Privacy Preserving Data Mining: A Detailed Survey

Vidisha M. Pradhan
Assistant Professor, Dept of Information Technology
VIE, Affiliated to G.T.U., Kotambi, Gujarat, India

Abstract- *Data mining is one of the most important topics in current scenario. There are various filed in data mining which needs focus of interest. One of these fields is privacy preserving data mining. With the increase of online data and analysis of online data, it is necessary to concern about the privacy of important data. Privacy preserving data mining offers a vast variety of algorithms which maintains the confidentiality of sesitive data and analysis only those data which is useful for mining task. In this paper, different types of privacy preserving data mining is discussed with its advantages as well as disadvantages.*

Keywords- Data Mining, Privacy Preserving, Sensitive attribute, Association Rule Mining

I. INTRODUCTION

Data mining is a field in which data extraction and knowledge mining is done. Data mining is the process of discovering the intensive knowledge from large amounts of data stored either in databases, data warehouses, or other information repositories. It is an interdisciplinary field and it involves an integration of techniques from multiple disciplines such as database and data warehouse technology, statistics, machine learning, pattern recognition, information retrieval, and spatial or temporal data analysis[3].

There are various levels in which the privacy of data is concerned. All these levels are having different hierarchical abstraction. There are various applications of privacy preserving data mining (PPDM). The most important use of PPDM is in health care area. It maybe possible that a hospital is providing patient's data to some IT professional for analysis task which can be useful for all medical personals, then its is necessary to take care of patient's private data like age, address, contact number etc. The mining task should be performed in such a way that sensitive data should not be released.

Three main levels in which PPDM algorithms must be applied to protect data integrity are shown in the figure. In the first level randomization and anonymization algorithms are used to preserve the privacy of data. In second level, crypto graphical techniques are applied As per level three, each company has to will analyze its own data by applying privacy preserving algorithms which will further result in rules. And at

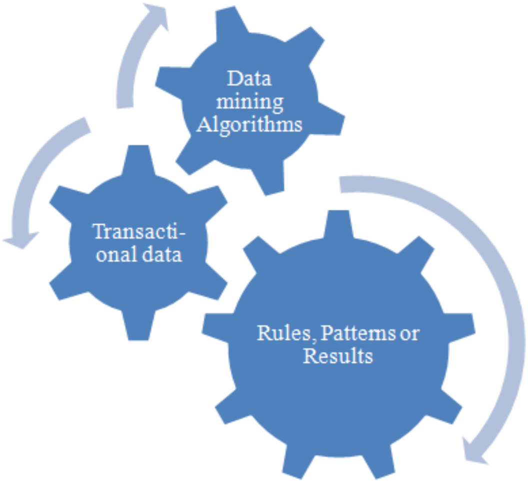

Fig. 1 Levels of Privacy Preserving Algorithms

And at the end, each company will share sanitizes view of its own dataset. Each end user may have different privacy concern while sharing the data. So the end user oriented privacy preserving data mining must be investigated[4]

Fig. 2 represents the taxonomy of privacy preserving data mining. It shows that, PPDM is related to various other disciplines also.

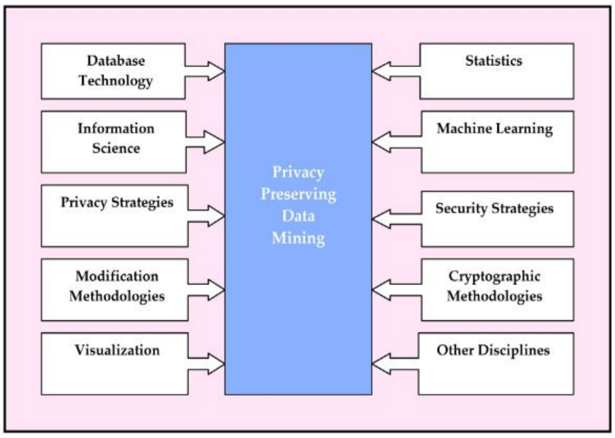

Fig. 2 Taxonomy of Privacy Preserving Data Mining

The rest of the paper is organized in the following way. Section 2 gives the overview of related work in the field of privacy preserving data mining. Section 3 contains conclusion and future work.

II. RELATED WORK

In [5] association rule hiding algorithms is applied to preserve the privacy of data. Two main approaches (Distortion & Blocking) for association rule hiding were discussed with its pros & cons. In this paper various communities' related data were analyzed.

Paper [6] has information related to all types of privacy preserving data mining techniques. The techniques which are studied mentions classification, clustering and association rule mining. Advantages and disadvantages of various data mining techniques are explained.

The paper [7] proposes a new algorithm for privacy preserving using k-means algorithm which is secure but it does not use cryptography. It uses multi party additive scheme. The work was carried out on horizontal partitioned data. The algorithm developed by [7] can tackle adversary's passive model.

[8] proposes advancement in mining association with secrecy constraints (MASK). Data perturbation technique used by MASK algorithm provides low degree of privacy preservation. The execution time required by MASK algorithm is very large so it can't be applied practically in the field of privacy preservation. So to overcome the problems of MASK algorithm, this paper proposes advancement in MASK, which contains data perturbation and query restriction (DPQR). For solving problem related to time, calculation to obtain inverse matrix is divided into blocks and set theory is used to reduce number of scanning in database. This new algorithm is evaluated theoretically as well as practically and hence it was proved that this advancement has better performance.

The system which is proposed in [9] consists of access control mechanism and privacy protection mechanism. Privacy requirement policy is applied on sensitive table i.e. database. Anonymization method was applied to allow only authentic users to see the data. Processing speed was increased by only showing limited rows from table. As location based method is less complicated, it is used for access control which restricts anonymous users to view data. The proposed strategy is capable of handling linking attack as well as it provides more accurate data. This technique can be applied in distributed environment.

The paper [10] presents ant colony system based algorithm known as ACS2DT (ant colony system to delete transaction), which reduces the side effect of sanitization process. This paper contains algorithm for each operator also. ACS2DT works like tradition ACS process which contains state transaction rule, pheromone updating rules and selection process to select particular route. There is no termination condition set by ACS algorithm to stop ant to go further as ant completes its route because no other nodes are selected or it reaches final destination. But in ACS2DT routing graph is defined very specifically and because of predefined termination conditions in those graph, ants are guided towards their destination. A very important heuristic function is generated in ACS2DT for hiding sensitive data from the item-set.

For providing security to sensitive data in the database, concept of cryptography is used in [11]. In this paper multi-party computation of privacy preserving is reduced to 2-party case. There are various advantages (like Trust, Independence of inputs, Communication, Privacy, Correctness, Efficiency, Guaranteed output delivery and Fairness) of using this kind of reduction. This new approach achieved very good results. This paper contains applications of using crypto graphical concept in the field of securing sensitive data in database.

Most of the privacy preserving algorithms assume that quasi-identifiers can be separated from sensitive attributes. But most of the attributes contains features which are sensitive attributes and quasi identifiers as well. [12] contains a new privacy model as well as a method that can treat sensitive quasi identifiers. This method contains two separate algorithms in which the first is anonymization algorithm and the second is reconstruction algorithm. This novel method is experimentally tested using real world data sets and the result of the test proves that this proposed algorithm can anonymize databases as well as reconstruct existing databases while focusing on keeping a high quality of data within a realistic periods.

III. APPLICATIONS OF PRIVACY PRESERVING DATA MINING

The area of privacy preserving has a lot of applications in:

- Medical Databases
- Bioterrorism Applications
- Genomic Privacy
- Homeland Security Applications

IV. CONCLUSION

Now-a-days data mining has become very popular because of its various applications. But as the database becomes available, for all the parties who are concern with data mining task, the privacy of personal data violates. So to control this violation privacy preserving techniques must be applied. In this paper detailed survey on privacy preserving is done. Each and every technique contains its advantages as well as some disadvantages. So conclusion from the survey is, as per the level as well as type of privacy is required for data mining task various different techniques should be applied on data warehouse.

REFERENCES

[1] Jiawei Han, Micheline Kamber and Jian Pei, “Data mining Concepts and Techniques”, Third Edition, Morgan Kaufmann Series in Data management Systems

[2] Charu C. Aggarwal, “Data Mining: The Textbook”, Springer, 2015

[3] K.Saranya, K.Premalatha, S.S.Rajasekar, "A Survey on Privacy Preserving Data Mining", IEEE Sponsored 2nd International Conference On Electronics And Communication System (Icecs 2015)

[4] A.S.Shanthi, Dr. M. Karthikeyan, "A Review on Privacy Preserving Data Mining", IEEE, 2012

[5] S.Vijayarani, Dr.A.Tamilarasi, R.SeethaLakshmi, "Privacy Preserving Data Mining Based on Association Rule- A Survey", Proceedings of the International Conference on Communication and Computational Intelligence – 2010

[6] K.Saranya, K.Premalatha,S.S.Rajasekar, "A Survey on Privacy Preserving Data Mining", IEEE Sponsored 2nd International Conference on Electronics and Communication System, 2015

[7] Zakaria Gheid, Yacine Challa, "Efficient and Privacy-Preserving k-means clustering For Big Data Mining", IEEE TrustCom/BigDataSE/ISPA, 2016

[8] Haoliang Lou, Yunlong Ma, Feng Zhang, Min Liu, Weiming Shen, "Data Mining for Privacy Preserving Association Rules Based on Improved MASK Algorithm", Proceedings of the 2014 IEEE 18th International Conference on Computer Supported Cooperative Work in Design

[9] Ahmed H I Lakadkutta, R V Mante, "Location Based Privacy Preserving Access Control for Relational Data", IEEE International Conference On Recent Trends In Electronics Information Communication Technology, May 20-21, 2016

[10] Jimmy Ming-Tai Wu, Justin Zhan, And Jerry Chun-Wei Lin, "Ant Colony System Sanitization Approach to Hiding Sensitive Itemsets", IEEE, 2017

[11] Anand Sharma, Vibha Ojha, "Implementation Of Cryptography For Privacy Preserving Data Mining", International Journal of Database Management Systems (IJDMS) Vol.2, No.3, August 2010

[12] Yuichi Sei, Hiroshi Okumura, Takao Takenouchi, Akihiko Ohsuga, "Anonymization of Sensitive Quasi-Identifiers for l-Diversity and t-Closeness", IEEE Transactions on Dependable and Secure Computing, 2017