ISSN 2320-9984
Volume 5 Issue 3| September 2017 ISSN: 2320-9984 (Online)
International Journal of Modern Engineering & Management Research
Website: www.ijmemr.org

Survey on Privacy Preserving Data Mining Techniques

Swati Goel
M. Tech. Research Scholar CSE
Oriental Institute of Science and Technology
Bhopal (M.P.), [INDIA]
Email: goel.swati02@gmail.com

Sanjay Sharma
Assistant Professor
Department of Computer Science Engineering
Oriental Institute of Science and Technology
Bhopal (M.P.), [INDIA]
Email: sanjaysharma@oriental.ac.in

***Abstract*—*Privacy preserving Data mining is an emerging branch of data mining using which data mining operations are performed in huge amount of data to extract meaningful patterns in such a way such that sensitive information is not revealed. Aim of PPDM algorithms is to cover sensitive information within the information in such a way that information miner will extract important data from the changed information with comparable accuracy like original information. The inspiration driving this work is the growing needs of individuals and organizations to share information publically such that protection is guaranteed.*

This survey paper provides a review of different PPDM technique which can be used to stop the information access from unauthorized users and focuses on proposing a novel concept that is combined approach of non homogenous anonymization and association rule mining techniques.

***Keywords:*—*Anonymization, Privacy preservation, data mining, Non Homogeneous generalization, Randomization.*

1. INTRODUCTION

Data mining [1] is a technique of discovering interesting, relevant and meaningful patterns from vast amount of data. As Data mining is an application oriented field of study, it has been applied effectively in numerous areas, such as, Web search, scientific discovery, digital libraries, etc.

Now days with the advancement in information technologies, a tremendous amount of individual or corporate data is being stored for the purpose of analysis and research. Such data is extremely critical from the perspective of data miner as it is utilized for decision making process. But as the data is personal, it can be misused for a variety of reasons. In this way, an ever increasing number of individuals do not want to share their genuine individual information. To deal such a privacy issues in data mining, a sub field of data mining, referred to as Privacy Preserving Data Mining (PPDM) has attracted numerous analysts in recent years.

So more or less, PPDM is a field of keeping up balance between the two issues i.e. information quality and information security. Variety of strategies and techniques are developed for privacy preserving data mining, however many of those strategies may end in data loss and hence reduction in information utility and downgrading the efficiency of knowledge mining. There are many approaches found for privacy preserving in data mining such as data distribution, data modification, rule hiding and many more. Anonymization technique aims at creating the individual record be indistinguishable among a bunch records by utilizing techniques of generalization and suppression. PPDM

strategies can be associated with an immense scope of utilizations, some of which are medical applications, country security applications and bio terrorism applications.

Definitive aim of PPDM algorithms is to hide sensitive information in the database in such a way that data miner can extract meaningful information from the modified database with comparable accuracy as with original database. So, outcome of any PPDM algorithm is a balance between Information loss and utility.

This paper presents a brief analysis regarding the study of different techniques used in the field of PPDM. In section II, a method of categorization of PPDM techniques is discussed. In section III, literature survey of existing PPDM techniques is given. Association rule mining used in privacy preservation is illustrated in section IV. Conclusion and future work is given in section V.

2. CATEGORIZATION OF PRIVACY PRESERVING TECHNIQUES

In recent years, the field of privacy preservation in data mining has been researched extensively. There are several existing approaches that are being used for privacy preservation in data mining. These techniques include anonymization, perturbation, data swapping, generalization and suppression, randomization, secure multiparty computation and much more. On a broad level, PPDM methods can be classified based on data distribution scenario. Based on how the data is distributed, PPDM techniques can be classified as Central Server techniques and distributed server techniques. Both of these techniques can be further classified [2] as shown in Figure 1.

A. Central Server Techniques:

In this scenario data is not distributed among multiple parties. Privacy preservation mechanism is incorporated before the data is published.

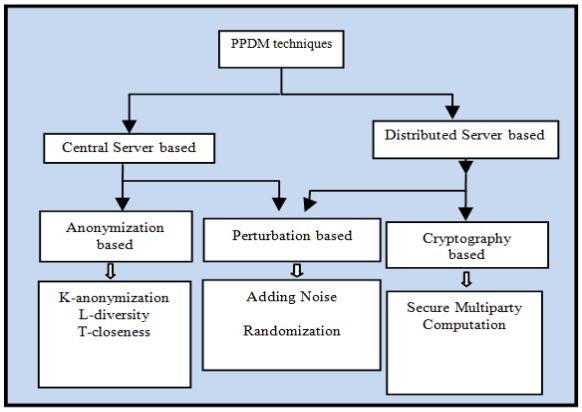

Figure 1: *Classification of PPDM Techniques*

Here data owners do not handle privacy related issues. Data is not handed over in encrypted form to third party rather some kind of precaution is taken care to make the identity of individuals indistinguishable. Anonymization and perturbation based methods comes under this category.

i. Anonymization Method

There are many circumstances when the data need to be shared in its original form for the purpose of analysis and research. But we cannot release the private data in its original form as it will hamper the individual's identity. So, a kind of mechanism is used in which some set of attributes are generalized to make the identity of individual indistinguishable. Anonymization is an example of such a generalization technique. In this way, we are able to share the sensitive information publically without releasing the individual's identity. K-Anonymity model is the basic framework of the data anonymization process.

K-Anonymity:

Most commonly, data set is stored in the form of two dimensional relational databases. For ease of understanding, attributes are divided into four main categories: Key attributes, Quasi attributes, Sensitive attributes and Non sensitive attributes.

Key Attribute: It is also known as identifier. It is used to uniquely identify an individual. As it can directly reveals the identity of the individual, it must be removed during the data

pre-processing phase before performing actual data mining operation. Example: AADHAR card number in India can be used to recognize an individual identity. Hence, it is a key attribute.

Quasi attributes: These are a combination of attributes which can be used to reveal one's identity even after key attribute has been removed from the data set. Quasi-identifiers are a specific sequence of attributes in the table that malicious attackers can take advantage of these attributes linking released dataset with other dataset that has been already acquired, then breaking privacy, eventually gaining sensitive information. Due to uncertainty of the number of quasi-identifiers, each approach of PPDP assumes the quasi-identifiers sequence in advance.

Sensitive attributes: These are set of attributes that contains touchy individual particular information, for example, illness, salary and so forth.

Non-Sensitive attributes: Non-quasi attributes have less effect on data processing. For this reason, sometimes, these attributes does not turn up in the progress of data processing which tremendously decrease memory usage and improve the performance of the proposed algorithm. It is an attribute or a combination of attributes that makes no problem if revealed even to the adversaries.

K-Anonymity model follows the homogenous approach of anonymization. Author [5] has proposed a novel idea of generalization that is based on non homogenous approach. Non-homogeneous anonymization is a technique that provides better data utility as compared to homogenous approach of anonymization. In given example, disease is sensitive attribute. Zip code, gender and age are quasi attributes.

Example given below clearly shows that non homogenous anonymization yields higher data utility.

Table 1. Original Data Table

Zip code	Gender	Age	Disease
701152	M	28	Flu
701157	F	30	Cancer
701578	M	15	Cancer
702398	M	20	AIDS
702301	M	48	None

Table 2. 2-Anonymity Using Homogenous Anonymization

Zip code	Gender	Age	Disease
701***	*	15-30	Flu
701***	*	15-30	Cancer
701***	*	15-30	Cancer
7023**	M	20-48	AIDS
7023**	M	20-48	None

Table 3. 2-Anonymity Using Non Homogenous Anonymization

Zip code	Gender	Age	Disease
70115*	*	15-28	Flu
701***	*	28-30	Cancer
701***	M	15-30	Cancer
7023**	M	20-48	AIDS
7023**	M	20-48	None

ii. Perturbation Based Method

In this method, the original data that needs to be shared publically is modified to ensure the privacy of the data. This technique distorts the data by using any mathematical operation such as addition, subtraction or multiplication. One of the notable techniques used for perturbation is to add a fixed value of noise from a known set of distribution values.

Another variation of the traditional perturbation method is Randomization. In this method, original values are masked by adding data to the original data in arbitrary manner. Initially the randomization method was used for distorting data based on probability distribution; author in paper [3] has extended

this idea to be used in privacy preserving data mining operations. One basic advantageous feature of the randomization method is its simplicity because it is not mandatory to gain knowledge of the distribution of other records in the data. This is in contrast with other methods such as k-anonymity which require the knowledge of other records in the data.

B. Distributed Server Techniques

In this scenario, data is not published on a single server site rather data is distributed among multiple parties. These parties perform privacy protection mechanism on their private databases. Here data owners perform the data mining operations and get the consolidated result over the union of their private databases. Privacy is ensured over the result so that no other information is revealed other than the desired and actual output. Cryptographic and perturbation based methods comes under this category.

i. Cryptographic Method:

If the data is dispersed over numerous sites which are legitimately denied from imparting their information to each other, it is as yet possible to build a data mining model.

Secure multiparty computation (SMC) is a system that can be utilized to keep up protection in various dispersed data mining situations. Now days Secure Multiparty Computation (SMC) is considered as the basic method of providing privacy preservation when multiple parties are involved in process in distributed manner.

Using SMC, every party just knows its own input and expected results. In Distributed scenario, data can be distributed either horizontally or vertically. In horizontally distributed data set, data is partitioned in different datasets that has the same set of attributes and these partitioned data sets are owned by individual parties.

In case of vertically distributed data sets, data set is partitioned such that each partition contains different set of attributes and the same data set. Different parties own those vertically partitioned data set. In order to perform a data mining operation, each party needs to have the consolidated data set which is the union of these individual data sets. At the same time; they do not want to disclose their own data sets. So, it makes the process of PPDM challenging. In order to handle this problem, researches have proposed different ideas. Vaidya and Clifton[4] have proposed a novel method based on association rule mining which is applied on vertically partitioned data sets. A high level of security may be achieved using SMC but its practical implementation is difficult as it involves high computation and communication cost.

3. STUDY OF EXISTING PPDM TECHNIQUES

Latanya sweeney [6] presented K-anonymity model of protection which is used as the basic framework in the field of privacy preservation. According to this framework, a set of records are said to be as K- anonymous if the individual record is indistinguishable from at least K-1 records. The probability of identity disclosure of individuals decreases as the value of 'K' increases. But this model is susceptible to attribute linkage attacks.

Machanavajjhala [7] presented another model known as l-diversity which was build to solve the attribute linkage attack problem of the existing k-anonymity model. Rather than being just a concept, it is practically possible to evaluate the l-diversity model experimentally. This model emphasizes on keeping the minimum number of unique values in the equivalence class of the sensitive attribute. This model is particularly suitable in those circumstances when the values of the sensitive attributes are similar. This model is able to solve the attribute linkage attack but it suffers from homogeneity and background knowledge attacks.

Xuenyun Li [8] studied different PPDM techniques and illustrated his reviews based on comparing advantages and disadvantages of

different techniques. He also pointed out various open issues related to the field. He has used condensation technique and data anonymization technique to safeguard the information.

Slava Kisilevich, Yuval Elovici, Bracha Shapira, and Lior Rokach [9] proposed a novel technique of using K-anonymity within the privacy preservation mechanism. They proposed swapping of the values instead of suppressing it. They have shown that the new method is more efficient than the suppression based method as it incurs less information loss.

G. Loukides, A. Gkoulalas-Divanis Liu [10] suggested a novel technique for data anonymization that is based on the utility requirement of the data publisher. So, in this technique data was anonymized in customized way based on the data utility requirement. They have proposed a unique measurement to calculate the information loss. Using this measurement they have shown that their proposed technique incurs less information loss.

W.K. Kong, Nikos Mamoulis and David W. Cheung [5] proposed a methodology for non homogeneous generalization, which improves utility while maintaining an adequate level of privacy. In their paper, authors have shown that non homogenous generalization technique results in less information loss in comparison to traditional generalization technique when applied on the same data set. Their study is mainly focused on the basic k-anonymity model. They have provided a methodology for verifying whether a non homogeneous generalization violates k-anonymity.

Ninghui Li et al.[11] proposed a technique which was able to solve the problems associated with the existing models k -anonymity and l-diversity. As with paper [7], it is clear that l – diversity was able to solve the linkage attack problem of k-anonymity model, but its practical implementation was challenging. So, there was a need of a novel approach which could solve the problem of both the models. Ninghui [8] proposed a novel privacy technique which is based on the distribution of sensitive values in a different table in order to achieve l-diversity and is able to prevent the attribute disclosure problem.

In paper [12], author attempts to avoid the disclosure of information of the data set on which K-anonymity has already been applied. In data mining, many algorithms have been proposed recently which provides privacy preservation. Author has proposed a novel association rule hiding technique that reduces the information loss by hiding the transactions that has the Support value of specific rule above the pre defined threshold.

4. PRIVACY PRESERVING IN ASSOCIATION RULE MINING

Association rule mining is a technique in data mining that recognizes the regularities or patterns found in huge amount of data. This technique may identify the sensitive patterns underlying the data set that may belong to an individual or corporation. In contrast to association rule mining, privacy preserving rule mining is a technique that is needed to avoid disclosure of personal or sensitive information from the actual or generalized data set.

Association rule hiding [13] is an area that has been researched extensively in recent years along with two major directions. The first kind of approaches is that which target at hiding specific association rules among the mined rules from the database. The second kind of approach is those that hide specific frequent item sets among the frequent item sets found by mining algorithm. By applying these methodologies data owner can be sure that his or her sensitive data is protected adequately. The common approaches used in association rule hiding algorithms are

1) Heuristic approaches

2) Border-based approaches.

The Heuristic approaches are utilized to adjust the chosen transactions from the

database for concealing the sensitive data. The Border-based approaches is used to hide the sensitive rules through the alteration of the origin a borders in the lattice of the frequent and the infrequent patterns in the data set.

5. CONCLUSION AND FUTURE SCOPE OF WORK

In our current paper, we have studied two broad categories of PPDM techniques which are central server based and distributed server based techniques. Out of these categories, we have focused on anonymization technique of privacy preservation. Further with the help of an example, we have shown that non homogenous anonymization provides better data utility while maintaining an adequate level data confidentiality. Even after the data has been generalized, further it is possible to find out sensitive rules from the anonymized data set using the concept of association rule mining. So, anonymized data further can be randomized arbitrarily in order to avoid the disclosure of sensitive rules. In our proposed work, we are going to extend the concept of non homogenous generalization and association rule mining. The proposed algorithm will be executed on ADULT data set taken from UCI machine learning repository. It consists of a limited number of records in thousands. But, in real time, database will contain huge amount of data. Based on our research, the proposed idea will reduce the Risk factor of disclosure while maintaining desirable amount of data originality.

REFERENCES:

[1] J. Han and M. Kamber, "Data Mining: Concepts and Techniques", 2nd ed.,The Morgan Kaufmann Series in Data Management Systems, Jim Gray, Series Editor 2006.

[2] Jisha Jose PaNackal, "Privacy Preserving Data Mining: An Exrensive Survey", ACEEE 2013.

[3] Oliveira S. R. M., Zaiane O.: Data Perturbation by Rotation for Privacy-Preserving Clustering, Technical Report TR04-17, Department of Computing Science, University of Alberta, Edmonton, AB, Canada, August 2004.

[4] C. Clifton, M. Kantarcioglu, J. Vaidya, X. Lin and M. Y. Zhu, "Tools for Privacy-Preserving Distributed Data Min- ing," ACM SIGKDD Explorations Newsletter, Vol. 4, No. 2, 2002, pp. 28-34.

[5] W. K. Kong, Nikos Mamoulis and David W. Cheung, "Non-homogeneous Generalization in Privacy Preserving Data Publishing",June 2010.

[6] L. Sweeney, "K-Anonymity: A Model for Protecting Privacy", International Journal on Uncertainty Fuzziness Knowledge based Systems, 2002.

[7] A. Machanavajjhala, J. Gehrke, D. Kifer and M.Venkitasubramaniam,"L -Diversity: Privacy beyond K-Anonymity", In the Proceedings of the IEEE ICDE 2006.

[8] Xuenun Li,"A review on Privacy Preserving Data Mining",IEEE International Conference On Computer and Infoemation Technology,2014.

[9] Slava Kisilevich, Yuval Elovici, Bracha Shapira, and Lior Rokach, "KACTUS 2: Privacy Preserving in Classification Tasks Using K-Anonymity", Springer-Verlag Berlin Heidelberg 2009.

[10] G. Loukides, A. Gkoulalas-Divanis, "Utility-Preserving Transaction Data Anonymization with Low Information Loss", Expert Systems with Applications, Elsevier 2012.

[11] Ninghui Li Tiancheng Li,Suresh

Venkatasubramanian, T-Closeness: Privacy Beyond K-Anonymity and L-Diversity, ICDE 2007, pp. 106–115.

[12] Dr. R. Sugumar, Dr. A. Rengarajan, M.Vijayanand. "Extending K-Anonymity to Privacy Preserving Data Mining Using Association Rule Hiding Algorithm". International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 6, June 2012.

[13] Amruta Mhatre, Durga Toshniwal, "Hiding Co-occurring Sensitive Patterns in Progressive Databases", ACM, March 22, 2010.

[14] Agrawal R., Srikant R. "Privacy-Preserving Data Mining". Proceedings of the ACM SIGMOD Conference, 2000.

[15] Aggarwal C. C. "On Randomization, Public Information and the Curse of Dimensionality". ICDE Conference, 2007.

[16] Oliveira S. R. M., Zaiane O. "Data Perturbation by Rotation for Privacy-Preserving Clustering, Technical Report TR04-17, Department of Computing Science, University of Alberta, Edmonton, AB, Canada, August 2004.

[17] Jun Lin Lin and Meng Chang, "An efficient clustering method for K-Annonymization", 2010.

[18] Dhanalakshmi. M, "Privacy Preserving Data Mining Techniques-Survey", ICICES 2014, ISBN No.978-1-4799-3834-6.

[19] Jian Wang, "A Survey on privacy Preserving data Mining", International workshop on database Technology and Applications, 2009.

[20] M. Naga Lakshmi, Dr. K. Sandhya Rani "Privacy Preserving Clustering by Hybrid Data Transformation Approach" International Journal of Emerging Technology and Advanced Engineering, ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 3, Issue 8, August 2013.