ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)
Vol. 2 Issue Special 1 Jan-March 2014
International Journal of Advanced Research in Computer Science & Technology (IJARCST 2014)

Data Separations for Providing Multilevel Trust in Privacy Preserving Data Mining

[I]S.Saranya, [II]Y.Arockia Jesuraj
[I,II]Dept. of CSE, Sethu Institute of Tech., Affiliated to Anna University, Kariapatti, Tamilnadu, India

Abstract
To preserve the privacy in data mining, random perturbation method for individual value is introduced. This Perturbation method is used before the data are published to third parties for mining purposes. The Existing (Privacy Preserving Data Mining) PPDM approach assumes that single level trust on data miners. From the single level trust, a data owner generates only one perturbed copy of its data with a fixed amount of uncertainty. The proposed approach of the PPDM introduces the multilevel trust on data miners. Here different perturbed copies of same data are available to data miner at different trust levels and may combine these copies to jointly add additional information about original data and release the data is called diversity attacks. To avoid these kinds of attacks, using the multilevel PPDM approach. Random Gaussian noise is added to the original data with arbitrary distribution. So, the data miners will have no diversity gain in their joint reconstruction of the original data. This allows data owners to generate perturbed copies of its data on demand at arbitrary trust levels.

Keywords
Random perturbation, multilevel trust, diversity attack, Gaussian noise, Data Preprocessing

I. Introduction

In general, privacy preservation occurs in two major dimensions: users personal information and information concerning their collective activity. The primary goal of data privacy is the protection of personally identifiable information. In general, information is considered personally identifiable if it can be linked, directly or indirectly, to an individual person. Thus, when personal data are subjected to mining, the attribute values associated with individuals are private and must be protected from disclosure. Miners are then able to learn from global models rather than from the characteristics of a particular individual. Protecting personal data may not be enough. Some- times, we may need to protect against learning sensitive knowledge representing the activities of a group. The protection of sensitive knowledge as collective privacy preservation [1]. The goal here is quite similar to that one for statistical databases, in which security control mechanisms provide aggregate information about groups (population) and, at the same time, should prevent disclosure of confidential information about individuals. However, unlike as is the case for statistical databases, another objective of collective privacy preservation is to preserve (hide) strategic patterns that are paramount for strategic decisions, rather than minimizing the distortion of all statistics (e.g., bias and precision). The goal here is not only to protect personally identifiable information but also some patterns and trends that are not supposed to be discovered. These miners always act legally in that they perform regular data mining tasks and would never intentionally breach the privacy of the data. The malicious data miners would purposely breach the privacy in the data being mined. Malicious data miners come in many forms. The focus on particular sub-class of malicious miners. That is, malicious data miners follow standards but are curious, they follow proper protocols and standard procedures, but they may perform some analysis to discover private information. Data perturbation [5] is a data security technique that adds 'noise' to databases to allow individual record confidentiality. This technique allows users to ascertain key summary information about the data while preventing a security breach. The data security is focused not on physical and technical access methods, but on statistically based methods that seek to protect confidential data by using data perturbation techniques. Data perturbation involves adding random noise to confidential, numerical attributes, thereby protecting the original data. Even while altering the original data, these methods allow users the ability to access important aggregate statistics (such as means, correlations and co variances, etc.) from the entire database, thus 'protecting' individual records. For the sales data, an employee may not be able to access what a particular individual purchased from a store on a given day, but that employee could determine the total sales volume for the store on the same day. The single-level trust PPDM problem via data perturbation [2] has been widely studied in the literature. In this setting, a data owner implicitly trusts all recipients of its data uniformly and distributes a single perturbed copy of the data. A widely used and accepted way to perturb data is by additive perturbation. This approach adds to the original data, X, some random noise, Z, to obtain the perturbed copy, Y, as follows:

$$Y=X+Z$$

The Diversity attacks referred to as 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. By addressing this challenge in enabling MLT-PPDM services. In particular, on the additive perturbation approach [12] 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 [3]. The data mining consists of various methods. Different methods serve different purposes, each method offering its own advantages and disadvantages. However, most data mining methods commonly used for this review are of classification category as the applied prediction techniques assign patients to find out the diabetes patient and generate rules for the same. Hence, the diabetes problems are basically in the scope of the widely discussed classification problems. In data mining, classification is one of the most important tasks. It maps the data in to predefined targets. It is a supervised learning as targets are predefined. The aim of the classification is to build a classifier based on some cases with some attributes to describe the objects or one attribute to describe the group of the objects. Then, the classifier is used to predict the group attributes of new cases from the domain based on the values of other attributes. The commonly

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

used methods for data mining classification tasks can be classified into the following groups.

II. Related Work

A. Multiparty Computation

To address the issue of privacy preserving data mining, in which two parties owning confidential databases wish to run a data mining algorithm on the union of their databases, without revealing any unnecessary information. The above problem is a specific example of secure multi-party computation [6] and as such, can be solved using known generic protocols. However, data mining algorithms are typically complex and the input usually consists of massive data sets. The generic protocols in such a case are of no practical use and therefore more efficient protocols are required.

B. Maintain Privacy to Individual Party

Several specific computations, such as database query, intrusion detection, data mining, geometric computation, statistical analysis, and scientific computations. These computations from another perspective secure multi-party computation perspective that is how to conduct these computations among multiple parties [7], while maintaining the privacy of each party's input. The results denied with a number of secure multi-party computation problems, among which some are well studied for decades, and some are just new problems.

C. Formalize Private Database Sharing

The data in each database can be revealed to the other databases. However, there is an increasing need for sharing information across autonomous entities in such a way that no information apart from the answer to the query is revealed. To formalize the notion of minimal information sharing across private databases [8], and develop protocols for intersection, equijoin, intersection size, and equijoin size and proved that these protocols disclose minimal information apart from the query result. Using protocol for computing equijoin size, but this protocol leaks some information about which tuples joined, based on the distribution of duplicates.

D. Data Anonymization Technique

Random perturbation is a classic data anonymization technique that has significant importance. To settle the problem with a novel algorithm for randomly perturbing a dataset at an infinite number of privacy levels [9]. The solution here is to robust even when recipients attempt to infer extra sensitive information by sharing their data together. The algorithm works in the online setting where the privacy levels of future data requests are unknown in advance, and can arrive at an arbitrary order. Finally, in addition to its rigorous and strong privacy guarantees, the proposed technique is also highly efficient, as its expected space and time complexities are asymptotically optimal.

E. Data Perturbation

The treatment of the randomization approach in the presence of public information. This also provides a framework for analysis of other future members of this privacy preserving methods. This framework to illustrate a number of insights of the randomization method [10]. That shows the degrading effect of the dimensionality curse, and quantifies the required perturbation level as a function of the dimensionality. A careless choice of the perturbing distribution can degrade the privacy behavior in subtle ways because of the presence of public information. This shows that privacy is an extremely elusive goal for the randomization method, when public information is injected into the analysis.

In Existing System, using additive perturbation technique provides the strongest level of privacy. In proposed System, the multi level trust is used. Here, more than one perturbed copies are generated by the data owners. It is only handling with linear attacks.

III. Proposed Approach

To address the problem of developing accurate models about data, without access to precise information in individual data records. To propose additive Random Rotation Perturbation based PPDM Approach provides multi-level trust in privacy preserving data mining [4]. To avoid nonlinear attacks, the privacy guarantee value is calculated. Based on those values, it allows data owners to generate perturbed copies of data at arbitrary trust levels, and gives more flexibility. It provides maximum flexibility for data owners. It is not easy to reconstruct the data. It generates more than one perturbed copies based on data miner request. Here, multilevel trust is used.

A. Dataset Preprocessing

Data pre-processing is the important step in the data mining process. Data-gathering methods are often loosely controlled, resulting in out-of-range values impossible data combinations missing values, etc. Analyzing data that has not been carefully screened for such problems can produce misleading results. Thus, the representation and quality of data is first and foremost before running an analysis. If there is much irrelevant and redundant information present or noisy and unreliable data, then knowledge discovery during the training phase is more difficult. Data preparation and filtering steps can take considerable amount of processing time. Here the dataset is preprocessed, i.e., the dataset is loaded into the database. After loading the data, extract the data, which is used for our process.

B. Perturbed Copy Generation

The data owner determines the M trust levels, and generates M perturbed copies of the data in one batch. In this case, all trust levels are predefined and all are given when generating the noise. It generates noise to the perturbed copies of the dataset. The Noise Generation is based on the Gaussian Noise process. Let G1 through GL be L Gaussian random variables. They are said to be jointly Gaussian. It follows linear combination of multiple independent Gaussian random variables. G1 through GL are jointly Gaussian. It is a linear combination of them and also a Gaussian random variable.

C. Data Miner Authentication

The trust level of the data miner is evaluated. The trust level evaluation is based on checking the miner details, when giving access request to the dataset. The trust level is evaluated based on the miner details, that the user id and some other details. The user id is generated when the user registration phase. And also the miner authentication will be carried, when the user gives access request to the owner, verify the user id and some other details given by the user from already stored in the user database. The trust level is classified as low, medium and high level.

D. On Demand Generation

The perturbed copies are generated based on the data miner request and trust level. If the miner trust level is very low means, add high level noise to the data. If the trust level of the user is medium means, add medium level noise to the data. If the trust level of the user is high means, add low level noise to the data. The noise adding procedure is done as on demand request from the user. The malicious data miners [11] always attempt to reconstruct a more accurate estimate of the original data given perturbed copies. To avoid this, the data owner wants to distribute a total of M different perturbed copies of its data, each for a trust level based on the data miner request.

E. Performance Result

The performance result for error estimation is based on considering the reconstructing process. The data miner want to reconstruct the data, they didn't get the original data. In reconstructing process, error estimation is calculated. In time and space complexity, we prove that the time space complexity is reduced compare to existing System. The performance gap between the proposed framework and other approaches is at the high level compare to other approaches. It provides better flexibility to the data owners.

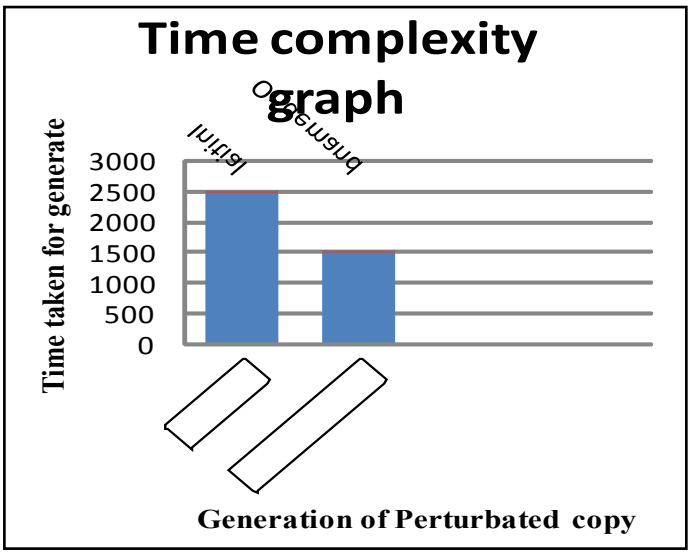

Fig. 1: Time complexity Graph

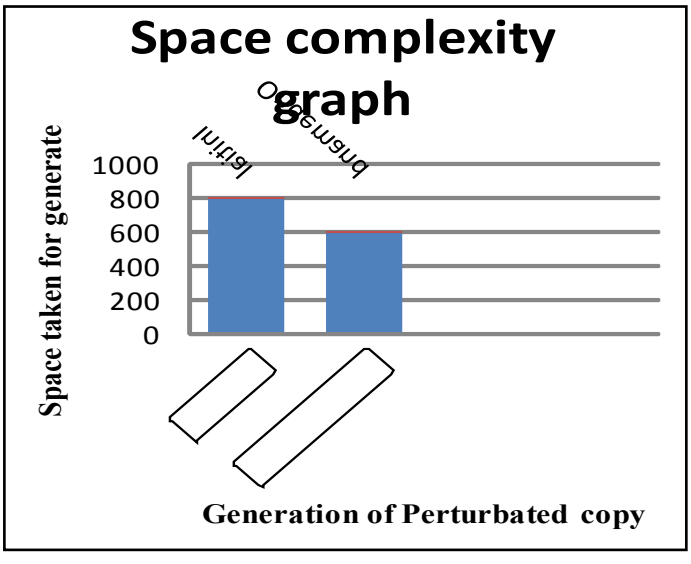

Fig. 2: Space Complexity Graph

IV. Conclusion and Future Enhancement

Using additive perturbation based PPDM approach for multilevel trust is used for providing better flexibility and security. MLT-PPDM allows data owners to generate differently perturbed copies of its data for different trust levels. This method address the challenge of 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. This challenge is addressed by properly correlating noise across copies at different trust levels. So, the data miners will have no diversity gain in their joint reconstruction of the original data. Finally, our solution allows data owners to generate perturbed copies of its data at arbitrary trust levels on-demand.

However the data miner is low, medium and high level and send that data to the data miner. The drawbacks of this system is defined as, if it is low trust means have to send that data. To avoid that, in future have to calculate privacy guarantee value of each and every user, and evaluate the user, in this module, the privacy guarantee value is calculated, based on the data miner request. By Random Rotation Perturbation Method the value is evaluated using threshold value. If the value is lower than the threshold value, the data miner is evaluated, and the data owner decides not to release the data.

V. Acknowledgement

References

[1] *D. Agrawal and 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.*

[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] *Y. Lindell and B. Pinkas, "Privacy Preserving Data Mining," Proc. Int'l Cryptology Conf. (CRYPTO), 2000.*

[7] *J. Vaidya and C.W. Clifton, "Privacy Preserving Association Rule Mining in Vertically Partitioned Data," Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, 2002.*

[8] *J. Vaidya and C. Clifton, "Privacy-Preserving K-Means Clustering over Vertically Partitioned Data," Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, 2003.*

[9] *A.W.-C. Fu, R.C.-W. Wong, and K. Wang, "Privacy-Preserving Frequent Pattern Mining across Private Databases," Proc. IEEE Fifth Int'l Conf. Data Mining, 2005.*

[10] *B. Bhattacharjee, N. Abe, K. Goldman, B. Zadrozny, V.R. Chillakuru, M.del Carpio, and C. Apte, "Using Secure Coprocessors for Privacy Preserving Collaborative Data*

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

Mining and Analysis," Proc. Second Int'l Workshop Data Management on New Hardware (DaMoN '06), 2006.

[11] C.C. Aggarwal and P.S. Yu, "A Condensation Approach to Privacy Preserving Data Mining," Proc. Int'l Conf. Extending Database Technology (EDBT), 2004.

[12] E. Bertino, B.C. Ooi, Y. Yang, and R.H. Deng, "Privacy and Ownership Preserving of Outsourced Medical Data," Proc. 21st Int'l Conf. Data Eng. (ICDE), 2005.

Author's Profile and Image

S.Saranya received her Bachelor Degree in Computer Science and Engineering from Sethu Institute of Technology ,India, in 2007.She is currently pursuing Master Degree in Computer Science and Engineering from Sethu Institute of Technology ,India. Her interest includes Data Mining and Data Warehousing.