International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 3 | ISSN : 2456-3307

Preserving Privacy in Data Mining

Hemlata*
Department of Computer Science, Maharshi Dayanand University, Rohtak, Haryana, India

ABSTRACT

Data Mining means the process of deriving new knowledge, rules and patterns from the existing database. Knowledge is extracted from unstructured, large amount of data for analysis. The process is also known as knowledge discovery. The derived data or the patterns provide valuable information in decision-making process and business strategy. The results of Data Mining should not reveal the sensitive data of users. The Data Mining techniques for preserving the privacy of data from malicious users are termed as Privacy Preserving Data Mining Techniques. This paper provides a review of privacy preserving Data mining architecture. It also presents the bands of Privacy Preserving Objectives. Approaches of Privacy preserving are summarized in the paper. This paper is intended for the researchers and scientists who work in the field of Privacy preserving Data Mining.

Keywords: Data Mining, Privacy Preserving Data Mining, PPDM Framework.

I. INTRODUCTION

Data Mining refers to the process of extracting meaningful knowledge from the unstructured data. It is the procedure of discovering new knowledge from the previously unknown data. Finding new meaning and pattern from the existing database can be termed as Data Mining.

While mining the data from the database, there exist many problems and challenges. Most important challenge is the protection of private data of user from the malicious users. The process of preserving the privacy of data while mining the data is known as privacy preserving data mining. It is the process of protecting the user data from misuse.

This paper elaborates the concept of Privacy preserving data mining. The paper is a detailed review of privacy preserving and its framework. The steps of privacy preserving are reviewed and explained in the form of bands. It can give a direction to the researchers and data miners how to extract knowledge without sacrificing the sensitivity and privacy of data.

The organization of this paper is as follows. In Section 2 (**Bands of Privacy Preserving Objectives**), the bands or steps of privacy preserving are explained. Section 3 (**Approaches of Privacy Preserving Data Mining**), presents various approaches or methods of privacy preserving. In Section 4 (**A PPDM Framework**), the framework or the architecture of PPDM is elaborated. Section 4 (**Conclusion**) conclusion and future scope is presented.

II. BANDS OF PRIVACY PRESERVING OBJECTIVES(BPPO)

Different bands of privacy preserving objectives means why privacy preserving is required in different dataset. For this requirement BPPO is designed (Figure 1).

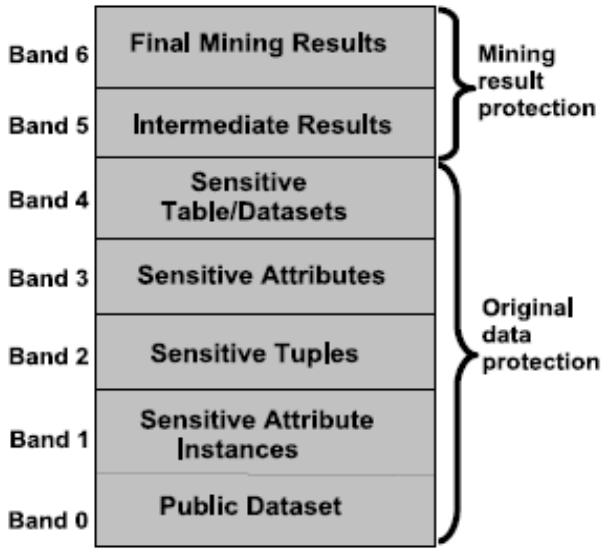

Fig 1: Bands of privacy preserving Objectives (BPPO)

The bands in figure 1 are divided in two groups:

1) **Original Data Protection**: It includes Sensitive Attribute Instances, Sensitive Tuples, Sensitive Attributes and Sensitive Datasets.
2) **Mining Result Protection**: It includes Intermediate Mining results and Final Mining results.

The objectives of privacy preservation of data mining algorithms are classified in the following bands:

1) **Band 0** - *Public Database Protection*: The data mining algorithms of the centralised database where the data is not scattered in different parties lies in this band.
2) **Band 1** - *Sensitive Attribute Instances Protection:* The privacy preserving data mining algorithms should protect the sensitive or private attribute values. Methods to achieve it are- modify data, encode data, swap data etc.
3) **Band 2** – *Sensitive Tuples Protection*: This band preserves the sensitive records or tuples. Horizontal partitioning of the dataset is considered under this band. Techniques used to sanitize data in this band are- distributed dataset techniques, randomization techniques, mathematical transformation techniques etc.
4) **Band 3** – *Sensitive Attributes protection:* The objective of this band is to protect the sensitive column or attribute. When the dataset is partitioned vertically, the data owners protect or hide the sensitive attribute.
5) **Band 4** – *Sensitive Table or Dataset protection:* This band protects the whole sensitive table or the sensitive dataset. Normally distributed dataset is involved in the band.
6) **Band 5** – *Intermediate Data Mining result Protection:* This band preserves or protects the data mining intermediate results. In other words, when data mining is done in steps, few intermediate data is not required to be shared among all the participants. The mid way data is protected in this band.
7) **Band 6** – *Final Mining Result Protection*: Protection of the results of data mining process are done in this band.

III. APPROACHES OF PRIVACY PRESERVING DATA MINING

While data mining the approaches to be followed are: -

1) Data Distribution: Privacy preserving of distributed dataset is more complex as compared to centralised dataset. There are two approaches of data distribution:
   a. Horizontal Partitioning- The dataset is partitioned in such a way that each site has all attributes but does not have all tuples.
   b. Vertical Partitioning- The dataset is partitioned in such a way that each site has all tuples but does not have all attributes.
2) Data Modification: For the security and privacy of original data it is modified by the owner. Data modification methods are:
   a. Perturbation: In this method, noise/ dummy data is added to the value of an attribute.
   b. Blocking: The value of the attribute is replaced by a symbol like "?".
   c. Aggregation: Several values are merged to a broader category.
   d. Swapping: The values of the records are swapped in order to protect privacy.

e. Sampling: Only a sample of data is released to the public.

3) Data Mining Algorithm: There are many data mining algorithms for mining or extracting knowledge form the database. These algorithms must include a technique of preserving the sensitive data from malicious users. Various existing algorithms are- Decision Tree, Association Rules, Clustering, K-means, etc.

4) Data or Rule Hiding: Data hiding means hiding the original or raw data. But large amount of data hiding leads to incorrect knowledge. So, there should be a trade-off between releasing and hiding of proper amount of data.

5) Privacy preservation: This method preserves the data by modifying the selected data. Existing Privacy Preserving methods are: which exists Different existing techniques of privacy preservation can be categorised as:
   a. Heuristic-based techniques
   b. Cryptography-based technique
   c. Reconstruction based technique

IV. PPDM FRAMEWORK

Privacy Preserving Data Mining Framework explains the overall properties of Data Mining. The framework is 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 2.

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.

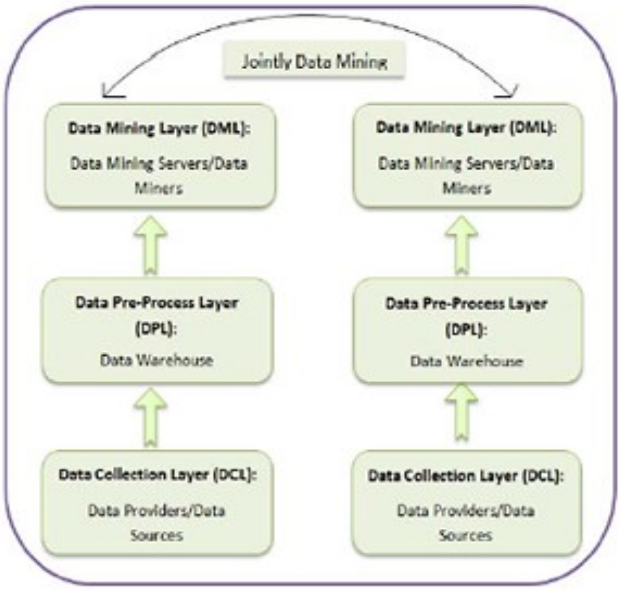

Fig 2: PPDM Framework

The second layer DPL contains data warehouse servers

which are responsible for storing and pre-processing the collected raw data from the data providers. The raw data stored in the data warehouse servers can 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 pre-processing for later data mining, and the other is the security of data access.

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.

V. CONCLUSION

Privacy Preserving Data Mining is the process of obtaining knowledge from the existing large dataset. The paper summarises the meaning and characteristic process of Privacy Preserving Data Mining. It also elaborates the approaches of Privacy Preserving Data Mining. The approaches are very useful on full proof Privacy of personal data. Privacy Preserving Data Mining Framework is depicted for better understanding of Privacy Preserving.

VI. REFERENCES

1. K Thearling, "Data Mining and Privacy: A Conflict in Making", DS, November 1998.
2. R Agrawal and R. Srikant. "Privacy Preserving Data Mining", ACM SIGMOD Conference on Management of Data, pp: 439-450, 2000
3. Y Lindell and B. Pinkas, "Privacy Preserving Data Mining", Journal of Cryptology, 15(3), pp.36-54, 2000.
4. C Clifton, M. Kantarcioglu, and J. Vaidya. "Defining Privacy for Data Mining", Proceedings of the National Science Foundation Workshop on Next Generation Data Mining, 2002. pp. 126-133.
5. Hemlata, Gulia, P. (2018). DCI3 Model for Privacy Preserving in Big Data. In Big Data Analytics (pp. 351-362). Springer, Singapore.
6. S R. M. Oliveira and Osmar R. Zaiane, "Toward Standardization in Privacy-Preserving Data Mining", DMSSP 2004 (In conjunction with SIGKDD 2004).
7. D Agrawal and C. Aggarwal, "On the Design and Quantification of Privacy Preserving Data Mining Algorithms", PODS 2001. pp: 247-255.
8. A Evfimievski, R. Srikant, R. Agrawal and J. Gehrke, "Privacy Preserving Mining of Association Rules", SIGKDD 2002. pp. 217- 228
9. Hemlata, Gulia, Preeti. "Novel Algorithm for PPDM of Vertically Partitioned Data." International Journal of Applied Engineering Research 12.12 (2017): 3090-3096.
10. S. Rizvi and J. Haritsa, "Maintaining Data Privacy in Association Rule Mining", VLDB 2002. pp: 682-693.
11. W. Du and Z. Zhan, "Using Randomized Response Techniques for Privacy Preserving Data Mining", SIGKDD 2003. pp. 505-510.
12. S. Agrawal and J. Haritsa, "On Addressing Efficiency Concerns in Privacy-Preserving Mining", DASFAA 2004. pp. 113-124.
13. Hemlata and Dr. Preeti Gulia, "Techniques and Algorithms of PPDM", International Journal for Scientific Research & Development, Vol. 3, Issue 04, 2015, pp. 3484-3487.
14. Stanley, R. M. O. and R. Z Osmar, "Towards Standardization in Privacy Preserving Data Mining", Published in Proceedings of 3rd Workshop on Data Mining Standards, WDMS' 2004, USA, p.7-17.
15. S.V. Vassilios , B. Elisa, N.F. Igor, P.P. Loredana, S. Yucel and T. Yannis, 2004, "State of the Art in Privacy Preserving Data Mining" Published in SIGMOD Record, 33, 2004, pp: 50-57.
16. Elisa, B., N.F. Igor and P.P. Loredana. "A Framework for Evaluating Privacy Preserving Data Mining Algorithms", Published by Data Mining Knowledge Discovery, 2005, pp.121-154.
17. Philip Chan, "An Extensible Meta Learning Approach for scalable and Accurate Inductive Learning", PhD Thesis, Department of Computer Sciences, Columbia University, New York, NY, 1996
18. Philip Chan, "On the accuracy of meta-learning for scalable data mining". Journal of intelligent Information Systems, 8:5-28, 1997.
19. Andreas Prodromidis, Philip Chan, and Salvatore Stolfo, : "Metalearning in distributed data mining systems: Issues and approaches". In "Advances in Distributed and Parallel

Knowledge Discovery", AAAI/MIT Press, September 2000.

20. S.V. Vassilios , B. Elisa, N.F. Igor, P.P. Loredana, S. Yucel and T. Yannis, 2004, "State of the Art in Privacy Preserving Data Mining" Published in SIGMOD Record, 33, 2004, pp: 50-57.
21. Gayatri Nayak, Swagatika Devi, "A survey on Privacy Preserving Data Mining: Approaches and Techniques", International Journal of Engineering Science and Technology, Vol. 3 No. 3, 2127-2133, 2011.
22. Aggarwal C, Philip S Yu, "A General Survey of Privacy-Preserving Data Mining Models and Algorithms", Springer Magazine, XXII, 11-52, 2008.
23. J. Vaidya, C. Clifton, "Privacy Preserving Association Rule Mining in Vertically Partitioned Data", In Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 639–644, 2002.