A Survey based on Privacy Preservation Data Mining Techniques

Hemal J. Kanjia[1], Mr. Devang Patel[2]
[1, 2] Department of Computer Engineering
[1, 2] SOCET, AHMEDABAD

Abstract- *Data mining is process of extracting useful knowledge/information from large amount of Data. In digital generation, data mining is an important tool to transform data into useful information. When we analyze the data mining algorithm, maintaining the privacy of data is become an issue. To solve that issue, privacy preservation data mining (PPDM) field is applied. Privacy preservation data mining is used to hide sensitive information or knowledge which is generated from database/data warehouse using data mining algorithm. Different techniques are used for PPDM.*

Keywords- Data mining, Privacy Preserving, Sensitive data.

I. INTRODUCTION

Data Mining refers to extracting or mining knowledge from large amounts of data. Data mining is the process of discovering the intensive knowledge from large amounts of data stored either in databases, data warehouses, or other information repositories. It is an interdisciplinary field and it involves an integration of techniques from multiple disciplines such as database and data warehouse technology, statistics, machine learning, pattern recognition, information retrieval, and spatial or temporal data analysis. Data mining, with its promise to efficiently discover valuable, non-obvious information from large databases, is particularly vulnerable to misuse. So, there might be a conflict among data mining and privacy. The another popular name of data mining is, Knowledge Discovery from Data, or KDD.

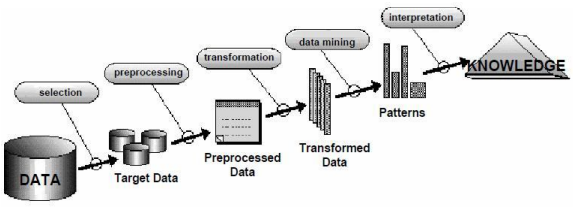

Figure-1 KDD Process

Knowledge discovery process is shown in Figure-1 and consists of an iterative sequence of the following steps:-

1. Data cleaning- In this step the noisy and the meaningless data are discarded from the database
2. Data integration- In this step one or more data sources are combined and generate single data source
3. Data selection- In this step the meaningful and data related to analysis is selected and retrieved from the data source.
4. Data transformation- In this step the relevant data is transformed into the appropriate form which is suitable to mining process.
5. Data mining- This is the most important step and in this step intelligent techniques are applied to generate the patterns.
6. Pattern evaluation- In this step the patterns are evaluated with selected measure.
7. Knowledge Presentation - Finally the knowledge which is identified is represented to user.

II. LITERATURE REVIEW

Privacy preservation data mining is novel research area where Data mining algorithms are analyzed for their side-effects they done on Data privacy. Privacy preservation data mining (PPDM) deals with the problem of hiding the sensitive information while analyzing data. Many techniques are available for PPDM like data distortion, data hiding, rule hiding, data modification etc. Privacy preserving data mining techniques are divided into two broad areas, data hiding and knowledge hiding. Data hiding is removal or modification of confidential information from the data before disclosing to others. Knowledge hiding is focus on hiding the sensitive knowledge which can be mined from the database using any data mining algorithm.

Need for Privacy in Data Mining

Security of private data is always been an important issue for mankind specially in this information area. When we analyze the data mining algorithm, maintaining the privacy of data is become an issue. To solve that issue, privacy preservation data mining (PPDM) field is applied. Privacy preservation data mining is used to hide sensitive information or knowledge which is generated from database/data warehouse using data mining algorithm. Different techniques are used for PPDM.

Privacy Preserving Data Mining

Techniques (basis of dimensions)

On the basis of dimensions, there are mainly five different PPDM techniques can be classified:

1. Data distribution
2. Data modification
3. Data mining algorithm
4. Data or rule hiding
5. Privacy preservation

1. Data distribution:

This dimension refers to the distribution of data. There are some of the approaches are developed for centralized data, while others refer to a distributed data scenario. Distributed data scenarios can be divided as horizontal data partition and vertical data partition. Horizontal distribution refers to these cases where different sets of records exist in different places, while vertical data distribution refers where all the values for different attributes reside in different places.

2. Data modification:

Data modification is used with the aim of change the unique values of a database that wants to be allowed to the public and in this way to guarantee high privacy protection. Methods of data modification include:

Perturbation: which is able to replacing attribute value by a new value (changing a 1-value to a 0-value, or adding noise)
Blocking: which is the replacement of an existing attribute value with a "?"
Swapping: This refers to interchanging values of individual record.
Sampling: This refers to losing data for only sample of a population.
Encryption: Many Cryptographic techniques are used for encryption.

3. Data mining algorithm:

The data mining algorithm for which the privacy preservation technique is designed:

i. Classification data mining algorithm
ii. Association Rule mining algorithm
iii. Clustering algorithm

4. Data or Rule hiding:

This dimension refers to whether raw data or grouped data should be hidden. Data hiding means protecting sensitive data values, e.g. names, social security numbers etc. of some people. And Rule hiding means Protecting Confidential Knowledge in data, e.g. association rule. The difficulty for hiding aggregated data in the form of rules is very difficult, and for this purpose, typically heuristics have been developed.

PPDM Techniques

There are mainly three Privacy Preserving Data Mining Techniques:

1) Heuristic-based techniques
2) Cryptography-based techniques
3) Reconstruction-based techniques

1) Heuristic-based techniques

It is an adaptive modification that modifies only selected values that minimize the effectiveness loss rather than all available values.

2) Cryptography-based techniques

This technique includes secure multiparty computation where a computation is secure if at the completion of the computation, no one can know anything except its own input and the results. Cryptography-based algorithms are considered for protective privacy in a distributed situation by using encryption techniques.

3) Reconstruction-based techniques

Where the original distribution of the data is reassembled from the randomized data.

Based on these dimensions, different PPDM techniques may be classified into following five categories:

I. Anonymization based PPDM
II. Perturbation based PPDM
III. Randomized Response based PPDM
IV. Condensation approach based PPDM
V. Cryptography based PPDM

I. Anonymization based PPDM

Anonymization refers to an approach where identity or/and sensitive data about record owners are to be hidden. It even assumes that sensitive data should be retained for analysis. It's obvious that explicit identifiers should be removed but still there is a danger of privacy intrusion when

quasi identifiers are linked to publicly available data. Such attacks are called as linking attacks.

II. Perturbation based PPDM

In the perturbation approach any distribution based data mining algorithm works under an implicit assumption to treat each dimension independently. Relevant information for data mining algorithms such as classification remains hidden in inter-attribute correlations. This is because the perturbation approach treats different attributes independently. Hence the distribution based data mining algorithms have an intrinsic disadvantage of loss of hidden information available in multidimensional records. Another branch of privacy preserving data mining that manages the disadvantages of perturbation approach is cryptographic techniques.

III. Randomized Response based PPDM

In Randomized response, the data is twisted in such a way that the central place cannot say with chances better than a pre-defined threshold, whether the data from a customer contains correct information or incorrect information. The information received by each single user is twisted and if the number of users is large, the aggregate information of these users can be estimated with good quantity of accuracy. This is very valuable for decision-tree classification. It is based on combined values of a dataset, somewhat individual data items. The data collection process in randomization method is carried out using two steps. During first step, the data providers randomize their data and transfer the randomized data to the data receiver. In second step, the data receiver rebuilds the original distribution of the data by using a distribution reconstruction algorithm.

Randomization method is relatively very simple and does not require knowledge of the distribution of other records in the data. Hence, the randomization method can be implemented at data collection time. It does not require a trusted server to contain the entire original records in order to perform the anonymization process.

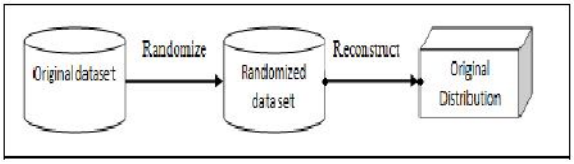

Figure-2 Randomization

IV. Condensation approach based PPDM

Condensation approach constructs constrained clusters in dataset and then generates pseudo data from the statistics of these clusters. It is called as condensation because of its approach of using condensed statistics of the clusters to generate pseudo data. It creates sets of dissimilar size from the data, such that it is sure that each record lies in a set whose size is at least alike to its anonymity level. Advanced, pseudo data are generated from each set so as to create a synthetic data set with the same aggregate distribution as the unique data. This approach can be effectively used for the classification problem. The use of pseudo-data provides an additional layer of protection, as it becomes difficult to perform adversarial attacks on synthetic data. Moreover, the aggregate behavior of the data is preserved, making it useful for a variety of data mining problems. This method helps in better privacy preservation as compared to other techniques as it uses pseudo data rather than modified data. Moreover, it works even without redesigning data mining algorithms since the pseudo data has the same format as that of the original data.

V. Cryptography based PPDM

Consider a scenario where multiple medical institutions wish to conduct a joint research for some mutual benefits without revealing unnecessary information. In this scenario, research regarding symptoms, diagnosis and medication based on various parameters is to be conducted and at the same time privacy of the individuals is to be protected. Such scenarios are referred to as distributed computing scenarios. The parties involved in mining of such tasks can be mutual un-trusted parties, competitors; therefore protecting privacy becomes a major concern. Cryptographic techniques are ideally meant for such scenarios where multiple parties collaborate to compute results or share non sensitive mining results and thereby avoiding disclosure of sensitive information. Cryptographic techniques find its utility in such scenarios because of two reasons: First, it offers a well-defined 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.

III. CONCLUSIONS

Currently privacy preserving in data mining is hot topic of research. The main objective of privacy preserving data mining is developing algorithm to hide or provide privacy to certain sensitive information so that they cannot be disclosed to unauthorized parties or intruder. Although a Privacy and accuracy in case of data mining is a pair of ambiguity. In this, we made an effort to review a good number

of existing PPDM techniques. Finally, we conclude there does not exists a single privacy preserving data mining algorithm that outperforms all other algorithms on all possible criteria like performance, utility, cost, complexity, tolerance against data mining algorithms etc. Different algorithm may perform better than another on one particular criterion.

REFERENCES

[1] Agarwal, R. and Shrikant, R. "Privacy Preserving Data Mining", Proceeding of Special Interest Group on Management of Dat pp. 439 - 450, 2000.

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

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

[4] 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.

[5] 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.

[6] Gayatri Nayak, Swagatika Devi, "A survey on Privacy Preserving Data Mining: Approaches and Techniques", ternational Journal of Engineering Science and Technology, Vol. 3 No. 3, 2127-2133, 2011.

[7] Benny Pinkas, "Cryptographic Techniques for Privacy preserving data mining", SIGKDD Explorations, Vol. 4, Issue 2, 12-19, 2002.

[8] Aggarwal C, Philip S Yu, "A condensation approach to privacy preserving data mining", EDBT, 183-199, 2004.

[9] Wang P, "Survey on Privacy preserving data mining", International Journal of Digital Content Technology and its Applications, Vol. 4, No. 9, 2010.

[10] Charu C. Aggarwal, Philip S. Yu “Privacy-Preserving Data Mining Models and algorithm”advances in database systems 2008 Springer Science, Business Media, LLC.