IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 18, Issue 4, Ver. V (Jul.-Aug. 2016), PP 47-51
www.iosrjournals.org

A Brief Survey on Privacy Preserving Techniques in Data Mining

Vinoth kumar J[1], Santhi V[2]
[1]*Research Scholar, SCOPE, VIT University, Vellore Tamil Nadu India*
[2]*Associate Professor, SCOPE, VIT University, Vellore Tamil Nadu India*

Abstract: *Data mining is a process of extracting the required information from large datasets. Privacy preserving data mining deals with hiding a person's sensitive identity without losing the usability of data. Sensitive identities include some private information about persons, companies, and governments that have to be suppressed before it is shared or published. Thus, privacy preserving data mining has become a vital field of research. The capability of privacy preserving data mining techniques is measured by using metrics such as performance in terms of time efficiency, data utility and level of uncertainty or resistance to data mining algorithms. In this paper, various privacy preserving techniques such as Data anonymization, Data Randomization, use of cryptography are presented.*
General Terms: *Data Mining, Privacy and Security.*
Keywords: *Anonymization, Cryptography, Perturbation, Privacy Preserving, Privacy Preserving Data Mining*.

I. Introduction

Data mining research focuses the extraction of potentially useful information from huge collections of data with a range of application areas from market basket analysis to customer relationship management. The mined information could of rules, patterns, clusters or classification models [1]. Knowledge Discovery is the extraction of patterns and relationships that are not known in advance. Data Mining is the process of discovering novel and interesting patterns as well as predictive understandable and descriptive models from large-scale data [2]. It is defined as the intelligent search technique for new knowledge in existing vast volume of data. Protecting sensitive data is a vital issue while accessing and sharing data. It is also a serious concern over individual confidentiality in data collection, processing, and mining and thus it is considered as an important issue in data mining.

As part of data mining, it is also important to locate valuable and refined information from huge databases which are vulnerable to misuse. Thus, the knowledge present in the data is extracted for use through securing the individual's privacy and protecting data holder against the leakage of the data. The rest of the paper is structured as follows. Section 2 introduces the related work. Section 3 describes the background of PPDM. Section 4 discusses various existing privacy-preserving approaches for data mining. Section 5 deals with evaluation criteria for privacy preserving techniques. Section 6 compares various privacy preserving approaches with the table. Section 7 puts forth the conclusion.

.

II. Background

Privacy is a vital concern while allowing access to different classes of the data set such as business and medical dataset for mining [3]. Privacy is so essential with respect to medical data since it contains private information such as the type of disease associated with patient ID, name, and address. In specific, while mining medical data the original data should be accessible for making precise predictions otherwise it leads to useless results. Any kind of release related to the person- specific information leads to several problems including moral issues. Thus, privacy can be defined as preventing unwanted expose of information while performing data mining on collective results.

2.1 Security Vs Privacy

Security is defined as the capability to manage access to the information, defend from unauthorized disclosure, modification and destruction of information [4]. A medical dataset consists of all information related to the patient. Privacy is a more specific term which is defined as the right of an individual to keep his personal information from being revealed. In medical datasets, a person's specific disease must not be revealed into public domain. Today several known PPDM techniques exist and these are comprehensively studied.

2.1 Privacy Issues and Policies

Privacy is the ability of an individual or group to protect information about them. The most significant issues are to provide confidentiality while preserving information and computational overhead. If cryptographic techniques were used for privacy preservation, it will add more computational complexity. In a distributed environment, when the number of stakeholders becomes larger, the communication cost develops exponentially

[5]. A privacy policy is a set of rules that discloses some of the ways a party gathers, manages, discloses and uses customer's data. In order to ensure privacy, the author must address various privacy attacks which need a high degree of deliberation. In data mining, Privacy attack occurs when one's precise privacy information are openly linked to him. Since it is difficult to identify all types of attacks occasionally, the privacy providers can track certain kind of policies delivered by different nations such as FHIPAA of US, Information Technology Act, 2000 of India and Data Protection Act of UK.

III. Framework

The framework of privacy preserving data mining consists of three elements namely dataset, privacy preserving technique and data mining algorithm. The framework for PPDM is shown in Figure 1. In knowledge discovery from databases or data mining, the transactional data is collected by single or multiple organizations and stored in corresponding databases. Then, it is changed to a format suitable for analytical purposes kept in large data warehouses and then data mining algorithms are applied to it for the generation of knowledge. To protect the privacy of individual, several approaches can be applied to data before or along the process of mining [6]. Privacy can be achieved through any one of the approaches such as data hiding or masking, encryption, secure multi-party computation suppression, generalization, anonymization, perturbation, randomization, condensation etc.

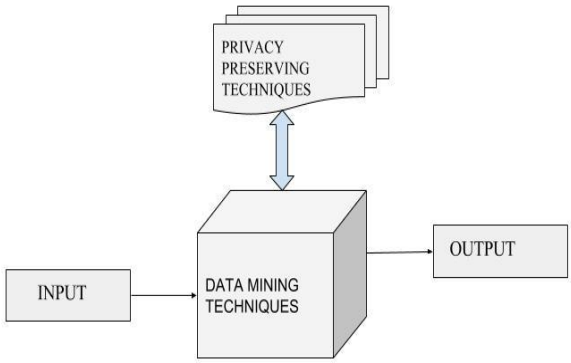

Fig 1. A Simple Privacy Preserving Model

3.1 Classification of PPDM

Basically, PPDM techniques can be classified either central server based or distributed based technique depending upon the environment.

3.1.1 Central Server Scenario

In central Scenario, the PPDM techniques deal with how data is secured before publishing it for the task of data mining. It is also called as Data Publishing scenario. Here data miners and data owners are independent of managing privacy issues. Anonymization and Perturbation are the best suitable approaches under this scenario.

3.1.2 Distributed Server Scenario

In distributed scenario, a PPDM technique must deal with the security against private databases involved in the process of mining. In this, the data owners can also be the data miners and get aggregate results on the union of their databases [7]. Cryptography based on Secure Multi-party Computation (SMC) principle and Perturbation can be used.

IV. PPDM Approaches

Approaches of PPDM are broadly classified into four major groups namely anonymization, perturbation, Randomization, and cryptography. Each group has several variants and sub-categories under them.

4.1 Data Anonymization

Anonymization decreases the risk of identity discovery whereas the data remains still realistic. Each such dataset consists of i) Personal identification like Name, Address or Social Security Number which uniquely identifies an individual ii) Sensitive Attributes like salary and disease iii) The values of Quasi Identifiers such as Gender, Age, Zip code will indicate the discovery of identity when taken together. Two main Privacy Preserving approaches are k-anonymity and l-diversity [8]. Anonymization can be obtained through

methods such as generalization, suppression, data removal, permutation, swapping etc. Some of the variants of k-anonymity are km-anonymization, (α, k) anonymity, p-sensitive k-anonymity, (k, e) anonymity, which is described [9]. Another anonymization technique was known, as Condensation is a statistical approach which creates constrained clusters in the dataset and then generates pseudo-data from the statistics of these clusters.

4.2 Perturbation

Disclosing a 'perturbed' version of a data before releasing it for data mining is one of the data distortion methods for privacy protection. Adding noise from a known distribution is one of the perturbation technique widely accepted [10]. In perturbation, the original values are changed with some false data values so that the numerical information computed from the perturbed data does not change the numerical information computed from the original data. The distorted data records do not agree to real-world record holders, so the attacker cannot achieve the thoughtful linkages or recover private knowledge from the available data. Perturbation can be done by using data swapping. Perturbation methods can be carried out in both the centralized and distributed environment. A variation of classical perturbation technique known as Randomization is a data distortion technique that covers the data by randomly modifying the data values. The randomized response is one of the statistical techniques introduced by Warner to solve survey problems [11]. In Randomized response, the data is jumbled in such a way that the central place cannot say with better probabilities than a pre-defined threshold, whether the data contains rightful information or false information. The information obtained from all individual users is jumbled and if the number of users is significantly greater, the aggregate information of these users can be more accurate.

4.3 Data Randomization

Randomization technique is a cheap and effective approach for privacy preserving data mining [12]. The randomization method has been traditionally used in the context of altering data by probability distribution for methods such as surveys which have an elusive answer bias because of privacy concerns.The technique of randomization can be explained as follows. Consider a set of data records denoted by $X = \{x1, \ldots xn\}$. For record $xi \varepsilon X$, we add a noise component which is taken from the discrete probability distribution FY(y). These noise components are drawn independently and are denoted y1....yn. Thus, a new set of distorted records is denoted by x1+y1…..xn+yn. This new set of records is denoted by z1…….zn. Some methods in randomization are numerical randomization and item set randomization. Noise can be introduced either by multiplying or adding random values to numerical records by removing real items and adding "false" values to the set of attributes. Some of the variants of randomization techniques are discussed subsequently.

4.3.1 Additive and Multiplicative Perturbations

The most common method of the randomization is that of additive perturbations. But, multiplicative perturbations can be used to good effect for privacy preserving data mining. Multiplicative perturbations can be suitably used for distributed privacy preserving data mining.

4.3.2 Data Swapping

An associated method is that of data swapping, in which values across different records are swapped in order to complete the privacy preservation [13]. One merit of these techniques is that the lower order marginal totals of the data are totally preserved and are not distorted at all. It is noted that this method does not follow the general principle of randomization which permits the value of the record to be altered independently of other records.

4.4 Cryptography Based PPDM

Consider a scenario where several medical institutions request to conduct a joint research for some common benefits without revealing unnecessary information. Cryptographic techniques are preferably meant for such scenarios where numerous parties cooperate to calculate results or share non-sensitive mining results and thereby preventing revelation of sensitive information. Cryptographic methods find its use in such scenarios because of two reasons. One is that it offers a definite model for confidentiality that contains methods for demonstrating and measuring it. Second, a vast set of cryptographic procedures and constructs to devise privacy preserving data mining procedures are available in this particular domain.

The data may be distributed among different collaborators either vertically across multiple sites or horizontally. All these methods are almost based on an encryption protocol known as Secure Multiparty Computation (SMC). SMC used in distributed privacy preserving data mining consists of a set of secure sub-protocols which are used in horizontally and vertically partitioned data: Secure scalar product, secure set union and secure sum [14]. Parties that each knows some of the private data participate in a protocol that produces the data mining results that guarantees no data items is exposed to other parties.

V. Evaluation Criteria Of PPDM Strategies

Some of the criteria based on which privacy preserving techniques can be evaluated are accuracy, completeness, consistency, scalability and data quality which are defined in the following sections.

5.1 Accuracy

The accuracy is strictly related to the information loss originating from the hiding strategy: If the loss of the information is less it means the data quality is better. Always a PPDM algorithm has to maintain high accuracy to reduce the loss of information [15].

5.2 Completeness and Consistency

Completeness evaluates the degree of unused data in the cleaned database. Incomplete data has an important effect on data mining results and affects the data mining algorithms from providing a precise representation of the basic data.

5.3 Scalability

Scalability refers to the efficiency trends when the size of data increases. Such factor concerns the increase of both storage and performance requirements as well as the costs of the communications required by a data mining technique with the growth of data size [16].

5.4 Data quality

High-quality data that has been prepared exactly for data mining tasks will result in useful data mining models and output. Alternatively, low-quality data has a substantial negative impact on the utility of data mining results.

VI. Comparison Of PPDM Approaches

The above-discussed ppdm techniques are given in a tabularized format that includes the type of data mining algorithm which with it can employ along with their suitable environment.

Table 1 Comparison of PPDM techniques

Technique	Methods	Scenario	Data mining scheme			
			Classifier	Cluster	ARM	Outlier
Anonymization	Generalization, Suppression, Permutation	Data publishing	✓	✓	✓	✓
Randomization	Adding noise	Data publishing, Distributed	✓	-	-	-
Perturbation	Adding noise, Swapping	Data publishing, Distributed	✓	✓	✓	✓
Cryptography	SMC	Distributed	✓	✓	✓	-

VII. Conclusion

In this paper, we have discussed several approaches which are used in privacy preserving data mining. As the collection of information about individuals and organizations are huge in nature, it is essential to maintain the confidentiality of their sensitive information. Each of the discussed approaches has its own merits and demerits. Most of the privacy attacks can be efficiently prevented by the advanced techniques. In distributed privacy preserving data mining, efficiency remains an important issue. Privacy and accuracy usually contradict each other as improving one undermines the other. All approaches are better to our goal of privacy preservation though there is a need to further improve those approaches to arrive at much more efficient techniques.

References

[1]. Han J, Kamber M. Data Mining: Concepts and Techniques. 2nd edn. Morgan Kaufmann Publishers, 2006.
[2]. Benjamin CM, Fung, Wang K, Chen, Philip Yu S. Privacy-Preserving Data Publishing: A Survey of Recent Developments. ACM Computing Surveys. 2010, June; 42(4), 523-553.
[3]. Majid BM, Asger GM, Rashid Ali. Privacy Preserving Data Mining Techniques: Current Scenario and Future Prospects. *Proceedings of 3rd ICCCT, India*, 2012, 26-32.
[4]. Kamakshi P, Vinaya BA. Preserving Privacy and Sharing the Data in Distributed Environment using Cryptographic on Perturbed data. Journal of Computing. 2010, April; 2(4), 115-119.

[5]. Benny P. Cryptographic Techniques for Privacy-preserving data mining. ACM SIGKDD Explorations. 2008, December; 4(2), 12-19.
[6]. Yogendra KJ, Vinod KY, Geetika SP. An Efficient Association Rule Hiding Algorithm for Privacy Preserving Data Mining. International Journal of Computer Science and Engineering. 2011, July; 3(7), 2792-2798.
[7]. Yaping L, Chen M, Li Q, Zhang W. Enabling Multilevel Trust in Privacy Preserving Data Mining. IEEE Transactions on Knowledge and Data Engineering. 2012, September; 24(9), 1598 – 1612.
[8]. Anil PD, Ravindar M. Privacy Preservation Measure using t-closeness with combined l-diversity and k-anonymity. International Journal of Advanced Research in Computer Science and Electronics Engineering. 2012, October; 1(8), 28-33.
[9]. Kristen LF, Raghu R. Workload-Aware Anonymization Techniques for Large-Scale Datasets. ACM Transactions on database systems. 2008, August; 33(3), 42-51.
[10]. Chen K, Ling L. Geometric data perturbation for privacy preserving outsourced data mining. ACM Journal of Knowledge and Information Systems. 2011, December; 29(3), 657-695.
[11]. Heikki M. Randomization Techniques for Data Mining Methods. *Proceedings of ADIBIS 12th East European Conference, Finland,* 2008, 16-24.
[12]. Jayanti D, Raghavendra K, Debadutta D. Privacy preservation in horizontally partitioned databases using randomized response technique. *Proceedings of IEEE ICICT, South Korea*, 2013, 835 – 840.
[13]. Li Liu, Kantarcioglu M, Thuraisingham B. Privacy Preserving Decision Tree Mining from Perturbed Data. *Proceedings of the 42nd HICSS, USA*, 2009, 1 – 10.
[14]. Blanton M. Achieving Full Security in Privacy-Preserving Data Mining. *Proceedings of 3rd IEEE SocialCom, USA*, 2011, 925 – 934.
[15]. Vedanayaki M. A Study of Data Mining and Social Network Analysis. Indian Journal of Science and Technology. 2014, November; 7(7), 185-187.
[16]. Naser A, Vahid R, Davood K, Sara TM. Information Disclosure by Data Mining Approach. Indian Journal of Science and Technology2012, April; 5(4), 2593-2602.