Review of Privacy Preserving Data Mining Techniques

Jyothi Mandala[#1], Suneetha Merugula[*2]

[1]*Asstistant Professor, GMRIT, Rajam, Srikakulam. AP, INDIA*
jyothirajb4u@gmail.com
[2]*Asstistant Professor, GMRIT, Rajam, Srikakulam. AP, INDIA*
sunita.merugula@gmail.com

Abstract— Data Mining is a process of discovering useful information in large data repositories. Data mining techniques have been used to enhance information retrieval systems. Privacy-preserving data mining (PPDM) refers to the area of data mining that is primarily concerned with protecting against disclosure of individual data records i.e., to safeguard sensitive information. To address about privacy researchers in data mining community have proposed various solutions. In this paper we present an extensive review of all privacy preserving data mining (PPDM) techniques. We use a classification scheme, which is adopted from earlier studies, to review the techniques.

Keywords— *PPDM, privacy; data privacy, Data Mining, data perturbation*

1. Introduction

Privacy preserving Data Mining concept is explained in detail by R. Agarwal and R. Srikant in 200[2]. Two or more parties owning confidential datasets wish to run a data mining algorithm on the union of their datasets without revealing any unnecessary information. In the information technology era, privacy refers to the right of users to conceal their personal information and have some degree of control over the use of any personal information disclosed to others. In paper [1] PPDM is defined as “getting valid data mining results without learning the underlying data values”. PPDM encompasses the dual goal of meeting privacy requirements and providing valid data mining results.

PPDM can be one of the three approaches: 1. Data hiding, in which sensitive raw data like identifiers, name, addresses, etc were altered, blocked or trimmed out from the original database in order for the users of the data not to be able to compromise another person’s privacy. 2. Rule hiding, in which sensitive data extracted from the data mining process be excluded for use, because confidential information may be derived from the released knowledge and 3. Secure multiparty computation (SMC), where distributed data are encrypted before released or shared for computations, so that no party knows anything except its own inputs and the results.

1.1 Data mining algorithms

Association analysis involves the discovery of association rules, showing the attribute value and conditions that are frequently in a given set of data.

Classification is the process of finding a set of models that describe and distinguish data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown.

Clustering analysis is a process of decomposing the data set into groups so that points in one group are similar to each other and are different from the points in other groups.

2. Classification of existing PPDM techniques

PPDM techniques are mainly classified into three major categories: data partitioning, data modification and data restriction as shown in the below Fig 1.

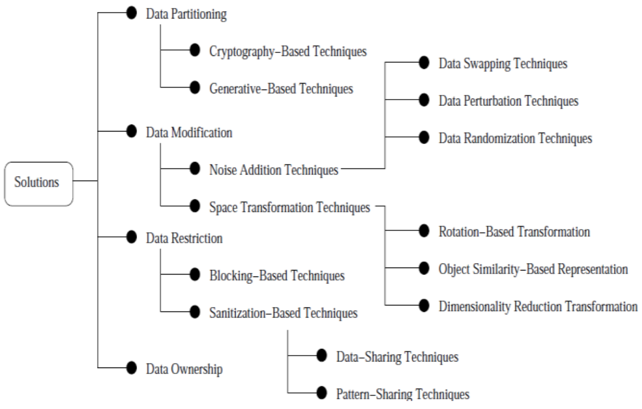
Solutions
Data Partitioning
Cryptography–Based Techniques
Generative–Based Techniques
Data Swapping Techniques
Data Perturbation Techniques
Data Randomization Techniques
Data Modification
Noise Addition Techniques
Space Transformation Techniques
Rotation–Based Transformation
Object Similarity–Based Representation
Dimensionality Reduction Transformation
Data Restriction
Blocking–Based Techniques
Sanitization–Based Techniques
Data Ownership
Data–Sharing Techniques
Pattern–Sharing Techniques

Fig1: classification of PPDM techniques

2.1 Data Partitioning Techniques:

These types of techniques are useful when databases used for mining are available in different locations. The data may be distributed either horizontally or vertically.

Horizontal Partitioning: Same schema is followed in all partitions Ex: Medical insurance records.

Vertically Partitioning: Attributes of the same entity are split across the partitions

Techniques are:

Cryptography-Based Techniques:

This technique is used if two or more parties want to perform data mining task on combined datasets. This problem is referred as Secure Multi-party Computation (SMC) problem. By using cryptographic techniques we can perform privacy preserving classification [3], privacy preserving association rule mining [4] and privacy preserving clustering [5].

Generative-Based Techniques:

In this technique instead of sharing the complete data set it such share a small portion of its local model which is used to construct the global model. This technique is used in horizontally partitioned data. By using this technique we can perform privacy preserving clustering [6].

2.2 Data Modification Techniques:

These types of techniques modify the original values of a database. This modified database will be used for data mining task. Techniques are:

Noise Addition Techniques: In this technique some noise is added to the original data to prevent the identification of confidential information relating to a particular individual. To hide the patterns the attribute values can be randomly shuffled.

These are of two types:

Data swapping techniques: In this technique the value of individual records will be interchanged [7].

Data distortion techniques: In this technique some distortion will be added to the data [8]. Most commonly used distortions are Uniform distribution over an interval $[-\infty, \infty]$ and Gaussian distribution with mean $\mu=0$.

Space Transformation Techniques: These techniques are used to protect the data values used to perform clustering without affecting the similarity between objects under analysis.

These are of three types:

Rotation Based Transformation: In this technique the idea is to that the attributes of a database are split into pair wise attributes selected randomly. One attribute can be selected and rotated more than once and the angle Θ between an attribute pair is also selected randomly. This technique is used for cluster analysis [9].

Object similarity based representation: In this technique the data owner share some data for clustering analysis by simply computing the dissimilarity matrix between the objects and share that with a third party.

Dimensionality reduction based transformation: This technique can be used when the attributes of objects are available at a central repository or split across many sites. By reducing the dimensionality of the database we can preserve the distance between data points i.e. preserving the similarity between data points.

Data Restriction Techniques These techniques are used to limit the access to mining results. This will perform either generalization or suppression (sanitization-based techniques of information or blocking the access to some patterns (blocking based techniques). These techniques are used for privacy preserving association rules and clustering rules [10].

Data Ownership Techniques

These techniques are used for two purposes:

To protect the ownership of data by people about whom the data were collected [11].

To identify the entities that receives confidential data when such data are shared or exchanged [12].

2.3 Data Distribution:

Privacy preserving Data Mining Algorithms can be applied on centralized database or distributed database. In centralized database environment data are all stored in a single database, where as in distributed database environment data is distributed in different locations. On centralized Data base we can perform Data Hiding and Rule Hiding privacy, whereas on Distributed database we can perform data hiding privacy only. To perform data hiding can be performed using classification, clustering and association rules algorithms. To perform rule hiding also we can use classification, clustering and Association rules algorithms [13].

Data hiding refers to the cases where sensitive data from original dataset like identity, name, and address that are linked directly or indirectly to an individual person to hide.

Rule hiding refers to the removal of sensitive knowledge derived from original databases after applying data mining algorithms.

The privacy preserving techniques used in a distributed database is mainly based on cryptography techniques.

3. Conclusion

PPDM is emerged as a new field of study. In this paper we have surveyed different privacy preserving data mining techniques. And we also identified which type of privacy will be provided by using which type of data mining technique. For a new comer, this paper provides a brief review about existing privacy preserving techniques.

References

[1] C. Clifton, M. Kantarcioglu, and J. Vaidya. Defining Privacy For Data Mining. In Proc. of the National Science Foundation Workshop on Next Generation Data Mining, pages 126-133, Baltimore, MD, USA, November 2002.

[2] R. Agarwal and R. Srikanth. Privacy- Preserving Data mining. In Proc. of the 2000 ACM SIGMOD International Conference on Management of Data, pages 439-450, Dallas, Texas, May 2000.

[3] M. Kantarcioglu and J. Vaidya. Privacy Preserving Naïve Bayes Classifier for Horizontally Partitioned Data. In Proc. of the IEEE ICDM Workshop on Privacy Preserving Data Mining, Pages 3-9, Melbourne, FL, USA, November 2003.

[4] J. Vaidya and C. Clifton. Privacy Preserving Association Rule Mining in Vertically Partitioned Data. In Proc. of the 8th ACM SIGKDD Intl. Cong. On Knowledge Discovery and Data Mining, Pages 639-644, Edmonton, AB, Canada, July 2002.

[5] J. Vaidya and C. Clifton. Privacy Preserving K-Means Clustering Over Vertically Partitioned Data. In Proc. of the 9th ACM SIGKDD Intl. Cong. On Knowledge Discovery and Data Mining, Pages 206-215, Washington, DC, USA, August 2003.

[6] S. Meregu and J.Ghosh. Privacy- Preserving Distributed Clustering using Generative Models. In Proc. of the 3rd IEEE International Conference on Data Mining (ICDM'03), pages 211-218, Melbourne, Florida, USA, November 2003.

[7] V. Estivill-Castro and L. Brankovic. Data Swapping: Balancing Privacy Against Precision in Mining of Logic Rules. In Proc. of Data Warehousing and Knowledge Discovery DaWaK-99, pages 389-398, Florence, Italy, August 1999.

[8] C. W. wu. Privacy Preserving Data Mining: A Signal Processing Perspective and A Simple Data Perturbation Protocol. In Proc. of the IEEE ICDM Workshop on Privacy Preserving Data Mining, pages 10-17, Melbourne, FL, USA, November 2003.

[9] S. R. M. Oliveira and O.R. Zaiane. Achieving Privacy Preservation When Sharing Data for Clustering. In Proc. of the Workshop on Secure Data Management in a Connected World (SDM'04) in conjuction with VLDB'2004, pages 67-82, Toronto, Ontario, Canada, August 2004.

[10] Y. Saygin, V. S. Verykios, and C. Clifton. Using Unknowns to Prevent Discovery of Association Rules. SIGMOD Record, 30(4):45-54, December 2001.

[11] A. P. Felty and S. Matwin: Privacy- Oriented Data Mining by Proof Checking. In Proc. of the 6th European Conference on Principals of Data Mining and Knowledge Discovery (PKDD), Pages138-149, Helsinki, Finland, August 2002.

[12] A. Mucsi-Nagy and S. Matwin. Digital Fingerprinting for Sharing of Confidential Data. In Proc. of the Workshop on Privacy and Security Issues in Data Mining, pages 11-26, Pisa, Italy, Septe,ber 2004.

[13] Xiaodan Wu, Yunfeng Wang, Chao-Hsien Chu. A Close Look at Privacy Preserving Data Mining Methods. In Proc. of The Tenth Pasific Asia Conference on Information Systems (PACIS 2006), pages167-173.

M.Jyothi received B.Tech degree in Computer Science and Information Technology from AITAM College, Tekkali, India and M.Tech degree in Computer Science and Engineering from JNTUK, Kakinada, India. She is pursuing her Ph.D from Acharya Nagarjuna University Guntur, in the area of Data Mining. Currently she is working as Assistant professor in Information Technology Department at GMRIT, Rajam.

M.Suneetha received B.Tech degree in Computer Science and Information Technology from AITAM College, Tekkali, India, M.Tech degree in Software Engineering from JNTUH, Hyderabad, India. She is pursuing her Ph.D from Acharya Nagarjuna University Guntur, in the area of Data Mining. Currently she is working as a Assistant professor in Information Technology Department at GMRIT, Rajam.