Techniques and Algorithms of PPDM

Hemlata[1] Dr. Preeti Gulia[2]

[1]Research Scholar [2]Assistant Professor

[1,2]Department of Master of Computer Application

[1,2]M.D.University, Rohtak

Abstract— Privacy preserving data mining (PPDM) is a method of protecting the privacy of data without sacrificing the utility of the data. In this present world of Internet people have become well aware that they should not share their personal data and sensitive information. This may lead to the negative results of the data mining. This paper provides a detailed review of different techniques and algorithms used in Privacy Preserving Data Mining (PPDM). This paper also presents an overview of PPDM Framework, PPDM techniques, PPDM tools and algorithms.

Key words: Data Mining, Privacy Preserving Data Mining, PPDM Framework, PPDM Algorithm

I. INTRODUCTION

Data mining refers to the techniques of extracting rules and patterns from data. It is also commonly known as KDD (Knowledge Discovery from Data).[2] In data mining knowledge are extracted through different technique such as classification, clustering, association etc. [1] Traditional data mining operates on the data warehouse model of gathering all data into a central site and then running an algorithm against that warehouse. This model works well when the entire data is owned by a single party who generates and uses a data mining model without disclosing the results to any third party [2]. The extracted knowledge patterns can provide insight to the data holders as well as be invaluable in tasks such as decision making and strategic business planning [3]. As a valuable technique, data mining is developing and is flourishing. But, at the same time, serious concerns have grown over individual privacy in data collection, processing and mining [4]. As a result, preserving privacy is a serious issue in data mining. It is perceived that data mining, with its promise to efficiently discover valuable, non-obvious information from large databases, is particularly vulnerable to misuse [5, 6]. So, there might be a conflict between data mining and privacy. But, in reality, there is no need to intrude privacy for data mining. The objective of data mining is to generalize across populations, rather than reveal information about individuals. But, the problem is that data mining works by evaluating individual data. Consequently, there might be privacy intrusion. Hence, the true problem does not lie with data mining, but with the way data mining is done. The goal of preserving privacy in data mining is to lower the risk of misuse of data and at the same time produce results same as that produced in the absence of such privacy preserving techniques.

II. PPDM FRAMEWORK

The framework for PPDM is shown in fig. 1. In data mining or knowledge discovery from databases (KDD) process the data (mostly transactional) is collected by single/various organization/s and stored at respective databases. Then, it is transformed to a format suitable for analytical purposes, stored in large data warehouse/s and then data mining algorithms are applied on it for the generation of information/knowledge. With the intent of protecting privacy the model has to be evolved.[1]

The figure given below suggests three levels where privacy concerns are taken care of. At level 1, the raw data collected from a single or multiple databases or even data marts is transformed into a format that is well suited for analytical purposes. Even at this stage, privacy concerns are needed to be taken care of. Researchers have applied different techniques at this stage but most of them deal with making the raw data suitable for analysis.

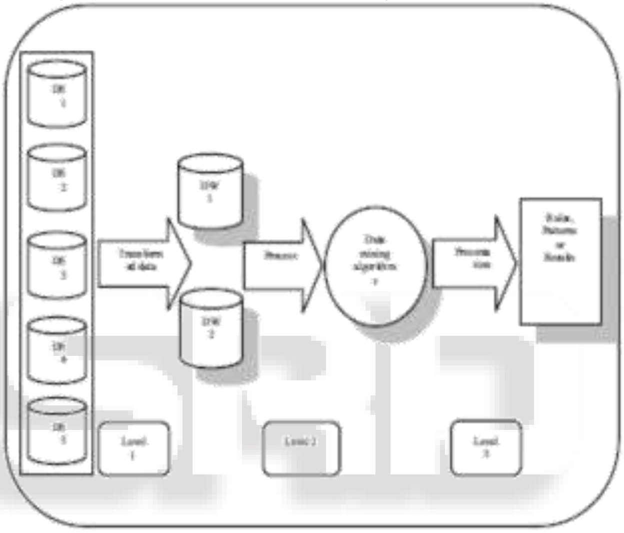

Fig. 1: Framework of PPDM

At level 2, the data from data warehouses is subjected to various processes that make the data sanitized so that it can be revealed even to untrustworthy data miners. The processes applied at this stage are blocking, suppression, perturbation, modification, generalization, sampling etc. Then, the data mining algorithms are applied to the processed data for knowledge/information discovery. Even the data mining algorithms are modified for the purpose of protecting privacy without sacrificing the goals of data mining. At level 3, the information/knowledge so revealed by the data mining algorithms is checked for its sensitiveness towards disclosure risks. We have described the embedding of privacy concerns at three levels, but any combination of these may be used.

III. PPDM TECHNIQUES[1]

As a research direction in data mining and statistical databases, privacy preserving data mining received substantial attention and many researchers performed a good number of studies in the area [7-15]. Since its inception in 2000 with the pioneering work of Agrawal & Srikant [7] and Lindell & Pinkas [8], privacy preserving data mining has gained increasing popularity in data mining research community. PPDM has become an important issue in data mining research [16-18]. As a result, a whole new set of approaches were introduced to allow mining of data, while at the same time prohibiting the leakage of any private and sensitive information. The majority of the existing approaches can be classified into two broad categories [3]**:**

1) methodologies that protect the sensitive data itself in the mining process, and
2) methodologies that protect the sensitive data mining results (i.e. extracted knowledge) that were produced by the application of the data mining.

The first category refers to the methodologies that apply perturbation, sampling, generalization / suppression, transformation, etc. techniques to the original datasets in order to generate their sanitized counterparts that can be safely disclosed to untrustworthy parties. The goal of this category of approaches is to enable the data miner to get accurate data mining results when it is not provided with the real data. Secure Multiparty Computation methodologies that have been proposed to enable a number of data holders to collectively mine their data without having to reveal their datasets to each other. The second category deals with techniques that prohibits the disclosure sensitive knowledge patterns derived through the application of data mining algorithms as well as techniques for downgrading the effectiveness of classifiers in classification tasks, such that they do not reveal sensitive knowledge. In contrast to the centralized model, the Distributed Data Mining (DDM) model assumes that the data sources are distributed across multiple sites. Algorithms developed within this field address the problem of efficiently getting the mining results from all the data across these distributed sources. A simple approach to data mining over multiple sources that will not share data is to run existing data mining tools at each site independently and combine the results [19-21]. However, this will often fail to give globally valid results. Issues that cause a disparity between local and global results include:

1) Values for a single entity may be split across sources. Data mining at individual sites will be unable to detect cross-site correlations.
2) The same item may be duplicated at different sites, and will be over-weighted in the results.
3) At a single site, it is likely to be from a homogeneous population. Important geographic or demographic distinctions between that population and others cannot be seen on a single site.

PPDM tends to transform the original data so that the result of data mining task should not defy privacy constraints. Following is the list of five dimensions on the basis of which different PPDM Techniques can be classified [22]:

- Data distribution
- Data modification
- Data mining algorithms
- Data or rule hiding
- Privacy preservation

Based on these dimensions, different PPDM techniques may be classified into following five categories [22, 23,24]

- Anonymization based PPDM
- Perturbation based PPDM
- Randomized Response based PPDM
- Condensation approach based PPDM
- Cryptography based PPDM

IV. Tools And Algorithms of PPDM

An overview of some of the commonly used tools and algorithms for PPDM are[2]:

A. *Secure Multi Party Communication*

Almost all PPDM techniques rely on secure multi party communication protocol. Secure multi party communication is defined as a computation protocol at the end of which no party involved knows anything else except its own inputs the results, i.e. the view of each party during the execution can be effectively simulated by the input and output of the party. In the late 1980s, work on secure multi party communication demonstrated that a wide class of functions can be computed securely under reasonable assumptions without involving a trusted third party. Secure multi party communication has generally concentrated on two models of security. The semi-honest model assumes that each party follows the rule of the protocol, but is free to later use what it sees during execution of the protocol. The malicious model assumes that parties can arbitrarily cheat and such cheating will not compromise either security or the results, i.e. the results from the malicious party will be correct or the malicious party will be detected. Most of the PPDM techniques assume an intermediate model, - preserving privacy with non-colluding parties.

B. *Secure Sum*

Distributed data mining algorithms often calculate the sum of values from individual sites. Assuming three or more parties and no collusion, the following method securely computes such a sum.

Let $v = \sum_{i=1}^{s} v_i$ is to be computed for s sites and v is known to lie in the range [0..N]. Site 1, designated as the master site generates a random number R and sends $(R+v_1)\bmod N$ to site 2. For every other site l = 2, 3, 4 … s, the site receives:

$$V = (R + \sum_{j=1}^{l-1} v_j) \bmod N \cdot$$

Site l computes:

$$(V + v_l) \bmod N = (R + \sum_{j=1}^{l} v_j) \bmod N$$

This is passed to site (l+1). At the end, site 1 gets:

$$V = (R + \sum_{j=1}^{s} v_j) \bmod N$$

And knowing R, it can compute the sum v. The method faces an obvious problem if sites collude. Sites (l-1) and (l+1) can compare their inputs and outputs to determine v_l. The method can be extended to work for an honest majority. Each site divides v_l into shares. The sum of each share is computed individually. The path used is permuted for each share such that no site has the same neighbors twice.

C. *Secure Set Union*

Secure set union methods are useful in data mining where each party needs to give rules, frequent itemsets, etc without revealing the owner. This can be implemented efficiently using a commutative encryption technique. An encryption algorithm is commutative if given encryption keys $K_1, K_2, ... K_n \in K$, the final encryption of a data M by applying all the keys is the same for any permuted order of the keys. The main idea is that every site encrypts its set and adds it to a global set. Then every site encrypts the items it hasn't encrypted before. At the end of the iteration, the global set will contain items encrypted by every site. Since encryption technique chosen is commutative, the duplicates will encrypt to the same value and can be eliminated from the global set. Finally every site decrypts every item in the global set to get the final union of the individual sets. One addition is to permute the order of the items in the global set to prevent sites from tracking the source of an item. The only additional information each site learns in the case is the number of duplicates for each item, but they cannot find out what the item is.

D. *Secure Size of Set Intersection*

In this case, every party has their own set of items from a common domain. The problem is to securely compute the cardinality/size of the intersection of these sets. The solution to this is the same technique as the secure union using a commutative encryption algorithm. All k parties locally generate their public key-part for a commutative encryption scheme. The decryption key is never used in this protocol. Each party encrypts its items with its key and passes it along to the other parties. On receiving a set of encrypted items, a party encrypts each item and permutes the order before sending it to the next party. This is repeated until every item has been encrypted by every party. Since encryption is commutative, the resulting values from two different sets will be equal if and only if the original values were the same. At the end, we can count the number of values that are present in all of the encrypted item sets. This can be done by any party. None of the parties can find out which of the items are present in the intersection set because of the encryption.

E. *Scalar Product*

Scalar product is a powerful component technique and many data mining problems can be reduced to computing the scalar product of two vectors. Assume two parties P_1 and P_2 each have a vector of cardinality n, $X = (x_1, x_2, ... x_n)$, $Y = (y_1, y_2, ... y_n)$. The problem is to securely compute $\sum_{i=1}^{n} x_i y_i$. There has been a lot of research and proposed solution to the 2 party cases, but these cannot be easily extended to the multi party case. The key approach to a possible solution proposed in [25] is to use linear combinations of random numbers to disguise vector elements and then do some computations to remove the effect of these random numbers from the result. Though this method does reveal more information than just the input and the result, it is efficient and suited for large data sizes, thus being useful for data mining

F. *Oblivious Transfer*

The oblivious transfer protocol is a useful cryptographic tool involving two parties, - the sender and the receiver. The sender's input is a pair (x_0, x_1) and the receiver's input is a bit $\sigma \in (0,1)$. The protocol is such that the receiver learns x_σ (and nothing else) and the sender learns nothing. In the semi-honest adversaries, there exist simple and efficient protocols for oblivious transfer.

G. *Oblivious polynomial evaluation*

This is another useful cryptographic tool involving two parties. The sender's input is a polynomial Q of degree k over some finite field F (k is public). The receiver's input is an element $z \in F$. The protocol is such that the receiver learns Q(z) without learning anything else about the polynomial and the sender learns nothing.

V. EVALUATION OF PPDM ALGORITHMS

Another important aspect of privacy preserving data mining algorithms is their evaluation against certain parameters like [17]:

- Performance: the performance of a mining algorithm is measured in terms of the time required to achieve the privacy criteria.
- Data Utility: Data utility is basically a measure of information loss or loss in the functionality of data in providing the results, which could be generated in the absence of PPDM algorithms.
- Uncertainty level: It is a measure of uncertainty with which the sensitive information that has been hidden can still be predicted.
- Resistance: Resistance is a measure of tolerance shown by PPDM algorithm against various data mining algorithms and models.

As such, all the criteria that have been discussed above need to be quantified for better evaluation of privacy preserving algorithms. But, two very important criteria are quantification of privacy and information loss. Quantification of privacy or privacy metric is a measure that indicates how closely the original value of an attribute can be estimated [11]. If it can be estimated with higher confidence, the privacy is low and vice versa. Lack of precision in estimating the original dataset is known as information loss which can lead to the failure of the purpose of data mining. So, a balance needs to be achieved between privacy and information loss. Dakshi Agrawal and Charu Agrawal in [11] have discussed quantification of both privacy and information loss in detail.

REFERENCES

[1] Majid Bashir Malik, M. Asger Ghazi, Rashid Ali, "Privacy Preserving Data Mining Techniques: Current Scenario and Future Prospects", 2012 Third International Conference on Computer and Communication Technology, © 2012 IEEE,

[2] Shuvro Mazumder, " Privacy Preserving Data Mining"

[3] Aris Gkoulalas-Divanis and Vassilios S. Verikios, "An Overview of Privacy Preserving Data Mining", Published by The ACM Student Magazine, 2010.

[4] The Economist. "The End of Privacy", May 1st, 1999. pp: 15

[5] C. Clifton and D. Marks, "Security and Privacy Implications of Data Mining", ACM SIGMOD

Workshop on Research Issues on Data Mining and Knowledge Discovery, 1996. pp: 15-19.

[6] K. Thearling, “Data Mining and Privacy: A Conflict in Making”, DS, November 1998.

[7] R. Agrawal and R. Srikant. “Privacy Preserving Data Mining”, ACM SIGMOD Conference on Management of Data, pp: 439-450, 2000

[8] Y. Lindell and B. Pinkas, “Privacy Preserving Data Mining”, Journal of Cryptology, 15(3), pp.36-54, 2000.

[9] 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.

[10] S. R. M. Oliveira and Osmar R. Zaiane, “Toward Standardization in Privacy-Preserving Data Mining”, DMSSP 2004 (In conjunction with SIGKDD 2004).

[11] D. Agrawal and C. Aggarwal, “On the Design and Quantification of Privacy Preserving Data Mining Algorithms”, PODS 2001. pp: 247-255.

[12] A. Evfimievski, R. Srikant, R. Agrawal and J. Gehrke, “Privacy Preserving Mining of Association Rules”, SIGKDD 2002. pp. 217- 228

[13] S. Rizvi and J. Haritsa, “Maintaining Data Privacy in Association Rule Mining”, VLDB 2002. pp: 682-693.

[14] W. Du and Z. Zhan, “Using Randomized Response Techniques for Privacy Preserving Data Mining”, SIGKDD 2003. pp. 505-510.

[15] S. Agrawal and J. Haritsa, “On Addressing Efficiency Concerns in Privacy-Preserving Mining”, DASFAA 2004. pp. 113-124.

[16] 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.

[17] 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.

[18] 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.

[19] 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

[20] Philip Chan, “On the accuracy of meta-learning for scalable data mining”. Journal of intelligent Information Systems, 8:5-28, 1997.

[21] 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.

[22] 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.

[23] 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.

[24] Aggarwal C, Philip S Yu, "A General Survey of Privacy-Preserving Data Mining Models and Algorithms", Springer Magazine, XXII, 11-52, 2008.

[25] 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.