© 2016 IJSRSET | Volume 2 | Issue 2 | Print ISSN : 2395-1990 | Online ISSN : 2394-4099
Themed Section: Engineering and Technology

A Survey: Privacy Preservation Data Mining Techniques and Geometric Transformation

Anjana K. Patel
Computer Engineering Department, Silver Oak College of Engineering & Technology, Ahmedabad, Gujarat, India

ABSTRACT

What is Privacy Preserving Data Mining is the process of hiding and protecting sensitive data of individuals. In the recent era, we use many applications which require personal sensitive data of individuals. Thus, people are more concern about sharing their personal sensitive information due to increase of privacy intrusions. Since last two decades many Privacy Preserving Data Mining techniques are used today. In this paper, we present a detail comparative study of various Privacy Preserving Data mining techniques and their pros and cons.
Keywords: Data Mining, Privacy Preserving, Privacy Preserving Data Mining Techniques**.**

I. INTRODUCTION

Data mining is one of the core processes in knowledge discovery of databases. Data mining research deals with the extraction of potentially useful information from large collections of data with a different type of application areas such as customer relationship management, market basket analysis. The mined data can be a patterns, rules, clusters or classification models. During the whole process of data mining these data, which typically hold sensitive individual information such as medical and Financial circumspect, often get exposed to different parties including collectors, owners, users and miners. The massive amount of data available means that it is possible to acquire knowledge of a lot of information about individuals from public data. Privacy preserving has start as an important concern with reference to the favourable outcome of the data mining. Privacy preserving data mining (PPDM) deals with keep safe the privacy of individual data or sensitive knowledge without sacrificing the usefulness of the data. People have become well familiar with of the privacy intrusions on their personal data and are very forced to share their sensitive information. In recent years, the area of privacy has become aware of fast advances because of the increases in the power to store data.

In particular, fresh advances in the data mining field have lead about privacy .The goal of privacy preserving data mining(PPDM) algorithms is to mined relevant information from vast amounts of data while protecting at the same time reflective information. The main goals a PPDM algorithm is:

1. A PPDM algorithm should have to Foil the discovery of sensible information.
2. It should be proof against to the various data mining techniques.
3. It should not settlement the access and the application of no sensitive data.
4. It should not have an exponential computational complexity.

Many secure protocols have been proposed so distant for data mining and machine learning techniques for decision tree classification, clustering, association rule mining, Neural Networks, Bayesian Networks. The most important concern of these algorithms is to preserve the privacy of parties' sensitive data, while they obtain useful knowledge from the complete dataset. One of the most studied difficulties in data mining is the process of discovering frequent item sets and, consequently association rules. Association rule mining are usually used in various field. Most of the privacy-preserving data mining techniques put in an application for a

transformation which reduces the used of the underlying data when it is applied to data mining techniques or algorithms. Privacy concerns can keep away from building of centralized warehouse – in scattered among several places, no one are allowed to transfer their data to diffrent place. In preserving privacy of data, the problem is how securely results are gained but not with data mining result but. As a example, suppose some hospitals want to get useful aggregated knowledge about a specific diagnosis from their patients' records while each hospital is not allowed, due to the privacy acts, to make known individuals' private data. Therefore, they call to run a common and secure protocol on their distributed database to reach to the required information. In many cases data is assign and fetching the data collected in one position for analysis is not possible due these privacy acts or regulation. Mining association rules requires iterative scanning of database, which is quite expensive, in processing. These techniques can be demonstrated in centralize as well as distributed environment where data can be varying among the different sites. Distributed database scenario can be classified in horizontally partitioned data and vertically partitioned data.

1. Horizontally partitioned data: It divides database into a number of seprate horizontal partitions. In this type of data different places have different record about same entities which are used for mining purposes. Many of these methods use specialized versions of the common approaches discussed for various problems.

2. Vertically partitioned data: In Vertically partitioned data sets; each site has different number of attributes with same number of transaction. The technique, of vertically partitioned mining has been expand to a variation of data mining applications such as decision trees, SVM Classification, Naïve Bayes Classifier, and k-means clustering.

II. METHODS AND MATERIAL

A. PPDM Framework

Data mining is the process of extracting or mining knowledge from large amounts of data [1]. The extracted knowledge can be used for decision making, process control, information management, query processing and so on.

Now-a-days, data mining is used widely in many applications and huge volume of data is collected. As data mining extracts information from large databases, which may make the data vulnerable and lead to misuse. Some examples of sensitive data are:- credit card/debit card details, criminal records, medical history, identity information etc. Thus, it's necessary to have some privacy policy to secure the sensitive personal data of individuals.

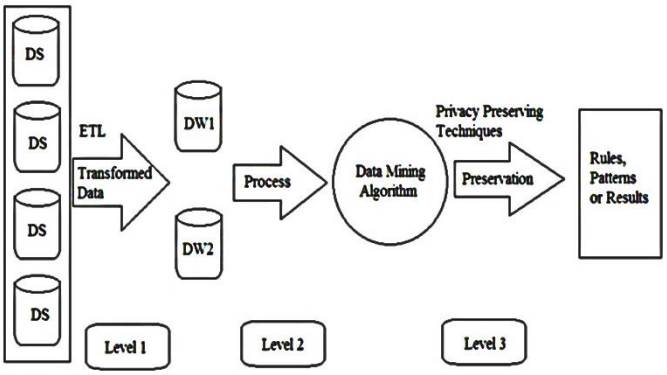

Figure 1. Framework for Privacy Preserving Data Mining

In recent era, Privacy Preserving Data Mining has emerged as an very important research area. Privacy Preserving Data Mining deals with giving protection to the individual's sensitive data.

This paper presents a detailed and comparative survey on recent algorithms developed for achieving Privacy Preserving Data mining. The following Privacy Preserving Data Mining techniques studied and analyzed in this paper:- Data Perturbation approach, Anonymization approach, Cryptographic approach, Blocking-Based approach, Condensation approach and Hybrid approach.

M. B. Malik, M. A. Ghazi and R. Ali introduced a ramework for Privacy Preserving Data Mining shown in Figure 1.

Data from different data sources are aggregated and pre-processed by using ETL tools. The transformed data from Level 1 are stored in data warehouses. In

Level 2, data mining algorithms are applied on the data in warehouse to find the patterns and discover knowledge. In Level 3, mining privacy preservation

techniques are used to protect data from unauthorized access.

B. Literature Study

PPDM tends to transform the original data so that the result of data mining task should not defy privacy constraints. Following is the record of five dimensions on the basis of which different PPDM Techniques can be classified:

i. Data distribution ii.
ii. Data modification
iii. Data mining algorithms
iv. Data or rule hiding
v. Privacy preservation Data or Rule Hiding

This dimension mention 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 private Knowledge in data, e.g. association rule. The difficulty for hiding aggregated data in the form of rules is very difficult, and for this reason, typically heuristics have been developed. Data Distribution: This dimension refers to the distribution of data. There are some of the technique 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 mention to these cases where different sets of records exist in different places, while vertical data distribution mention where all the values for different attributes reside in different places. 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:

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

Data Mining Algorithm: The data mining algorithm for which the privacy preservation technique is designed.

1. Classification data mining algorithm
2. Association Rule mining algorithms
3. Clusering algorithm

Privacy Preserving Techniques:

1. Heuristic-Based Techniques: It is an adaptive modification that change 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. Cryptographybased algorithms are considered for protective privacy in a distributed situation by using encryption techniques.

3. Reconstruction-Based Techniques: where the native distribution of the data is reassembled from the randomized data. Based on these dimensions, different PPDM methods may be classified into following five categories.

1 Perturbation based PPDM
2. Anonymization based PPDM
3. Randomized Response based PPDM
4. Condensation approach based PPDM
5. Cryptography based PPDM we discuss these in detail in the following subsections

3.1 Data Perturbation Approach:- In data perturbation technique, the sensitive data values are modified by using mathematical formula such as addition, subtraction etc. In case of discrete data, first preprocessing is done [3].

Preprocessing include of two steps- attribute coding and obtaining sets of coded data sets. Different methods such as data transpose matrix, addition of unknown values, addition of noise are used for distortion [4].

Major issue of these perturbation methods is difficulty in preserving the original data. T. Jahan, G.Narsimha and C.V Guru Rao introduced a new perturbation method based on Singular Value Decomposition (SVD) and Sparsified Singular Value Decomposition (SSVD) [4]. This technique is more efficient than that of other perturbation approaches.

Perturbation techniques can be used for achieving privacy in data publishing and also in the process of data mining these have certain limitations: i) Since this model uses distributions instead of original records, it restricts the range of algorithmic methods that can be used on the data. ii) And another limitation is the loss of implicit information available in multidimensional records. A variation of classical perturbation technique known, as *Randomization* is a data distortion technique that masks the data by randomly modifying the data values.

Disclosing a 'perturbed' version of a data before releasing it for data mining is one of the data distortion method for privacy protection. Adding noise from a known distribution is one of the perturbation technique widely accepted. Before conducting a data mining operation, the miner should reconstruct the perturbed version to obtain the original data distribution.

Perturbation methods can be used in both scenarios, central server as well as in distributed scenario. These methods use some sort of data distortion techniques like adding noise, randomization or condensation.

3.2 Cryptographic Approach :- Cryptographic techniques are preferably meant for such scenarios where multiple parties collaborate to compute results or share non sensitive mining results and thereby avoiding publishing of sensitive information. Cryptographic techniques find its utility in such scenarios because of two cause: First, it offers a welldefined model for privacy that includes methods for proving and quantifying it. Second a huge set of cryptographic algorithms and constructs to implement privacy preserving data mining algorithms are obtainable in this domain. The data may be distributed among different collaborators vertically or horizontally.

In this approach the sensitive data are encrypted by various algorithms [5]. There are two main drawbacks of this method, those are:- The outputs of computation lacks in privacy and this approach is not efficient enough in case of large databases.

3.3 Condensation Approach :- Condensation approach was introduced by C. Aggrawal and P.S. Yu [7]. This method is used in dynamic data update. In this method, data are condensed into multiple groups of predefined size. Then they generate corresponding pseudo- data by using statistics.Condensation approach constructs constrained clusters in dataset and then result in 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. This method helps in better privacy preservation as compared to other methods as it uses pseudo data rather than modified data.The main disadvantage of this method is loss of sensitive data occurs.

3.4 Anonymization Approach :- 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.

This is a kind of generalization of some attributes which protects against identity disclosure. Anonymization can be obtained through techniques such as generalization, suppression, data removal, permutation, swapping etc. k-anonymity method is treated as the classical anonymization method and most of the studies are based on k-anonymity.

Although the anonymization techniques ensure that the transformed data is true but suffers heavy information loss. Moreover it is not immune to homogeneity attack and background knowledge attack practically. Limitations of the k-anonymity model stem from the two conventions. First, it may be very tough for the owner of a database to decide which of the attributes are available or which are not obtainable in external tables. The second limitation is that the k-anonymity model adopts a certain method of attack, while in real situations; there is no reason why the attacker should not try other methods. However, as a research direction, kanonymity in combination with other privacy preserving methods

needs to be investigated for detecting and even blocking k-anonymity violations.

3.5 Hybrid Approach :- There are many algorithms proposed to achieve privacy preservation of sensitive data. Hybrid approach is the combination of two or more approaches to preserve sensitive data.

S. Lohiya and L. Ragha combined randomization and generalization techniques together [8]. First Sensitive data are randomized and then these randomized data are generalized. The original data can also be retrieved.

4. Privacy Preserving Using Geometric Transformation

4.1 Translation Based Pertrubation :- In this method the noise term applied to each confidential attribute is constant and can be either positive or negative. The set of operations takes only the value {Add} corresponding to an additive noise applied to each confidential attribute.

Translated value = Original value + Noise.

4.2 Rotation Based Perturbation:-In this method [9] a rotation matrix is used to rotate two attributes at a time. For the sake of simplicity a 2D rotation matrix is considered. The rotation of a point by an angle θ in a 2D discrete space can be seen as a matrix representation V_ = Ro(θ)×V, where V is the column vector containing the original coordinates, and V_ is a column vector whose coordinates are rotated coordinates and Ro(θ) is a 2×2 rotation matrix,

$$R0(\theta) = \begin{vmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{vmatrix}$$

4.3 Shearing Based Perturbation: - A noise is added to each confidential data in the Shearing Data Perturbation Method Such that each confidential data will takes only multiplicative noise.

Sheared value = Original value + (Noise * Original value)

5. Evaluation Criteria of Privacy Preserving Algorithm

Privacy preserving data mining an important characteristic in the development and evaluation of algorithms is the identification of suitable evaluation criteria and the development of related principles. In some case, there is no privacy preserving algorithm exists that beats the other entire algorithm on all possible measures. Relatively, an algorithm may perform better that another one on specific measures, like performance and/or data utility. It is vital to deliver users with a set of metrics which will allow them to select the best suitable privacy preserving technique for the data; with respect to some specific parameters. An introductory list of evaluation parameters to be used for evaluating the quality of privacy preserving data mining algorithms is given below:

(i) Performance: the performance of a mining algorithm is measured in terms of the time required to achieve the privacy criteria.

(ii) Data Utility: Data utility is basically a measure of information loss or loss in the functionality of data in providing the results, which could be generated in the absence of PPDM algorithms.

(iii) Uncertainty level: It is a measure of uncertainty with which the sensitive information that has been hidden can still be predicted.

(iv) Resistance: Resistance is a measure of tolerance shown by PPDM algorithm against various data mining algorithms and models.

Table 1. Advantages and Limitations of PPDM Techniq**ues**

Technique	Advantages	Limitations
Anonymization based PPDM	Identity or sensitive data about record owners are to be hidden.	Linking attack. Heavy loss of information.
Perturbation based PPDM	In this technique different attributes are preserved independently.	Original data values cannot be regenerated. Loss of information.

Condensation Approach based PPDM	Use pseudo data rather than altered data. This method is very real in case of stream data.	Huge amount of information lost. It contain same format as the original data.
Cryptography Based PPDM	Transformed data are exact and protected. Better privacy compare to randomized Approach.	This approach is especially difficult to scale multiple parties are involved.

III. CONCLUSION

As people are very concerned about sharing their sensitive data while using different applications, many Privacy Preserving Data Mining techniques are developed. These techniques are used to protect sensitive data from unauthorized access and prevented from being misused.

In this paper we focused on the existing Privacy Preserving Data Mining techniques and also analyzed their advantages and disadvantages. According to this survey, there is no technique which is consistent and applicable in all domains. The above two Hybrid approaches are more efficient than that of other mentioned techniques.

IV. REFERENCES

[1] J. Han and M. Kamber, "Data Mining: Concepts and Techniques", 2nd edition.

[2] M. B. Malik, M. A. Ghazi and R. Ali, "Privacy Preserving Data Mining Techniques: Current Scenario and Future Prospects", in proceedings of Third International Conference on Computer and Communication Technology, IEEE 2012.

[3] H. Kargupta and S. Datta, Q. Wang and K.

[4] Sivakumar, "The Privacy Preserving Properties of Random Data Perturbation Techniques", in proceedings of the Third International Conference on Data Mining, IEEE 2003.

[5] T. Jahan, G.Narsimha and C.V Guru Rao, "Data Perturbation and Features Selection in Preserving Privacy" in proceedings of 978-1-4673-1989-8/12, IEEE 2012.

[6] Y. Lindell, B.Pinkas, "Privacy Preserving Data mining", in proceedings of Journal of Cryptology, 5(3), 2000.

[7] A. Parmar, U. P. Rao, D. R. Patel, "Blocking-Based Approach for Classification Rule hiding to Preserve the Privacy in Database", in proceedings of International Symposium on Computer Science and Society, IEEE 2011.

[8] C. Aggarwal , P.S. Yu, "A Condensation Approach to Privacy Preserving Data mining", in proceedings of International Conference on Extending Database Technology (EDBT), pp. 183–199, 2004.

[9] S. Lohiya and L. Ragha, "Privacy Preserving in Data Mining using Hybrid Approach", in proceedings of Fourth International Conference on Computational Intelligence and Communication Networks, IEEE 2012.

[10] P.Deivanai, J. J. Nayahi and V.Kavitha," A Hybrid Data Anonymization integrated with Suppression for Preserving Privacy in mining Multi-Party Data" in proceedings of International Conference on Recent Trends in Information Tech