Privacy Preserving In Data Mining: A Survey

Ratna Kendhe[*], Lahar Mishra[**], Janhavi Bhalerao[**]

[*] Department of Computer Science, NMIMS University, Mumbai, India
[**] Department of Computer Science, NMIMS University, Mumbai, India

***Abstract*-**Privacy preserving has becomes important issue in the past decade due to the immergence of various data mining techniques. Privacy preserving data mining has become extremely essential because it allows sharing of critical data for analysis purposes. The availability of personal data has made the problem of privacy preserving data mining very critical. This paper aims at giving an overview into data mining and the concept of privacy preserving. It gives an insight in to privacy preserving models, framework and techniques.

***Index Terms*-** Data mining, Privacy preserving data mining.

I. INTRODUCTION

In today's world there is a continuous advent of advanced technology, a tremendous amount of data is being generated my various organization. As the data is increasing different data mining techniques to analyze this data need to be adopted. Data mining is technique used to extract useful information from large amount of data. The extracted data might provide information about an individual's private details. Hence individuals are faced with the task of releasing data which does not compromise privacy. The idea is to release data, and mine this data to analyze trends while preserving privacy. Privacy preserving in data mining is extremely important to preserve the data holder's confidential information. Various techniques are implemented to preserve the same. This paper is formatted in the following manner, the second section an introduction to data mining, the third section gives us a the classification of the data mining model, the third section tells us about privacy preserving, the fourth gives an insight into the models of privacy preserving in data mining and the final section tells us about the techniques for preserving privacy in data mining.

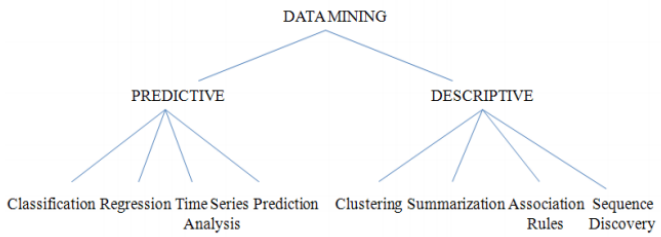
II. DATA MINING

There has been an exponential rise in the generation of data in the past decade. Data mining in layman terms is information and knowledge discovery, it is the process of analyzing large amount of data summarizing it and extracting useful information from it. Various software's are available for implementing data mining techniques, such tools allow users to analyze data by multiple ways, categorize the data and establish relationships. It is process of examining relational databases in or to extract business critical information. The basic objective is to apply a model for one problem to which the answer is known and then apply that model to a different problem in which the answer is required.

Consider a product company. Data mining is used by such a company to determine to relationships between internal factors like price, staff efficiency etc. And external factors like competition, economy trends. It permits them to understand the impact each factor has on the sales of the product, understand the gross profits of the company and get a detailed summary of the transactional data.

III. DATA MINING MODEL

The data mining model can be classified as Predictive or Descriptive in nature.



The predictive model: As the name suggests, this model predicts the values of data based on the previous values of the data. [1]
The descriptive model: This model analyses the relationship between the data. In this model all the properties of the data can be scrutinized.[1] A Data mining software analyses the relationships between the data stored in a data warehouse based on user generated queries. The types of relationships aimed are:

1. Classes: The stored data is located within predefined groups
2. Clusters: The data items are mined according to logical representations.
3. Associations: Mining can be done through association.
4. Sequential Discovery: Data mining is performed to anticipate behavioral trends.

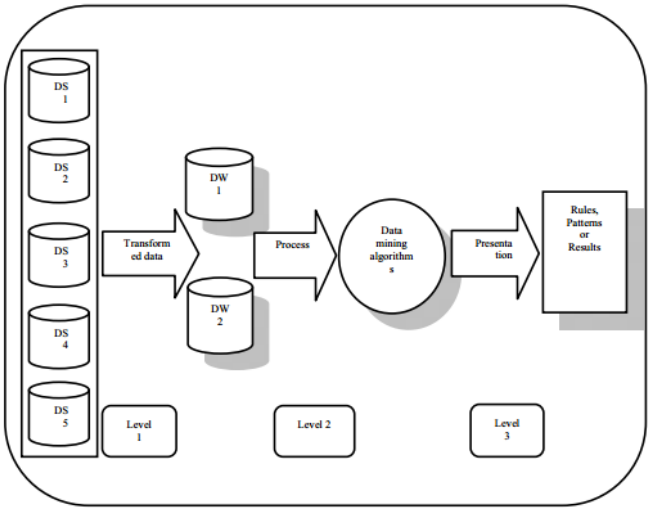
IV. PRIVACY PRESERVING IN DATA MINING

Data mining can be looked at as a threat to privacy. The confidentiality of the data holder comes into question. Privacy preserving is concerned with applying certain algorithms on confidential data that is not supposed to be disclosed. For example, medical information from the database of hospital can be crossed linked with voters ID database to find out confidential information like address of the data holder. So sensitive data like

addresses, names, IDs should be trimmed or modified from the original database, in order to not compromise privacy.
There are two types in privacy preserving:
Individual privacy preserving is the protection of data which if retrieved can be directly linked to an individual when sensitive tuples are trimmed or modified the database. When such data is mined the protected data should not be disclosed.
Collective privacy preserving includes protecting not only individual specific information but also trends and patterns relating to a particular organization. This is similar to a statistical database. The aim of collective privacy preserving is to protect strategic patters which are the most important agenda of the strategic plan of any organization.

V. Privacy preserving framework

Privacy preserving follows a particular framework. The data collected from various sources is first stored in a data warehouse, 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.
The figure below shows the framework of privacy preserving.



The levels of the privacy preserving framework are:
Level 1: At this stage the data from various sources is collected and checked whether it is suitable for further processing. At this stage to it is checked if the privacy is not compromised.
Level 2: At this level the data is sanitized. Processes like blocking, sampling, perturbation, generalization etc. are applied. The data mining algorithms are altered to preserve privacy.
Level 3: At this level the information revealed after data mining is checked in order to maintain privacy.

VI. Privacy preserving Models

There three main models of privacy preserving are:

1. Trust third party model: It is assumed that there exists a third party that to whom all the data is given. It is understood that no one other than that particular third party has any access to the data. The aim is to implement privacy preserving techniques to maintain the confidentiality of this third party.
2. Semi honest model: Very party follows certain protocols using correct formats but it is free to use any protocol during the execution if it feels the security is threatened.
3. Malicious model: Since the semi-honest model does not provide protection for all applications the malicious model is used. The malicious model is free to use any protect it wishes to protect privacy. It is difficult to develop efficient protocols for this model.

VII. 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:

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.

Anonymization is an approach where sensitive information about an individual is to be hidden. Quasi identifiers when combined to a publically available database can reveal sensitive information. For e.g. if a voters ID database is combined with the employee database of a company, sensitive information like the salary of a particular employee can be revealed. So k-anonymization involves hiding or the modification of certain quasi identifiers such that when the data is dent in for data mining the quasi identifiers are not disclosed. Where if there is 1 quasi identifier then the data sent for data mining will have k-1 tuples. Thus protecting privacy. This is accomplished by generalization and suppression. Although the anonymization method ensures that the transformed data is correct but it suffers heavy information loss. This method is not immune to similar attack and background knowledge attack. Limitations of the k-anonymity model are, first, 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. The second limitation is that the k-anonymity model assumes a certain method of attack, while in real scenarios; there is no reason why the attacker should not try other methods.
Perturbation based PPDM

The dictionary meaning of perturbed is being unsettled or upset. Similarly in data mining perturbation means replacing or upsetting the original values with some synthetic data values such that the statistical information of the data is preserved. The data records do not correspond to the individual's actual data. So in a malicious attack, cross linking cannot be done to retrieve sensitive information. Thus preserving privacy. Therefore the perturbed data are meaning less and contain only statistical information. Perturbation is done by adding noise to the original data, data swapping or synthetic data generation.

Randomized Response based Privacy Preserving

Randomized response is a statistical technique. In this 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.

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 very useful for decision-tree classification since decision-tree classification is based on aggregate values of a dataset, rather than individual data items. The data collection process is a twostep process. During first step, the data providers randomize their data and transmit the randomized data to the data receiver. In second step, the data receiver reconstructs the original distribution of the data by employing 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 limitation of randomization technique is that it treats all the records same even if they are of different local density. 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 base 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 a solutions of this problem in semi-honest models but in case of malicious models not much studies have been made.

Condensation Approach base Privacy Preserving

This methods constructs a 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 behaviour 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.

VIII. CONCLUSION

In this paper we have documented a survey on privacy preserving in data mining. In today world every second tremendous amounts of data is generated, it is extremely essential to remove valuable information from this data. But while data mining we have to make sure that the confidentiality of the data is maintained. This paper 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.

ACKNOWLEDGMENT

We would sincerely like to thank our professors for the constant mentorship. We are grateful towards our peers for the encouragement and constructive criticism.

REFERENCES

[1] Analysis of Privacy Preserving K-Anonymity Methods and Techniques S.Vijayarani#1, A.Tamilarasi#2, M.Sampoorna#3 #1, #3School of Computer Science and Engg., Bharathiar University, Coimbatore, Tamilnadu,India #2Dept. of MCA, Kongu Engg. College, Erode, Tamilnadu, India #1vijimohan_2000@yahoo.com #3m.sampoorna87@gmail.com.

[2] W.-K. Chen, *Linear Networks and Systems* (Book style). Belmont, CA: Wadsworth, 1993, pp. 123–135. Privacy Preserving Data Mining Techniques: Current Scenario and Future Prospects Majid Bashir Malik Department of Computer Sciences, BGSB University, Rajouri, J & K. majid.malik@rediffmail.com M. Asger Ghazi School of Mathematical Sciences and Engineering, BGSB University,

Rajouri, J & K. m_asger@yahoo.com Rashid Ali College of Computers and Information Technology, Taif University, Taif, Saudi Arabia. rashidaliamu@rediffmail.com

[3] Security in Privacy Preserving Data Mining Neha Kashyap1 , Dr. Vandana Bhattacharjee2 1Birla Institute of Technology, Department of Computer Science & Engineering, Extension Centre, Lalpur, Ranchi-834001, Jharkhand, India Kashyap9.neha@gmail.com 2Birla Institute of Technology, Department of Computer Science & Engineering, Extension Centre, Lalpur, Ranchi-834001, Jharkhand, India vbhattacharya@bitmesra.ac.in

AUTHORS

First Author – Ratna Kendhe, B.Tech Computer Science(Currently Pursuing),NMIMS University,Mumbai ,India, Ratna.kendhe@gmail.com.

Second Author- Lahar Mishra B.Tech Computer Science(Currently Pursuing),NMIMS University,Mumbai ,India,laharm@gmail.com.

Third Author- Janhavi Bhalerao, B.Tech Computer Science(Currently Pursuing),NMIMS University,Mumbai ,India,janhavi.bhalerao@gmail.com.