacy preserving. A data perturbation procedure can be simply described as follows:

Before the data owner publishes the data, they change the data in certain way to mask the sensitive information while preserving the particular data property that is critical for building meaningful data mining models.

The main consideration in privacy preserving data mining is twofold. First, sensitive raw data like identifiers, name, addresses, should be changed or trimmed out from the original database, in order for the receiver of the data not to be able to negotiation another person’s privacy. Second, sensitive information which can be mined from a record by using data mining algorithms should also be excluded, because such knowledge can equally well compromise data privacy.

In this paper we study the number of algorithm for the preserving privacy in data mining and classify the privacy preserving technique on that basis. Basically we study the cryptographic technique and random rotation to maintain the privacy in data mining. Random rotation are of the two types addition method and multiplicative. Multiplicative method is more reliable than additive method. The cryptographic technique can be used for the multi-party (distributed) computation while the random rotation for the single as well as multi site privacy preserving.[5,12]

The trade-off between data utility and disclosure risk is evident when we consider the different data release policies. When no information is released, clearly there is no risk of any type of disclosure from data release, but there is no data usefulness either. Releasing aggregate data increases the risk of disclosure, but also provides greater utility for users compared to no release of data. Releasing micro data provides users with the highest possible level of data utility, and has also traditionally been assumed to pose the highest risk of disclosure. However, it is possible to develop masking methods that are capable of provide high levels of data utility without a corresponding increase in disclosure risk. [1]

In this paper, First section gives theoretical introduction of disclosure risk and utility of data. Second section gives in depth theoretical study of the cryptographic technique and perturbation technique. In last section we present the conclusion on this basis.

2. Definition of data utility and disclosure risk :

The most important objective of perturbation method is to provide users with access to micro data. Ideally, the released micro data should satisfy two major criteria, namely, maximum data utility and minimum disclosure risk. In this section, we define data utility and disclosure risk.

2.1. Data utility :

Data usefulness refers to the level to which the uniqueness of the released data be similar to the original dataset. Ideally, the uniqueness of the released data should be alike to the original data so that results of any analysis performed on the perturbed data are precisely the same as the results of the analysis performed on the original data. However, given the natural world of perturbation, this may not be possible since modify entity values will result in some change for some random analysis. More often, a statistically based meaning of data utility is used whereby the outcome of analysis perform on the perturbed data should be the same as that using the original dataset, except for sampling error. The bigger the dataset, the less the diversity between analyses performed on the original and perturbed datasets.

In this study, we assume that the purpose of perturbation is to provide users with a dataset whose statistical characteristics closely look like the original dataset, that is, we assume that the available data set represents a finite population.[1,2,3] An alternative perspective is provided by multiple imputations, which assumes that the available data is a sample from an unknown population. In the case of multiple imputations, data usefulness arises from the ability to turn up at valid estimates of specific unknown population parameters when using the masked data for specific types of analysis. [7]

2.2. Disclosure risk :

In terms of disclosure risk, the perturbation technique should guarantee that a snooper (a devious user or a data spy) would not be able to suppose the identity of an person (in a dataset where the identity of the individual is protected) or obtain an exact estimation of the value of a sensitive variable provided some early and universal definitions of disclosure risk. The definition of disclosure risk by Duncan and Lambert is based on the grow in knowledge and the succeeding reduction in vagueness that results from having access to specific data. Dalenius defines disclosure to have occurred when the release of data allow the user to improve the approximation of an unknown secret value.[5] Other authors have proposed precise measures of disclosure risk that can be considered as practical measurements of the general definitions [7]. Note that even prior to micro data being released, a snooper can guess the confidential variables using relationships between the non-confidential variables and aggregate data. This disclosure risk can be measured by the quantity of information concerning the confidential variables that is available from the non-confidential variables. In situations where this risk of disclosure is measured high, then yet aggregate data concerning relationships between variables may not be released. When access to micro data is provide, the snooper may use the masked micro data and non-confidential variables to guess the original values of the confidential. In other words, the release of masked micro data increase the amount of information available to users and could potentially increase disclosure risk.[8,9]

Hence, in evaluate disclosure risk, we focus on the incremental information that results from the micro data release. We assume that users have maximum preceding information in the form of aggregate data concerning all confidential variables, relationships between confidential and non-confidential variables, and micro data access to non-confidential variables. In this case, the maximum information available to the snooper is based on the distribution of the confidential variables trained on the non-confidential variables. This definition of a snooper is reliable with definitions of a snooper (or intruder) who has "verified information" prior to micro data release.[4] Then, our definition of disclosure risk can be described as an increase in identity and value disclosure resulting from the incremental information provided by access to masked micro data, given knowledge of the distribution of the real micro data.

3. A theoretical basis for perturbation methods:

We consider a dataset as consisting of both secret variables (X) as well as non-secret variables (S). The secret and non-secret variables are either numerical or categorical and X and S together account for all numerical and categorical variables. Let f (·) and F (·) represent the probability density and cumulative density functions, respectively. While the data may consist of a set of key identification variables, since they are frequently non-numerical in nature and/or may not permit numerical operation, they will not be considered further.

The purpose of perturbation is to produce a set of masked values Y such that the following requirements are satisfied.

1. Data Utility or accuracy requirements: The statistical uniqueness of Y are the same as that of X (i.e., f (Y) = f (X)), and the relationship between Y and S is the same as that between X and S (i.e., f (Y, S) = f (X, S)).

2. Disclosure Risk requirement: The confidentiality of X is maintained and the released micro data (Y, S) does not increase disclosure risk, (i.e., f (X / S,Y) = f (X / S)).[9]

3.1. General procedure for perturbation

A general procedure for generating perturbed micro data values that satisfies the data utility and disclosure risk requirements can be stated as follows.

Generate an observation y_i from the conditional distribution f (X / S= s_i) such that, given

S = s_i , Y is independent of X. Thus,

$y_i \sim f\,(X\ /\ S = s_i\,)$, --------------- (1)

and

$f\,(X,Y\ /\ S = s_i\,) = f\,(X\ /\ S = s_i\,)\ f\,(Y/\ S = s_i\,)$. ----------------- (2)

Repeat the process for every observation i in the dataset.

Under this procedure, the actual values of the i^{th} observation y_i is an independent realization from the conditional distribution of f (X / S = s_i).

This approach by means of the conditional distribution of X / S to cause perturbed micro data has also been previously investigate, in the perspective of categorical data [5] Our objective in this study is to first to establish the appropriateness of this procedure and secondly to relate this procedure to existing perturbation methods. We can show that generating the perturbed values in this manner satisfies both the data value and disclosure risk necessities.

3.2 Data utility requirements :

Mathematically, we can state the model data effectiveness requirements in terms of the marginal and joint distributions as follows:

$f\,(Y) = f\,(X)$ and $f\,(Y, S) = f\,(X, S)$. ------------------------ (3)

Since, y_i is generated from f (X| S), f (Y / S) = f (X / S).

Hence,

$f\,(Y, S) = f\,(Y\ /\ S)\ f\,(S) = f\,(X\ /\ S)\ f\,(S) = f\,(X, S)$. ---------------------------- (4)

Furthermore, $f\,(Y) = \int_s f\,(Y, S)\,ds = \int_s f\,(X, S)\,ds = f\,(X)$.

Thus, if the perturbed values are generated using the general procedure in Section 4.1, the data utility requirements will be satisfied.

3.3 Disclosure risk requirements

In Section 3.2, we defined disclosure risk as the incremental information that is provided due to micro data access. We can formalize this ideal definition as follows. We assume that users already have the maximum information regarding X and S, namely, f (X / S), as well as micro data access to the non-secrete variables S. Hence, disclosure risk prior to access to perturbed micro data is defined by the ability of a snooper to predict X using the conditional density f (X / S). When users are provided access to the perturbed micro data, they have additional information, and could improve their predictive ability by using the conditional density f (X / S,Y) to predict X. For example, consider the simple perturbation technique of adding a noise term to the values of the confidential variables to generate the perturbed variables (of the form Y = X + e). In this case, we can easily show that f (X / S,Y) is different from f (X / S), and thereby providing access to micro data will improve the predictive ability of the snooper.[10]

When the perturbed values are generated using the general procedure in Section 4.1, using equation (2) we obtain:

$f\,(X\ /\ S,Y) = f\,(X\ /\ S)$. (5)

In other words, providing users access to the perturbed micro data values Y provides snoopers with no additional information regarding the confidential variables X. Therefore, releasing Y does not increase disclosure risk since it does not provide the snooper with “additional” information when f (X / S) is known by the user prior to micro data release.

4. Cryptographic:

Basically cryptography technique is used for the multi-party data or the data that is distributed over the different sites and one party can share the data of other party.

Cryptographic research typically considers two types of adversaries: A semi-honest opponent (also known as a passive, or honest but interested opponent) is a party that correctly follows the protocol specification, yet attempts to learn extra information by analyzing the messages received during the protocol execution. On the other hand, a malicious opponent may arbitrarily deviate from the protocol specification. It is easier to design a solution that is secure against semi-honest opponent, than it is to design a solution for malicious adversary. A common approach is therefore to first design a secure protocol for the semi-honest case, and then transform it into a set of rules that is secure against malicious adversary. This transformation can be done by requiring each party to use zero-knowledge proof to prove that all steps that it is taking follows the specification of the protocol. So the semi honest adversarial model is often realistic one.[15]

4.1 Classification, decision trees and ID3.

Classification is a common problem in data mining, which is commonly solved by means of decision trees. ID3 is a basic algorithm for construct decision trees. The input to a classification problem is a structured database comprise of attribute-value pairs. Each row of the database is a transaction and each column is an attribute taking on different values. One of the attribute in the database is selected as the class attribute (e.g., it could denote whether the patient has a certain disease). The aim is to use the database in order to guess the class of a fresh transaction by performance only the non-class attributes.

A decision tree is a rooted tree contain nodes and edges. Each inner node is a test node and correspond to an attribute. The edges leaving a node correspond to the possible values taken on by that attribute. The last node of the tree contain the expected class value for transactions matching the path from the root to that leaf. Given a decision tree, one can guess the class of a new transaction by traversing the nodes from the root down, subsequent the edges that correspond to the attribute values of the transaction. The value of the leaf is the predictable class value of the new transaction. The ID3 algorithm is used to design a decision tree based on a given database. The tree is construct in top-down recursive fashion. At the root, each attribute is tested to determine how well it alone classifies the transactions. The “best” attribute (to be defined below) is then chosen and the remaining transactions are partitioned by it. ID3 is then recursively called on every partition (which is a smaller database containing only the appropriate transactions and without the splitting attribute).[13,14]

The central principle of ID3 is to choose the best predicting attribute based on information theory. The idea is to check which attribute reduce the information of the class attribute to the greatest degree. Namely, to choose the attribute that provides the maximal information gain, where this value is defined as the difference between the entropy of the class attribute, and the entropy of the class attribute given the value of the chosen attribute. This decision rule results in a greedy algorithm that searches for a small decision tree consistent with the database. (Note that we only discuss the basic ID3 algorithm, and assume that each attribute is categorical and has a fixed set of possible values.)

4.2 Privacy preserving distributed computation of ID3.

We are paying attention in a situation concerning two parties, each one of them holding a database of dissimilar transactions, where all the transactions have the similar set of attributes (this scenario is also denoted as a “horizontally partitioned” database). The parties wish to calculate a decision tree by applying the ID3 algorithm to the union of their databases

4.3 Computing information gains.

Let T be a set of transactions. The exact test for determining the best attribute is defined as follows. Let $c_1,\ldots\ldots c_l$ be the class-attribute values and let T(ci) indicate the set of transactions with class c_i. Then the information needed to recognize the class of a transaction in T is the entropy, given by

$$HC(T) = \sum^{m`}_{j=1} - (((|T(c_i)| / |T|) \log (|T(c_i)| / |T|))\text{-----------}(5)$$

Let C be the class attribute and A be some non-class attribute. We wish to quantify the information needed to recognize the group of a transaction in T given that the value of A has been obtained. Let A obtain values a1, ..., am and let T(aj) be the transactions obtain value aj for A. Then, the conditional information of T given A equals

$$HC(T/A) = \sum^{m}_{j=1} (((|T(a_j)|) / |T|) * H_C(T(c_j)))\ .\ \text{-------------------}(6)$$

Now, for each attribute A the information-gain is defined as Gain(A) = HC(T) - HC(T/A). The attribute A which has the most gain over all attributes is then selected.

The algorithm needs only to find the name of the attribute A which minimizes HC(T/A); the actual value is immaterial. Therefore, the coefficient 1/|T| can be ignored, and natural logarithms can be used instead of logarithms.[13] Let TA and TB be the transactions in Alice’s and Bob’s databases, respectively. The values |TA(aj)| and |TA(aj , ci)|, which are a function of the first database alone, can be computed by Alice independently, and a similar argument holds for Bob. Therefore the value HC(T/A) can be written as a sum of expressions of the form (vA + vB) · log(vA + vB), where vA is known to Alice and vB is known to Bob (e.g., vA = |TA(aj)|, vB = |TB(aj)|).[6,11]

The protocol computes the information gain of every attribute, such that at the end of the computation Alice and Bob hold two random shares, whose sum is equal to the information gain. None of the parties learn the information gain themselves, but they can later compare the sum of the dissimilar shares and find the attribute with the maximum gain. Protocols with this security assurance might seem weaker than protocols that are secure against collusions of say, any coalition of less than one half of the parties. After all, there is a coalition of just two parties – the two special parties, is able to crack the security of the system. Consider however a scenario where most of the parties are users (e.g. bidders) that have not established faith relationships between themselves, and there are one or more central parties that are more established.

5. Conclusion:

Generating an independent realization from the conditional distribution f (X/S = si) in the general case is not always straight forward , nor easy .

The objective of our study is to theoretical of the security and awareness of security in the data mining. As we know, it is more important in data mining. Since the perturbation is the best technique of security maintain in the single party where as cryptography is weaker.

References :

[1] Chris Clifton, Murat Kantarcioglou, Xiadong Lin, and Michael Y. Zhu, “Tools for privacy preserving distributed data mining”, SIGKDD Explorations.

[2] Dakshi Agrawal and Charu C. Aggarwal, “On the design and quantification of privacy preserving data mining algorithms”, In Proceedings of the 20th ACM Symposium on Principles of Database Systems (2001), 247–255.

[3] Rakesh Agrawal and Ramakrishnan Srikant, “Privacy-preserving data mining”, In Proceedings of the ACM SIGMOD Conference on Management of Data (2000), 439–450.

[4] Vassilios S. Verykios1, Elisa Bertino2, Igor Nai Fovino2 Loredana Parasiliti Provenza2, Yucel Saygin3, Yannis Theodoridis1 1Academic and Research Computer Technology Institute, Athens, GREECE.

[5] Keke Chen ¤ Gordon Sun y Ling Liu z “Towards Attack-Resilient Geometric Data Perturbation “

[6] Aggarwal, C. C., and Yu, P. S. “A condensation approach to privacy preserving data mining.” Proc. of Intl. Conf. on Extending Database Technology (EDBT) 2992 (2004), 183-199.

[7] Agrawal, D., and Aggarwal, C. C. “On the design and quantification of privacy preserving data mining algorithms.” Proc. of ACM PODS Conference (2002).

[8] Agrawal, R., and Srikant, R. “ Privacy-preserving data mining.” Proc. of ACM SIGMOD Conference (2000).

[9] Charu C. Aggarwal_ Philip S. Yu Statistics and Computing 13: 329–335, 2003 C _ 2003 Kluwer Academic Publishers. Manufactured in The Netherlands. “On Privacy-Preservation of Text and Sparse Binary Data with Sketches.”.

[10] KRISHNAMURTY MURALIDHAR* and RATHINDRA SARATHY† *School of Management, Gatton College of Business & Economics, University of Kentucky, Lexington, KY 40506-0034, USA
“A theoretical basis for perturbation methods”.

[11] Keke Chen Ling Liu Georgia Institute of Technology fkekechen, lingliug@cc.gatech.edu
“A Random Rotation Perturbation Approach to Privacy Preserving Data Classification”

[12] [12] AGRAWAL, D., AND AGGARWAL, C. C. “On the design and quantification of privacy preserving data mining algorithms.” Proc. of ACM PODS Conference (2002).

[13] AGRAWAL, S., AND HARITSA, J. R. “A framework for high-accuracy privacy-preserving mining.” In Proc. of IEEE Intl. Conf. on Data Eng. (ICDE) (2005), pp. 193–204.

[14] HUANG, Z., DU, W., AND CHEN, B. “Deriving private information from randomized data.” Proc. Of ACM SIGMOD Conference (2005).

[15] Cryptographic techniques for privacy preserving data mining Benny Pinkas HP Labs benny.pinkas@hp.com