A Technique to improve Security of Data in Multilevel Trust

A. Viji Amutha Mary
Assistant Professor, Dept of CSE, Faculty of Computing,
Sathyabama University, Old Mahabalipuram Road,
Chennai, 600119, Tamil Nadu, India
vijiamumar@gmail.com
http://www.sathyabamauniversity.ac.in

Dr. T. Jebarajan
Professor and Head, Dept of CSE, Rajalakshmi Engg College, Thandalam,
Chennai, Tamil Nadu, India
drtjebarajan@gmail.com

Abstract

The Privacy Preserving Data Mining technique that is used widely to conserve security of data is a random perturbation method. The original data is modified and many copies are created according to the trust levels in each field. The addition of noise level also varies with each trust level. The amount of noise added to the lower order trust level is less, whereas it is high in the higher order trust level. There is a chance for the hackers to reconstruct the original data with some non linear techniques. This challenge is addressed by the proposal of a novel non linear technique. Here, the noise level of each copy is checked after perturbation of the actual data. If any similarity is found in the noise level, reconstruction of the original data is possible. Therefore, additional noise is included in the perturbed data. This process is repeated till zero percent of similarity of data is attained.

Keywords: multilevel trust; perturbation; privacy preserving; data mining

1. Introduction

Today we survive in a world where a large amount of data is available. We dig out useful information from the vast data repository. There arises a need to conserve the confidential information form the mined data. Therefore, before outsourcing the data in the world, such private data must be protected. For example, patients' medical records in a hospital may be used for data analyzers to build a classification model using patients' epoch, smoking habits and obesity condition to predict their lifetime or to study the characteristics of various diseases. On the other hand, data publishers are often prohibited by law from revealing any person-specific information that compromises an individual's privacy.

Mining data sets that include information about people in a population is a great way of exhibiting intellectuality. This will also help in statistical analysis and in taking the survey of people's records. Applications include observing the results of treatments on diseases and dealing with disease outbreaks. Besides this fruitful information, the data sets also include sensitive information like the disease of an individual, the salary of an employee, etc. The goal is to prevent hackers from identifying the original sensitive data of an individual.

2. Overview of Existing System

Let us take the example of health care data system for the purpose of detecting bioterrorism. A detailed study on clinical data, its records and the pharmaceutical related transactions of the data and also of the drugs is necessary. But there is a probability of violating the laws of security if clustering is done on different data sets. The major organizations are not allowed to disclose the private confidential information. The hackers may combine the data sets and may reconstruct the original data. Therefore, a highly efficient technique that mainly focuses on preserving the private information should be devised. The interrelated characteristics and dependencies of data that are needed for the field of data mining should also be hidden.

Multiplicative noise is used widely to maintain the data privacy. There are mainly two ways by which the multiplicative noise is included in the data set. One way is to include the multiplicative noise to each confidential attribute in the data set. The noise must contain a lower Gaussian distribution with a mean value one and a less variance factor. The other method is to find the logarithmic value of the secret attribute, add a constant Gaussian noise value and then find the logarithmic inverse of the modified data. If only minute changes are to be made to the data, the first method works well. Otherwise, the alternate method can be employed. This method guarantees a higher security privilege compared to the first method. The only drawback is that the access of data becomes tough since the value of data lies in the logarithmic range value.

In the traditional perturbation techniques such as addition and multiplication, each private data attribute is modified independently. The records are not checked in a pairwise manner. Therefore, still the privacy of data is not assured. Moreover, these conventional techniques consider only numeric data. The owner of the data replicates the private data into multiple copies after the addition of noise. The noise level of each copy is found out with the application of the linear technique. Then the noise levels are compared to find out the similarity. If similarity is found in the noise levels, additional noise is included before publishing the data. In this system, the diversity attacks from the intruders are addressed with the help of linear techniques. But the intruders must only perform nonlinear techniques to avoid the attacks. This issue can be handled properly with the framing of non linear techniques.

3. Proposed Non Linear Technique

A network is developed to protect the privacy of person specific data. This proposed network system is useful in the medical field. The patient records are kept confidential and can be used in research studies by analyzing the patients' habits of life, epoch and obesity condition to predict their lifetime and to observe the characteristics of several diseases. The patients' diseases are kept secret because they don't like to expose it. Therefore the disease attributes of patients is distorted by the addition of noise. In the existing system, there is a possibility that the confidential data can be intruded by the authorized party by making diversity attacks with the help of non-linear techniques. To address this issue, the proposed network system is trained with nonlinear techniques using Sequential Generation Algorithm. Therefore, the attackers are not unable to trace the original secret data with diversity attacks.

The level of noise between the original data and the distorted data is found out. If the noise level is within the minimum threshold value, the third party can easily hack the data. Therefore, the server requests the data holder to add some more noise to the data. After the addition of noise, the noise level is again compared. This process is repeated until the noise level falls within the minimum threshold value. Thus, there will not be many similarities between the actual and the distorted data. The original and the modified data shows high deviations, and the privacy level of the data is ultimately raised. This technique prevents the similarity of data. The hackers also cannot misuse the data either by linear or by non-linear techniques.

The basic steps required for the functional architecture of the proposed system are below.

Algorithm for the proposed system:

(1) Loading the data to the server.
(2) Obtain the sensitive data.
(3) The data miner registers with the server.
(4) Add noise to the private data.
(5) Compare the noise level with the minimum threshold level.
(6) If the noise level is higher than the threshold level, add some more noise.
(7) Repeat steps (5) and (6) till the noise level falls within the threshold level.
(8) Publish the data.

4. Related Work

In the year 2006, Xiaokui Xiao developed a framework for implementing personalized anonymity. A new generalization based framework is represented with the help of quasi-identifier attributes and K-anonymity. A table is said to be k-anonymity, only if the QI values of each tuple are similar in values to those of at least $k-1$ other tuples [Xiao X and Tao Y (2006)]. Privacy protection is not guaranteed if there are multiple tuples in the data. The existing algorithm is not an optimized one and there is loss of information.

Ada Wai-Chee Fu suggested the concept of mining frequent item sets where the data are residing in multiple sites [Fu (2005)]. Here, the secrecy of an individual is not exposed. The mining of frequent patterns has significance not only in itself, but also for other data mining tasks such as mining of association rules for the data, correlations, sequences, classifiers and clusters in the related data. The difficulty that occurs in a database structure across multi dimension sites is handled in the frequent pattern mining. These sites are considered to be semi honest and distrust. This problem is solved by the use of a star schema by the concept of semi-join for preserving privacy.

In [Liu K et al. (2006)], the data is random projected and it is distorted using Multiplicative Data Perturbation technique. The statistical features and the properties related to distance are well preserved without disturbing the dimensionality and the actual values of the data. The random projection technique may become more powerful when it is combined with other geometric transformation techniques.

Daniel Kifer proposed Log-linear models and logistic regression models which are the popular techniques for analyzing tabular data [Kifer D and Gehrke J.E. (2006)]. They provide a compact and interpretable representation of high dimensional probability distributions. The K-Anonymity and l-diversity are

weaker privacy definitions they do not protect against adversaries with arbitrary amounts of Background knowledge but they provide considerably more utility.

The random perturbation technique was proposed which is a popular method of producing protected data for preserving privacy [Xiao et al. (2009)]. Its simplicity and good privacy protection allow mining of a variety of data patterns. Random perturbation mainly focuses on a detailed random perturbation procedure, which refers to as uniform perturbation.

An effective anonymization algorithm was suggested to thwart the attacks while publishing data [Fung C. M. et al (2008)]. In our practical life, we ought to publish data enormously and successively and the same data may be modified and sent to another recipient. In such cases, even when all the new released data are properly *k*-anonymized, the vagueness of an entity may be inadvertently compromised if the recipient rechecks all the received data. In the paper, we systematically characterize the correspondence attacks. Here each release contains the new data as well as previously collected data. Finally, it is proved that the detection and the anonymization methods are extendable to deal with multiple releases and other privacy requirements.

Li and Chen in the year 2012 suggested that avoiding diversity attacks are the main challenge of providing multi level trust PPDM services This challenge is addressed by efficiently correlating distortion across copies at various levels of trust. The drawback is that it considers only linear attacks. But it offers maximum flexibility to the data owner. In the future, nonlinear techniques may be applied to derive original data and recover more information.

An approach that has been proposed to maintain secrecy is to anonymize the dataset that includes private or secret information about a particular field [Li and Chen (2012)]. The two common manipulation methods by which k-anonymity of a data set can be achieved are i) Generalization and ii) Suppression. Generalization is defined as substituting a value with a less specific, but a semantic coherent value. Suppression is referred as not to release a data value at all. It has higher predictive performance when compared to existing methods. In the future, the proposed method can be extended to other data mining tasks and to another anonymity measures.

5. Implementation of Proposed System

The detailed information regarding how the network for preserving the privacy of data is developed is elaborated here. The developed system framework contains a server and the users. The types of users are data owner and consumer.

5.1. *User Authentication*

This phase contains details about network deployment and user authentication. The network in a medical field consists of three types of users, researchers, doctors and nurse. The server has the responsibility to control all the users and the network. Then the data owner and the consumer registers with the server and get their authentication certificates.

5.2. *Generation of Noise*

The data holder uses nonlinear techniques to avoid diversity attacks. He generates noise using the sequential generation algorithm. Add generated noise in every trust level using nonlinear techniques. Compare the modified data and the original data and check the similarity of data. Input the noise data to non-linear techniques to find the similarity level. If the noise is similar, add some more amount of noise that is generated by the sequential generation algorithm. If it is not similar, the data can be published.

5.3. *Modifying data at every trust level*

If the noise is similar, then the data owner generates noise using the sequential generation algorithm. Then add the generated noise in every level input data. Input the noisy data to nonlinear techniques. This technique finds the noise and predicts the similarity. Based on the similarity, the data owner adds additional noise to prevent the similarity. The noise level is compared again. If the noise is not similar, then the data owner uploads modified data.

5.4 Process of user and server

In this phase, the user sends a request for a particular data to the server. Then the server verifies the details and send the required file to the user based on priority level. Here the user who may be a researcher, doctor or nurse receives the data.

6. Performance Evaluation

The proposed system is trained to overcome both the linear and non-linear attacks from the hackers. The existing system deals only with linear attacks. The graph, Fig. 1 shows the security level of the linear and non-linear perturbation techniques. The x axis represents the perturbation techniques and the y axis, the security level of these techniques. From the graph, it is observed that the non-linear technique gives a higher security level compared to the other existing techniques, additive and multiplicative perturbation.

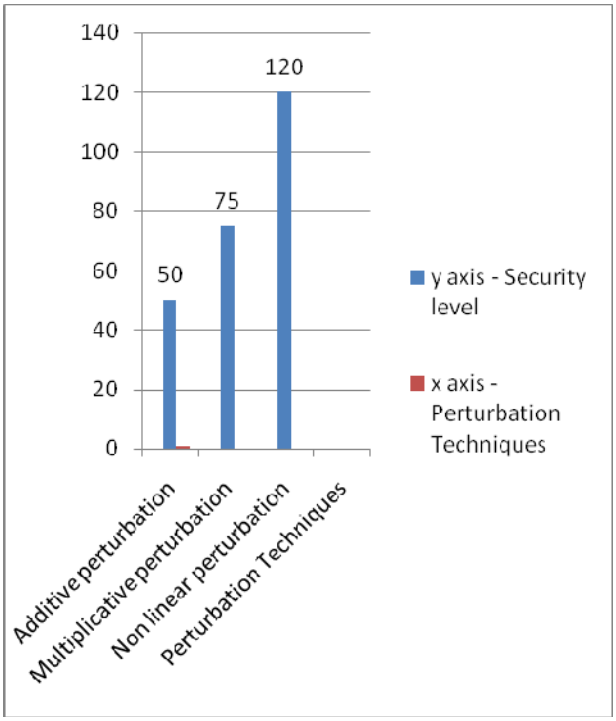

Fig. 1. Security level of the perturbation techniques

7. Conclusion and Future Work

The main challenge is to restrict the third party from reconstructing the actual data by combining the data copies at different levels of trust. The existing systems have used the linear techniques, which can be easily intruded. This issue is overcome by properly comparing the noise level repeatedly across copies at different trust levels with nonlinear techniques. The system promotes maximum prevention from reconstructing the original copy from the distorted one. The security of data transmission also can be improved in the future.

References

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

[2] Fung B, Wang K, Fu A, and Pei J (2008): Anonymity for Continuous Data Publishing, Proc. Int'l Conf. Extending Database Technology(EDBT).

[3] Kifer D and Gehrke J.E. (2006): Injecting Utility Into Anonymized Datasets, Proc. ACM SIGMOD Int'l Conf. Management of Data.

[4] Kisilevich; Rokach and Shapira. (2010): Efficient Multidimensional Suppression for K- Anonymity, IEEE Transactions on Knowledge and Data Engineering, Vol 22, No 3.

[5] Li and Sarkar (2006): A Tree-Based Data Perturbation Approach for Privacy Preserving Data Mining, IEEE Transactions on Knowledge and Data Engineering.

[6] Li and Chen (2012): Enabling Multilevel Trust in Privacy Preserving Data Mining, IEEE Transactions on Knowledge and Data Engineering, Vol. 24.

[7] Liu K, Kargupta H, and Ryan J (2006): 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.

[8] Xiao X and Tao Y (2006): Personalized Privacy Preservation, Proc. ACM SIGMOD Int'l Conf. Management of Data.

[9] Xiao X, Tao Y and Chen M (2009): Optimal Random Perturbation at Multiple Privacy Levels, Proc. Int'l Conf. Very Large Data Bases.