IJCSE International Journal of Computer Sciences and Engineering Open Access
Survey Paper Volume-6, Issue-1 E-ISSN: 2347-2693

Privacy Preserving In Data Mining: A Survey

Smita Rani[1], Akash Saxena[2]

[1*]Dept. of Computer Science, Compucom Inst. of Tech & Mgmt. Jaipur, Rajasthan Technical University, Jaipur, India
[2]Dept. of Computer Science, Compucom Inst. of Tech & Mgmt. Jaipur, Rajasthan Technical University, Jaipur, India

Corresponding Author: ersmita12@gmail.com, 91-9549135900

Available online at: www.ijcseonline.org

Received: 22/Dec/2017, Revised: 31/Dec/2017, Accepted: 19/Jan/2018, Published: 31/Jan/2018

Abstract— Todays scenario conversion of data from the databases or data warehouse to avail the users is one of the challenging tasks in data mining. There is high risk of data loss and these losses of data sometimes create high risk for users for their sensitive data; because large amount of data gets publish on daily basis. Data mining comes now a day has lots of necessary techniques for privacy preserving. In the past decennary the evolution of various data mining techniques, privacy preservation in data mining becomes an important issues. Basically privacy preservation of data mining provides the facility of sharing of critical data for analysis purposes. The problem of privacy preserving data mining becomes very crucial due to the possibility of occurrence of personal data. Essential parameter used for preserving the privacy of data mining is efficiency, time, cost, accuracy. To achieve the high privacy user have to compromise accuracy, time and cost. This survey paper mainly discussed the introduction of Data Mining, some of the proposed algorithm for privacy preserving in data mining and framework of privacy preservation. Several privacy preservation techniques in data mining based upon different parameters to measure Information Loss Rate (ILR) and Privacy Ratio (PR) are also discussed in this paper.

***Keywords*—**Data Mining, Privacy preserving Data Mining, Clustering.

I. Introduction

Now a day's the consistent growth in a technology resulting the generation of large amount of data from various sources. There are number of data mining techniques are implemented for analysing this data due to increasing of tremendous amount of data. Data mining is technique used to extract useful information from large amount of data. The data which is extracted by data mining techniques may give the private details of individual, due to which individual might face the problem of releasing data without holding privacy. To overcome this problem the idea of releasing and mining of data for analysis while preserving privacy is developed. For preserving the confidential data of user is extremely important, that can be possible by using privacy preserving data mining techniques. Numbers of different privacy preserving data mining techniques are implemented.

The paper is organized as follows, Section I contains the introduction of Privacy preserving Data Mining, Section II contain introduction of data mining, Section III contain the literature survey of proposed algorithm for privacy preserving data mining techniques, Section IV contain the framework of privacy preserving data mining, section V contains the privacy preserving techniques and Section VI provides the conclusion.

II. DATA MINING

In the past decades generation of tremendous amount of data has been observed. Layman defines the data mining in terms of knowledge and information discovery. Basically the data mining is the process of extracting valuable information from the huge amount of data and performs the process of analysis. Various tools are provided for the implementation of data mining techniques. These tools provide multiple ways for analyzing data as well as create relationship among data by categorizing the data. It is process of examining relational databases in or to extract business critical information. The main objective is to implement a model to one problem for which solution is already predicted and then apply that model to another problem which required the solution.

The process of knowledge extraction from large data sets is basically the term addressed as Data mining [1]. Data is passed through many phases during the life cycle of data management. As the data consists of multiple sensitive information that are very important for users so the privacy of data is one of the essential task at each stage of life cycle.

Now a days privacy, security and data integrity are considered as challenging problem in data mining. Extraction of important information from the large dataset can be done using the data mining techniques. Different techniques and algorithms are implemented for the data mining. One of the most necessary conditions for secure communication is privacy preservation using data mining techniques. Data mining techniques also have raised a number of ethical issues. Some such issues include those of privacy, data security, and many others. The functional constituent for obtaining the information and knowledge is privacy preservation using data mining techniques. Data mining incorporate privacy as a functional component for gaining the information and knowledge. Clustering is widely used data mining techniques such as customer behavior analysis, targeted marketing and many others. The challenging task is to achieve the privacy while sharing the data for clustering. To address this problem, the system must not only meet privacy requirement of data owners but also guarantee valid clustering results [2].

Data mining may not hold the confidentiality of data while extracting the useful information among the large datasets, which is one of the challenging task. Privacy preserving is mainly involved with using certain algorithms on private information which cannot be disclosed. For example, using phone directory phone number can be accessed which is linked with unique ID database, which helps to find out confidential information like bank account, address etc. So sensitive data like names, IDs should be trimmed or modified from the original database, which helps to maintain the privacy even after the data mining is performed on the datasets.

Basically privacy preservation can be classified into:

i. Individual privacy preserving: in this category when the sensitive rows are waived off or the database get modified then the data can be protected if accessed can be directly linked to an individual. When such data is mined the protected data should not be disclosed.

ii. Collective privacy preserving: when the trends and patterns have referenced to specific organization is protected including the protecting individual specific information is known as collective privacy preserving. This is same as a statistical database. The collective privacy preserving mainly focuses on the protection of strategic patterns of individual organization which are the most important program of the strategic plan of any organization.

III. LITERATURE SURVEY

Shikha Sharma et al [3] proposed a work which is based on the reduction of support and confidence of sensitive rules. In this work, algorithm is used in some modified form to hide the sensitive association rule without any side effect. To hide the sensitive element, algorithm repeatedly increases the hiding counter of the rule until confidence goes below a minimum specified threshold rather than checking all transactions again and again and ordering them in increasing or decreasing order.

Arpit Agrawal et al [4] proposed privacy-preserving data mining technique that is used to find the right balance between maximizing analysis results In this paper author proposed a new heuristic algorithm that improves the privacy of sensitive knowledge as item sets by blocking more inference channels. The item count and increasing cardinality techniques based on item-restriction that hides sensitive item sets. They demonstrate the efficiency of the algorithm, and also propose an efficient protocol that allows parties to share data in a private way with no restrictions and without loss of accuracy and demonstrate the efficiency of the protocol.

Tagaram Soni Madhulatha et al [5] proposed the k-medoid algorithm. The algorithm begins with arbitrary selection of the K objects as medoid points out of n data points (n>K). After selection of the K medoid points, associate each data object in the given data set to most similar medoid. After selection, non-medoid object O is randomly selected. The total cost of swapping for non medoid object O is calculated. If S>0, then swap initial medoid with the new one. Repeat steps until there is no change in the medoid.

Zhang et al [6] proposed an improved PAM clustering algorithm which involved the calculation of the initial medoids based on a distance measure. Points that were closer to each other were put in separate sets and the points closest to the arithmetic means of the sets were chosen as initial medoids. Authors made use of a minimal spanning tree for the pre-treatments of the data. A minimal spanning tree of the data was constructed and then split to obtain k sub trees which resulted in the formation of k clusters. The results produced by the proposed technique are found better than existing technique in terms of outliers detected and time complexity.

Susana et al [7] discussed a review of the most common partition algorithms in cluster analysis. A comparative study is discussed in this work. The main objective of this study is to compare several partition methods in the context of cluster analysis, which are also called non-hierarchical methods. In this work authors performed a simulation study to compare the results obtained from the implementation of the algorithms k-Means, k-Medians, PAM and CLARA when continuous multivariate information is available. Additionally, a study of simulation is presented to compare partition algorithms qualitative information, comparing the efficiency of the PAM and k-modes algorithms. The efficiency of the algorithms is compared using the Adjusted Rand Index and the correct classification rate. Finally, the algorithms are applied to real databases with predefined classes.

Bhat et al [8] proposed K-medoids clustering using partitioning around medoids for performing face recognition. It explores a novel technique for face recognition by performing classification of the face images using unsupervised learning approach through K-Medoids clustering. Partitioning Around Medoids algorithm (PAM) has been used for performing k-medoids clustering of the data. The results are suggestive of increased robustness to noise and outliers in comparison to other clustering methods. Therefore the technique can also be used to increase the overall robustness of a face recognition system and thereby increase its invariance and make it a reliably usable biometric modality. Comparing these two algorithms, K Medoids algorithm provides better result.

Md Zahidul Islam et al [9] presented a framework for adding noise to all attributes in two steps; in the first step they add noise to sensitive class attribute values, which are also known as labels. Additionally, in the next step they add noise to all non-class attributes to prevent re-identification of a record with high certainty and disclosure of a sensitive class value. Noise addition to non-class attributes also protect the attributes from being disclosed. The main goal of noise addition technique is to provide high level of security while preserving a good data quality by using a novel clustering technique. As mentioned in above the similarity between this method and our proposal method in the using additive noise privacy technique, but authors applied it only on the selected cluster and only on the numerical attributes because the nature of datasets that is used which it numerical.

Ratna Kendhe et al [10] documented a survey on privacy preserving in data mining. This work gives us an overview of the privacy preserving models in data mining, the framework and the techniques of privacy preserving. Different organization use different techniques and models depending on their requirements. The major function of privacy preserving data mining is developing methods to cover or provide privacy to specific sensitive information so that they can't be revealed to unauthorized one.PRIVACY PRESERVING IN DATA MINING .

IV. PRIVACY PRESERVING FRAMEWORK

A particular framework is followed for privacy preserving. The collected data from various sources is first stored in a data warehouse, and then it is converted to a suitable format for analytical purposes. And then data mining techniques are applied to it. Throughout this process privacy preserving has to be implemented at each step.

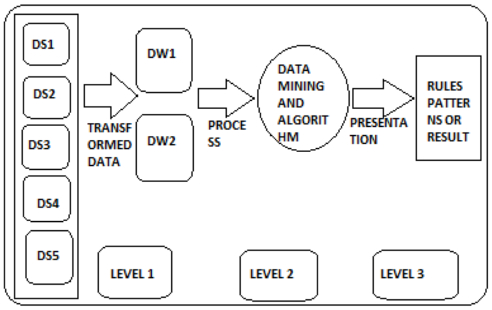

Figure 1. Framework of privacy preserving

The levels of the privacy preserving framework are:

Level 1: At the level 1 of privacy preserving, all the data collected from the various sources were examined whether it is suitable for further processing or not.

Level 2: Sanitization of data is performed at level 2 of privacy preservation. Sanitizing of data deals the operations like blocking, sampling, perturbation, generaliza- tion etc. that are applied at level 2. For the privacy preservation data mining algorithm are updated.

Level 3: This is the final level of privacy preserving framework. At this level the information acknowledged after data mining is checked in order to maintain privacy.

V. PRIVACY PRESERVING TECHNIQUES

The privacy preserving techniques are classified into five categories:

ANONYMIZATION BASED PRIVACY PRESERVING:

The data in the table consists of 4 different attributes:[11]

1. Explicit Identifiers: It is a set of attributes that identify an owner record explicitly.
2. Quasi Identifier: These are a set of attributes if combined with a publicly available tuple would identify the owner.
3. Sensitive Identifiers: An attribute which contains sensitive information about the owner e.g. salary.
4. Non-Sensitive Identifiers: If revealed such attributes create no privacy problems.

The process of hiding the sensitive information of the individual is known as anonymization. If the Quasi identifiers linked with the publically available database will result the disclose of sensitive information. For e.g. if a unique ID database is combined with the bank account database of a, sensitive information like the account number, address of a particular customer can be revealed. The aim is not to disclose the quasi identifiers while data undergone for data mining process. This can be achieved by anonymization, which provides the facility of hiding and modifying of certain quasi identifiers. The number of tuples in a database for preserving privacy is dependent upon the number of quasi identifiers, like if database having 1 quasi identifier then k-1 tuples requires for protecting privacy. This is accomplished by generalization and suppression. Due to anonymization correctness of transformed data is assured but it suffers high information loss.

Drawbacks of the k-anonymity model are: it may be very hard for the owner of a database to determine which of the attributes are available or which are not available in external tables. Another disadvantage is the k-anonymity model assumes a certain method of attack.

PERTURBATION BASED PPDM:

Perturbation means disruption or distress according to oxford dictionary. Similarly in data mining perturbation implies the replacement of original values with some artificial values in such a way that the original information of data is preserved. The data records do not correspond to the individual's actual data. During the vicious attack, retrieval of sensitive information cannot be done because of lack ness of crosslinking, which results the privacy preservation. Therefore only statistical and less means of information is included in the perturbed. Perturbation is done by adding noise to the original data, data swapping or synthetic data generation.

RANDOMIZED RESPONSE BASED PRIVACY PRESERVING:
In this technique the data is scrambled in such a way that the central place cannot tell if the data that is coming from the user contains true or false information. Basically it is statistical technique. The information received from each individual user is scrambled and if the number of users is more, the cumulative information of all users can be estimated with a pretty good accuracy. This is mainly beneficial for decision tree classification because it is based on aggregate values of dataset instead of individual data items. Collection of data involves two steps: firstly the data are randomized by the providers and transmitted that data to data receiver. In the next step, the data receiver collects the randomized data and reconstructs to get the original distribution of the data by using a distribution reconstruction algorithm. Randomization method can be implemented at data collection time, therefore it does not require a trusted server to keep all the records to perform the anonymization process. The drawback of randomization method is that even having a different local density it considers all the record same. This leads to a problem where the outer records become more vulnerable to adverse attacks as compared to records in more inner regions in the data.

CRYPTOGRAPHY BASED PRIVACY PRESERVING:
Cryptographic techniques are based on the fundamentals of distributed computing, where multiple parties come together to compute results or share non sensitive mining results and avoiding disclosure of sensitive information. Cryptographic techniques are beneficial because of two reasons: First, it offers a model for privacy that includes methods for proving and quantifying it. Second a vast set of cryptographic algorithms and constructs to implement privacy preserving data mining algorithms are available in this domain. The data may be distributed among different collaborators vertically or horizontally. In vertically partitioned data between different collaborators, the individual entities may have different attributes of same set of records and in case of horizontally partitioned data, individual records are spread out across multiple entities, each of which has the same set of attributes.
Although cryptographic techniques ensure that the transformed data is exact and secure but this approach fails to deliver when more than a few parties are involved. Moreover, the data mining results may breach the privacy of individual records. There exist solutions of this problem in semi-honest models but in case of malicious models not much study has been made.

CONDENSATION APPROACH BASED PRIVACY PRESERVING:
This method constructs constrained clusters in dataset and then generates pseudo data from the statistics of these clusters. It is called as condensation because it uses condensed statistics of the clusters to generate pseudo data. It constructs groups of non-homogeneous size from the data. Subsequently, pseudo data is generated from each group so as to create a synthetic data set with the same aggregate distribution as the original data. This approach can be effectively used for the problem of classification. The use of pseudo-data provides an additional layer of protection, as it becomes difficult to perform adverse attacks on synthetic data. The aggregate behavior of the data is preserved, making it useful for a variety of data mining problems. This approach helps in better privacy preservation as compared to other techniques as it uses pseudo data rather than modified data. It works even without modifying data mining algorithms since the pseudo data has the same format as that of the original data.

VI. CONCLUSION

Due to increasing rate of data generation, it is intensely important to extract useful information from this large amount of dataset. But it is one of the most important task that the privacy of data must be maintained while data mining. In this paper different privacy preserving techniques proposed by authors has been discussed. Different preserving techniques are used by the organization according to their requirement. As all the privacy preserving techniques proposed are only approximate to our goal of privacy preservation, it is required to further improvement of proposed approaches or develop some efficient methods.

REFERENCES

[1] R Natarajan, "A survey on Privacy Preserving Data Mining, "International Journal of Advanced Research in Computer and Communication Engineering Vol.1, Issue.1, pp. 103-112, 2012.

[2] Kalita, M., D. K. Bhattacharyya, and M. Dutta,"Privacy Preserving Clustering-A Hybrid Approach." Advanced Computing and Communications, 2008. ADCOM 2008. 16th International Conference, Chennai, India. IEEE, 2008.

[3] Shikha Sharma & Pooja Jain, "A Novel Data Mining Approach for Information Hiding", International Journal of Computers and Distributed Systems, Vol.1, Issue 3, October 2012.

[4] Arpit Agrawal," Security based Efficient Privacy Preserving Data Mining" International Journal of Application or Innovation in Engineering & Management, Vol. 2, Issue 7, pp- 225,July 2013.

[5] Tagaram Soni Madhulatha, "Compa rison between K-Means and K-Medoids Clustering Algorithms", Communications in Computer and Information Science, vol. 198, pp. 472-481, 2011.

[6] Zhang L, Yang M, Lei D. "An improved PAM clustering algorithm based on initial clustering centers" 2012. Applied Mechanics and Materials, Vol.135-136, pp. 244-249, 2012.

[7] A. Susana, Leiva-Valdebenito, J. Francisco, Torres-Aviles, " A review of the most common partition algorithms in cluster analysis: a comparative study", Colombian Journal of Statistics , Vol. *33*, No. *2*, pp.321–339, 2010.

[8] Bhat, A., "K-Medoids Clustering using Partitioning Around Medoids for Performing Face recognition", Inter -national Journal of Soft Computing, Mathemetics and Control(IJSCMC), Vol. 3, No. 3, pp. 1-12. August 2014.

[9] Md Zahidul Islam, Ljiljana Brankovic "Privacy preserving data mining: A noise addition framework using a novel clustering technique", Elsevier, Vol. 24, Issue. 8, pp. 1214-1223, 2011 .

[10] Ratna Kendhe, Lahar Mishra, Janhavi Bhalerao,"Privacy Preserving In Data Mining: A Survey", International Journal of Scientific and Research Publications, Vol 5, Issue 10, pp.1-4 , October 2015.

[11] V. Rajalakshmi, "Anonymization based on nested clustering for privacy preservation in data mining", Indian Journal of Computer Science and Engineering (IJCSE), Vol. 4 ,No.3 ,Jun-Jul 2013.

Authors Profile

Ms. Smita Rani pursed Bachelor of Technology from Jagannath University in 2012 and pursuing Master of Technology from Rajasthan Technical University. She has published a paper in ACM WIR-2018. Her research work focuses on clustering techniques, Data Mining. She has 2 years of teaching experience.

Dr. Akash Saxena pursued Bachelor in Engineering from UPTU, and Master of Science from udaipur university. He completed his PhD. Form Bikaner university in 2013. He is a member of IJCSI and IEEE. He has more than 15 years of teaching experience.