ISSN 2395-1621

Privacy Preserving Data Mining

[#1]Paridhi D. Agrawal, [#2]Monika V. Dongre

[1]paridhiagrawal17@gmail.com
[2]monika.dongre22@gmail.com

[#12]Department of Computer Engineering

Savitribai Phule University, Pune, India.

ABSTRACT

Data mining has been widely studied and applied in many fields such as Financial Data Analysis, Intrusion Detection, Telecommunication and Retail Industry, etc. However, data mining techniques occur serious challenges due to sensitive information disclosure and privacy violation. Privacy-Preserving Data Mining (PPDM) plays a very important role of performing data mining operations on private data and then forwarding data in a secured way in order to maintain the sensitivity of data. Thus, it is a very important branch of Data Mining. In addition to Information Extraction and Information Retrieval from large amounts of data, PPDM also protects private and sensitive data disclosure without the permission of data owners or providers. This paper reviews main PPDM techniques and compare the merits and demerits of different PPDM techniques. We will also discuss different issues and future research trends in PPDM.

ARTICLE INFO

Article History

Received: 2nd January 2017

Received in revised form :

2nd January 2017

Accepted: 5th January 2017

Published online :

15th January 2017

Keywords: Privacy Preserving Data Mining (PPDM), Cryptographic, Condensation.

I. INTRODUCTION

Nowadays a huge amount of data are being produced. This data are stored in a central repository for organized data called Data warehouse. Data mining deals with the extraction of useful information from such repositories and representation of this information in the form of patterns, rules, clusters or classification models i.e. the process of data mining starts with the gathering of data to discovery of knowledge. Customer relationship management, market basket analysis, healthcare, intrusion detection are a few of the many applications of data mining. During this whole process of data mining sensitive information about individuals, such as medical and financial information often gets exposed to several unauthorized users, including collectors, owners, miners, etc. Much advancement are made in the field of data mining and techniques have been proposed for preserving privacy without compromising data security.

Privacy Preserving Data Mining (PPDM) aims at maintaining the ratio between privacy protection and knowledge discovery. PPDM is related to Secure Multi-party Computation (SMC) i.e. sharing information among multiple parties without compromising the privacy. On one side, the data provider is aware about the privacy disclosure and intrusion. On the other side, businessman is concerns about the security of their private information. For example, consumers do not like social networking sites to disclose their private information and businessman does not want to share their business secrets to other partners. Therefore, PPDM plays a very important role in data analysis in such cases. Many algorithms have been proposed in the literature for the same.

In this paper, we review main techniques for Privacy Preserving Data Mining. Based on comparison and analysis of these techniques, we further discuss different issues and future research trends. Rest of the paper is organized as follows. Section 2 presents the PPDM framework and introduces PPDM techniques. Section 3 discusses PPDM methods in brief. Section 4 compares different PPDM methods and discuss their merits and demerits. Section 5 includes the future research trends and conclusion.

II. PRIVACY PRESERVING DATA MINING

The framework for PPDM is basically divided into three parts:

1. Collection of raw data.
2. Data mining algorithm.
3. Representation of Result in the form of patterns and graphs.

In the data mining process the data are collected from various sources and stored in databases. It is then transformed to a suitable format and stored in the data warehouse. Once this is done data mining algorithms are applied for information generation/knowledge discovery. During this process the privacy and sensitivity of data is given priority.

There are various methods which can be used for achieving Privacy preserving data mining like Random data perturbation techniques, cryptographic algorithms, condensation methods etc. The main reason behind using these techniques is to render a trade-off among accuracy and privacy. Each of these methods has their own merits and demerits as discussed below.

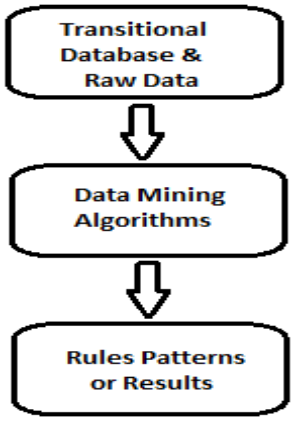

Fig 1. Privacy Preserving Data Mining Framework.

III. Privacy Preserving Data Mining Methods

PPDM techniques may be classified into following categories.

1. Data Perturbation
2. Anonymization Technique
3. Randomized Response
4. Condensation Method
5. Cryptography
6. PPDM with Distributed Data Mining Model.

A. Data Perturbation

Perturbation is the most easy and effective way of privacy preserving data mining. In this method for PPDM the original value is replaced by some arbitrary value. While making these changes, statistical property of this value is maintained. Thus, the results computed by using this perturbed value do not differ from the original to a large extent. Using this technique the attackers are not able to recover the sensitive information, thereby maintain the privacy of an individual.

There are basically two ways in which data perturbation can be performed.

1. Probability distribution approach - This type of perturbation is done by replacing the original value of the same distribution sample.
2. Value distortion approach – This type of perturbation is done by adding external noise to the data.

B. Anonymization Technique

Anonymization technique hides sensitive information about the person/organization from the dataset so that privacy is maintained. There are basically four types of attributes:-

1. Explicit Identifiers: This identifier contains set of attribute that can identify owner record explicitly.
2. Quasi Identifiers: This identifier contains set of attribute that can identify owner record when combined with publicly available data.
3. Sensitive Identifiers: This identifier contains set of attribute that have persons sensitive information.
4. Non-Sensitive Identifier: This identifier contains set of attribute such that even if they are revealed to unauthorized user it will cause no problem.

Linking attack is the most common attack suffered by anonymization based PPDM. When quasi identifiers are linked with publicly available data causes the linked attack. In order to prevent this k-anonymity model was proposed by Latanya Sweeney[12]. Using this model each individual is distinguishable from at least k-1 individual others with respect to quasi identifier attribute in any dataset. Even though this technique is immune to linking attack it fails to protect background knowledge attack and homogeneity attack practically. The main disadvantage of this method is a heavy data loss suffered during data transformation [13].

C. Randomized Response

This method of randomized response for PPDM was first proposed by S. L. Warner in 1965. The model for randomized response contains two entities the sender and the receiver. The sender randomizes their data and transmit it to the receiver, whereas the receiver reconstruct the original data using a reconstruction algorithm for a central or distributed system as per once requirement. The data from each sender is twisted and if the number of sender is large, the aggregate of this information is estimated.

This technique does not require a trusted server as the manipulation are done at the time of data collection. The main disadvantage of using this technique is that it treats each source equally irrespective of their local density. There are many algorithms developed using this technique one such is discussed in [10].

D. Condensation Method

Condensation based PPDM was first introduced by Charu C. Aggrawal and Philip S. Yu[11]. This technique uses condensed static property of clusters to generate pseudo data i.e. condensation based PPDM works on pseudo data besides modifying the original data. As the original data is untouched the privacy achieved using this technique is maximized. The format for pseudo data is kept same as the original data so that there is no need for redesigning.

E. Cryptography

Cryptography is the process of hiding data and then transmitting it in such a way that only intended user is able to read and modify. This technique is basically used for

establishing communication between two parties in the presence of the third party. Zhiquiang Yang, Sheng Zhong and Rebecca N. Wright[15] propose cryptography technique based on horizontal partition and vertical partition which is discussed further in this paper. This technique maintains the privacy without the cost of accuracy. The efficiency of this technique decreases as the number of parties involved in data sharing increases. There is a vast set of cryptography algorithm available in this domain. This algorithm is mainly famous for two important reasons, firstly, well defined model for privacy and secondly, huge toolsets are available for this algorithm. The only major disadvantage with this algorithm is that it slows down when huge dataset is involved.

F. PPDM with Distributed Data Mining Model

Data Mining refers to knowledge discovery from single dataset. Distributed Data Mining refers to knowledge discovery from dataset which is distributed across different resources. Distributed data are mainly of two types:

1. Horizontal distribution: In this type of distribution different records lies in different places.
2. Vertical distribution: In this type of distribution values of different attributes lies in different places.

Distributed data mining is very useful when multiple parties come together and share their information for some research. To give an example, Cancer hospitals in Mumbai came together and share patient's information for joint research. The source of data is distributed, but PPDM algorithms are used in order to maintain the privacy of patients.

IV. EVALUATION CRITERIA, MERITS AND DEMERITS OF PRIVACY PRESERVING DATA MINING

Ideally there is no algorithms which guarantees 100% privacy. Every algorithm have their own advantages and disadvantages. One may perform better when data utility is concerned, whereas others may perform better when performance is taken into consideration. There are various parameters which are used for the evaluation of privacy preserving algorithms. Some of them are listed below:

1. Performance Measure: Performance measure calculates the time required for attaining privacy criteria.
2. Data Utility: Data utility is the measure of data loss suffered due to privacy preserving algorithm.
3. Tolerance: Resistance shown by privacy preserving algorithms against various data mining models.
4. Quantification of privacy: It is privacy matrix that measure how closely the original value of attributes can be measured.
5. Information loss: Information loss is termed as lack of precision for original data estimation.

The algorithm which is able to achieve balance between privacy and information loss is most effective when privacy preserving is concerned. Randomized response provides better efficiency, then Cryptography based PPDM. Whereas Cryptography based algorithm provides better privacy than any other algorithm. All other techniques studied so far suffers heavy loss of data except cryptography based approach for PPDM. Charu Agrawal and Dakshi Agrawal in [14] have made the comparison of various algorithms with respect to privacy and information loss in brief.

V. CONCLUSION

The main aim of privacy preserving data mining is to develop an algorithm to provide privacy to user's sensitive data. There are many algorithms as discussed above proposed for the same. Privacy and accuracy are a pair of ambiguity success in one leads to compromise in another. There is no such algorithm which works best for all evaluation criteria like performance measure, cost, resistance, etc. One need to choose an algorithm as per their own requirement and specification.

REFERENCES

[1] Rajesh N, Sujatha K and A. Arul Lawrence, "Survey on Privacy Preserving Data Mining". International Journal of Computer Applications, January 2016.

[2] Charu C. Aggrawal and Philip S. Yu, "Privacy Preserving Data Mining Models and algorithms". Advances in database systems, 2008.

[3] Hina Vaghashia and Amit Ganatra, "A Survey: Privacy Preserving Techniques in Data Mining". International Journal of Computer Applications, June 2015

[4] Ronica Raj, Veena Kukarni, "A Study on Privacy Preserving Data Mining: Techniques, Challenges and Future Prospects". International Journal in Computer Engineering, Vol. 3, November 2015.

[5] S. Selva Ratna, Dr. T. Karthikeyan, "Survey on Recent Algorithms for Privacy Preserving Data Mining". International Journal of Computer Science and Information Technologies, 2015.

[6] M. Keyvanpur and S. S. Moradi, "Classification and evaluation the privacy preserving data mining techniques by using a data modification based framework". International Journal on Computer Science and Engineering, February 2011.

[7] Y. Li, M. Chen, Q. Li and W. Zhang, "Enabling multilevel trust in privacy preserving data mining". IEEE TRANSACTION ON KNOWLEDGE AND DATA ENGINEERING, 2012.

[8] Grljevic O., Bosnjak Z., Mekovec R., "Privacy Preserving in Data Mining- Experimental research on SMEs data". IEEE 9th International Symposium on Intelligent Systems and Informatics (SISY), 2011.

[9] D. Mittal, D. Kaur, A. Aggarwal, "Secure Data Mining in Cloud using Homomorphic Encryption". IEEE

International Conference on Cloud Computing in Emerging Markets (CCEM), 2014.

[10] A. S. Shanthi and Dr. Karthikeyan, "A Review on Privacy Preserving Data Mining".

[11] Charu C. Aggrawal and Philip S. Yu, "A Condensation Approach to Privacy Preserving Data Mining". Springer, 2004.

[12] Lantanya and Sweeney, "Achieving k- anonymity privacy protection using generalization and suppression", International journal, Vol. 10, 2002.

[13] Gayatri Nayak, Swagatika Devi, "A Survey on Privacy Preserving Data Mining: Approaches and Techniques". International Journal of Engineering Science and Technology, Vol. 3, 2011.

[14] D. Agrawal and C. Agrawal, "On the Design and Qualification of Privacy Preserving Data Mining Algorithms". 2001.

[15] Zhiqiang Yang, Sheng Zhong and Rebecca N. Wright, "Privacy Preserving Classification of Customer Data without loss of Accuracy". International Conference on Data Mining, 2005.

