A REVIEW ON PRIVACY PRESERVATION TECHNIQUES FOR DATA MINING

[1]Er. Amandeep Kaur,[2]Dr. Dinesh Kumar
Department of Computer Engineering
Guru Kashi University
Talwandi Sabo Bathinda,Punjab(151001) India
Department of Computer Engineering
Guru Kashi University
Talwandi Sabo Bathinda,Punjab(151001) India

***Abstract:** Data Mining has been widely used process for extracting useful data from large storage of data. This data mining can be used for different types of large repository of the databases. Like IOT and BIG DATA, Data Warehouse etc. but while extract of mining the data the sensitiveness of the data will be lost. Because it present the data in formatted way to the client who has requested the data. Privacy preservation is the important branch in the data mining field. This field has gained special attention in the now a days. Privacy preserving also protect private and sensitive data from disclosure without the permission of data owners or providers. In this paper various PPDM techniques have been compared on the basic of its advantages and disadvantages. Also discuss the future research that can be taken place in this field.*

***Keyword:** Privacy preservation, Data mining, Data Perturbation, K-anonymity.*

I. INTRODUCTION

Data mining is a series of methods used for data analysis in order to extract useful information and reveal underlying patterns in huge amounts of data. It has become an emerging and rapidly growing field in recent years. Many services and organizations have benefited from invaluable knowledge or patterns obtained from the data mining. For example, in telecommunications, data mining tools help operators to analyze consumer behaviors in order to pursue customer interests for developing new service packages. A commercial bank can utilize the data mining to obtain investment or deposit patterns in order to adjust its interest rate. The data mining brings insights into mess data and investigates explicit information to assist decision-making.

Apart from supporting business development, data mining is also applied in daily applications. For instance, a system or website can use data mining tools to recommend applications and contents to users according to their behaviors of data access and sharing over the Internet. In terms of Internet of Things, there are many applications based on data mining to offer intelligent services in a pervasive manner in wide fields such as healthcare, anomalous behavior detection, security and safety assurance, surveillance, and so on. Data mining is a valuable technique and it is continually being developed with newly raised demands. In recent years, concerns are gradually growing over individual privacy and sensitive data protection. This is particularly because large amounts of personal data could be misused without permission. Individual privacy could be intruded during data mining since it could figure out implicit private information unwilling to be disclosed. This kind of private data disclosure and misuse needs to be addressed by adding privacy preserving features into the process of data mining. Privacy-Preserving Data Mining (PPDM) aims to support data mining related computations, processes or operations with expected privacy preservation [2]. PPDM is highly related to Secure Multi-party Computation (SMC). It is also an interesting research topic in the field of privacy preservation.

PPDM is designed to protect personal data and sensitive information from disclosure to the public in the process of data mining. Such privacy protection is generally required in practice. On one hand, data providers are aware of the problems of privacy disclosure and intrusion. On the other hand, business holders concern the security of sharing their secret information. For example, consumers do not like telecommunication operators to disclose their phone call information, and a business partner does not want to share secret information with other entities during jointly data mining. Therefore, PPDM is a very important issue that should be solved in the applications related to data analysis, especially in the area of business development and future Internet of Things.

Much research has been carried out in the area of PPDM. Many algorithms have been proposed in the literature. Existing techniques mainly focus on preserving private information in different stages of a data mining process. In this paper, we review main PPDM techniques according to a PPDM framework that has three layers: Data Collection Layer (DCL), Data Pre-Process Layer (DPL) and Data Mining Layer (DML).

II. A PPDM FRAMEWORK

PPDM framework by referring to [6] to guide our PPDM technical review. The framework was constructed according to the stages in the data mining process, from data collection, pre-process, to final data mining procedure. The PPDM framework contains three layers: Data Collection Layer (DCL), Data Pre-Process Layer (DPL) and Data Mining Layer (DML), as shown in Figure 1.

The first layer DCL contains a huge number of data providers that provide original raw data that could contain some sensitive information. The privacy-preserving data collection can be carried out during the data collection time. All the data collected from the data providers will be stored and processed in the data warehouse servers in DPL.

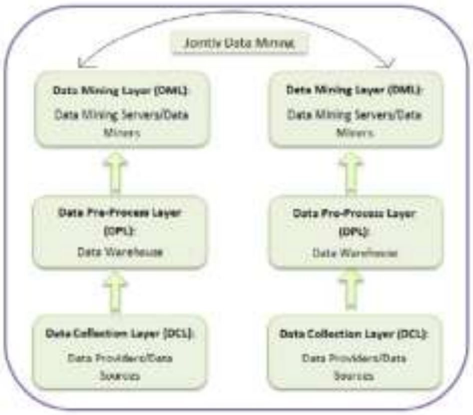

Jointly Data Mining
Data Mining Layer (DML):
Data Mining Servers/Data Miners
Data Mining Layer (DML):
Data Mining Servers/Data Miners
Data Pre-Process Layer (DPL):
Data Warehouse
Data Pre-Process Layer (DPL):
Data Warehouse
Data Collection Layer (DCL):
Data Providers/Data Sources
Data Collection Layer (DCL):
Data Providers/Data Sources

Fig. 1: A PPDM Framework[1]

The second layer DPL contains data warehouse servers that are responsible for storing and pre-processing the collected raw data from the data providers. The raw data stored in the data warehouse servers could be aggregated in sum, average etc., or pre-computed using privacy-preserving methods in order to make the data aggregation or fusion process more efficient. The privacy preservation in this layer concerns two aspects. One is privacy-preserving data preprocessing for later data mining, and the other is the security of data access.

In this paper, we focus on the privacy preserving data pre-computing methods, whereas the secure data access control is beyond the scope of this paper. The third layer DML consists of data mining servers and/or data miners located mostly in the Internet for conducting actual data mining and providing mining results. In this layer, privacy preservation concerns two aspects. One is improving or optimizing data mining methods to enable privacy-preserving features. The other is collaborative data mining based on the union of a number of data sets owned by multiple parties without revealing any private information.

III. PRIVACY PRESERVATION IN DATA COLLECTION LAYER

Data privacy preservation for data providers aims to protect raw data from disclosure without the providers' permission. Because the raw data is collected directly from the data providers, the privacy preservation in the DCL can be seen as the privacy preservation during data collection. Generally, there are two methods used to hide the raw data from its original value. The first method is to encrypt all the raw data so that no one can access the plain data except authorized data processors or miners. However, large amounts of data would lead to big computational cost for both data providers and data miners. It is infeasible to encrypt all the raw data in real-life applications. The second method is to perturb original data values in order to hide real private information. Currently, there are many data modification methods proposed for the privacy-preserving data collection. Most data modification methods used during data collection in

the DCL can be classified into two groups: value-based methods and dimension-based methods [26]. In what follows, we review the two groups of techniques in data perturbation.

A. Data Perturbation

i. Value-based methods

Random Noise Addition is the most common data perturbation method in the value-based group [6]. It is regarded as a method of Value Distortion. Random Noise Addition is described in [6] as:

$X=Xi+r$ (1)

where Xi is the original data value of a one-dimensional distribution, and r is a random value drawn from a certain distribution. This method distorts the original data values by adding random values as random noise and returning the processed value. The distributions used in this method are usually Uniform or Gaussian. The authors of [6] proved that the Random Noise Addition as a method of data perturbation can be used for data pre-processing before performing actual mining while at the same time preserving the data privacy by reconstructing the distributions, but not the individual values of the original data set. They also showed that Gaussian perturbation does better than Uniform perturbation in terms of achieving the same privacy preservation level. But the Uniform perturbation is easier to deploy than the Gaussian.

Therefore, which distribution is selected as random noise depends on application requirements.

ii. Dimension-based methods

The dimension-based methods were proposed to overcome the disadvantages of the value-based methods. In real-life applications, data sets are usually multi-dimensional, which could increase the difficulty of data mining process and affect the data mining results, especially in the tasks where multidimensional information is crucial for the data mining results. However, most value-based perturbation methods only concern preserving the distribution information of a single data dimension. Therefore, they have an inherent disadvantage to provide accurate mining results in a data-mining task that requires information over multiple correlated data dimensions. The dimension-based methods concern keeping multidimensional information when doing data perturbation to preserve data privacy. The most common dimension-based methods used during data collection are Random Rotation Transformation and Random Projection.

Random Rotation Transformation was proposed to decrease the loss of privacy while not affecting the quality of data mining [7], and it is often used for privacy-preserving data classification. The authors in [7] achieved this by multiplying a rotation matrix to a data set matrix as:

$g(X) = RXi$ (2)

where *R* represents the rotation matrix and *X* is the original data set. The rotation is conducted in a way to preserve the multi-dimensional geometric properties, such as Euclidean distance or inner product, of the original data set. Because the rotation does not perturb data points equally, and the points near the rotation center can have few changes, it could make privacy protection over these points weak. In order to solve this problem that is vulnerable to rotation oriented attacks, the rotation center is randomly selected in a normalized data space to make the weakly perturbed points unpredictable [7, 3].

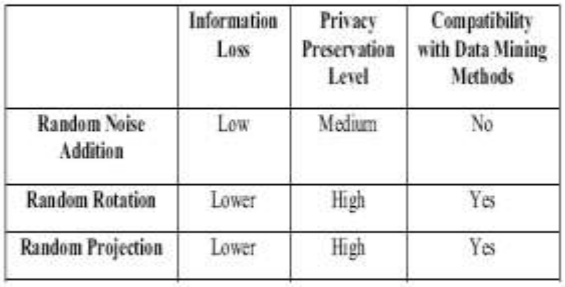
This Random Rotation Transformation method can provide high level of privacy protection and assure expected accuracy of data mining results.

Random Projection is a promising dimension-based data perturbation method. The original idea was proposed by Johnson and Linden Strauss[11] in order to reduce the dimensionality of original data set by projecting the set of data points from a high-dimensional space to a randomly chosen lower-dimensional subspace. In [7], the authors provided several properties of the random matrix and random projection, which are good for maintaining a data utility, based on Johnson-Linden Strauss Lemma [11]. They showed that the projection can preserve the inner product, which is directly related to several distance-related metrics, by conducting row wise and column-wise projection of the sample data. These properties guarantee that both the dimensionality and the exacted value of each element of the original data are kept confidential as long as the data and random noise are from a continuous real domain and all involved participating parties are semi-honest.

B. Analysis and Comparison

The data perturbation methods are often evaluated by measuring privacy preservation level and information loss[4]. The privacy preservation level herein indicates the level of difficulty of estimating original data from perturbed data [7]. The information loss refers to the loss of critical information of the original data set after perturbation. TABLE I gives a comparison of three kinds of data perturbation methods as described above. For value-based data perturbation including Random Noise Addition, data values are often perturbed independently. Thus, the information loss depends on the amounts of data values perturbed to achieve a certain privacy preservation level of a specific data-mining task. Compared to the value-based data perturbation methods, the dimension-based methods usually achieve lower information loss because they preserve more statistical information across different dimensions through certain transformations, rotations or projections when the data values are perturbed. But their perturbation algorithms are generally more complicated than the value-based methods.

TABLE I. COMPARISON OF THE DATA PERTURBATION METHODS



	Information Loss	Privacy Preservation Level	Compatibility with Data Mining Methods
Random Noise Addition	Low	Medium	No
Random Rotation	Lower	High	Yes
Random Projection	Lower	High	Yes

Generally, the privacy preservation level of the value based data perturbation methods depends on the reconstruction of the perturbed data values. In this kind of methods, the distribution of the actual data (instead of the individual data values) can be re-constructed from the perturbed data, thus individual data privacy can be well preserved [6]. However, a spectral filter was proposed in [2] to reconstruct some individual data points, which makes the privacy preservation level of the value-based data perturbation methods questionable. The dimension-based methods preserve data privacy in multi-dimensions, instead of a single-dimension conducted by most of the value-based methods. The dimension-based methods (e.g., Random Rotation and Random Projection) are believed to be difficult to attack for extracting individual private information if the attacker has no prior knowledge of the original data set. Another issue needed to be considered when choosing a proper perturbation method is whether the data perturbation methods are compatible with the existing data mining algorithms used in the upper layers of the PPDM framework.

When using the value-based perturbation methods, the data mining algorithms usually need to be modified in order to reconstruct the distribution of the original data and reach a required privacy preservation level at the same time. Thus, this kind of methods normally is not compatible with existing data mining methods. On the other hand, the dimension-based methods keep the statistical information of original data sets, thus they are mostly compatible with the data mining methods.

IV. CONCLUSION

From the discussions of various techniques for privacy preserving techniques for data mining, it is clear that the PPDM is the most important component as far as secured data mining is concerned. Because data mining is being used in different types of applications frameworks like hospitals, telecommunication, Adhar card data mining etc. these types of sensitive data required to be protected from various illegal access. As data mining has three stages one is data collection phase, second is data pre-process stage and last is data mining phase. In all the phases data mining with privacy preservation is required. While collecting the sensitive information from the user data will be coded, also while pre processing the data again it be coded and finally it is secured while data mining purpose.

V. FUTURE WORK

In current time various techniques are being followed for privacy preservation in data mining. These techniques are being applied in different stages of data mining process, like while data collection, data pre-processing and then data mining etc. K-Anonymity will be used for securing the data. This type of technique will be applied while collection the data. So that data will be stored in secured way.

REFERENCES

[1] C. Aggarwal, C. C. Aggarwal, and P.S. Yu, “A condensation approach to privacy preserving data mining,” Advances in Database Technology- EDBT, pp.183–199, 2004.

[2] C. C. Aggarwal, “On randomization, public information and the curse of dimensionality,” 2007 IEEE 23rd International Conference on Data Engineering, pp.136–145, April 2007.

[3] C.C.Aggarwal and P.S.Yu, “On variable constraints in privacy preserving data mining,” in Proceedings of the 2005 SIAM International Conference on Data Mining, pp.115-125, 2005.

[4] C.C.Aggarwal and P. S. Yu, Privacy-Preserving Data Mining: Models and Algorithms, Springer Publishing Company, Incorporated, 2008.

[5] G. Aggarwal, T. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy, D. Thomas, and A. Zhu, “Anonymizing tables,” Database Theory-ICDT 2005, pp.246–258, 2005.

[6] R. Agrawal and R. Srikant, “Privacy-preserving data mining,” in SIGMOD ’00 Proceedings of the 2000 ACM SIGMOD international conference on Management of Data, pp.439–450, May 2000.

[7] K. Chen and L. Liu, “A random rotation perturbation approach to privacy preserving data classification,” in Proceedings of International Conference on Data Mining, 2005.

[8] G. Aggar-wal, T. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy, D. Thomas, and A. Zhu, “Approximation algorithms for k-anonymity,” Journal of Privacy Technology, 20 Nov 2005.

[9] A. Hyvärinen, J. Karhunen, and E. Oja, Independent Component Analysis. Wiley-Blackwell, 2001.

[10] A. Inan, Y. Saygyn, E. Savas, A. Hintoglu, and A. Levi, “Privacy preserving clustering on horizontally parti-tioned data,” 22nd International Conference on Data Engineering Workshops, pp.95-104, 2006.

[11] W. Johnson and J. Lindenstrauss, “Extensions of Lipschitz mappings into a Hilbert space,” Conference in modern analysis and probability (New Haven, Conn., 1982), vol.26 of Contemporary Mathematics, pp.189–206. American Mathematical Society, 1984.

[12] H. Kargupta, S. Datta, Q. Wang, and K. Sivakumar, “Random-data perturbation techniques and privacy-preserving data mining,”

Knowledge and Information Systems, vol.7, pp.387-414, May 2005.

[13] M. Keyvanpour and S. S. Moradi, “Classification and evaluation the privacy preserving data mining techniques by using a data modificationbased framework,” International Journal on Computer Science and Engineering, vol.3, pp.862-870, Feb 2011.

[14] N. Li, T. Li, and S. Venkatasubramanian, “t-closeness: Privacy beyond k-anonymity and l-diversity,” in IEEE 23rd International Conference on Data Engineering, pp.106-115, April 2007.

[15] Y. Li, M. Chen, Q. Li, and W. Zhang, “Enabling multilevel trust in privacy preserving data mining,” IEEE Transactions on Knowledge and Data Engineering, vol.24, pp.1598-1612, 2012.