A Survey on Privacy Preserving Association Rule Mining of Outsourced Databases

Priyanka D. Salunkhe
Department of IT, Walchand College of Engineering Sangli, Maharashtra, India
priyanka.salunkhe@walchandsangli.ac.in

Abstract—Data mining finds useful patterns from the large dataset. Data analysis techniques that are frequent itemset mining and association rule mining are two popular and broadly utilized for different applications. Personal or sensitive information of individuals, industries or organizations must be kept private before it is shared for the data mining. Hence privacy preserving data mining has become an important issue in the outsourcing databases. Nowadays, with the data storage and data processing technologies, privacy preservation has been one of the greater concern in the data mining. During last decade, lots of data mining techniques have been proposed. This paper aims to focus on comparative study of privacy preserving data mining techniques.

Index Terms— Association rule mining, frequent itemset mining, privacy preserving data mining.

1 Introduction

DATA mining is the process of finding interesting information from the large datasets. Frequent itemset mining and association rule mining are two data analysis techniques generally used for discovering frequently co-occurring data items and interesting association relationships between data items in large transaction databases. There are many applications in which these techniques are used such as health care [1], market basket analysis [2], prediction [3] etc.

A transaction database is a set of transactions and each transaction is a set of data items with a unique transaction ID (TID).An itemset Z is regarded frequent if and only if $Supp(Z) \geq Ts$, where Ts is a threshold specified by the data miner. $Supp(Z)$ is Z's support, which is defined as Z's occurrence count in the database. An association rule is expressed using $X \Rightarrow Y$, where X and $Y$are two disjoint itemsets. $X \Rightarrow Y$ indicates that X's occurrence implies Y's occurrence in the same transaction with a certain confidence. $X \Rightarrow Y$ is regarded as an association rule if and only if $Supp(XUY) \geq Ts$ and $Conf\ (X \Rightarrow Y) \geq Tc$, where Tc is threshold defined by data miner. Confidence is defined as the chance of the rule's left-hand side when the transaction data also hold on the right-hand side. For Example supermarket's transaction database, where each transaction contains some customer's shopping list. A customer buying "bread" and "butter" will also buy "milk". Then {bread, butter} ⇒ milk is a possible association rule. $X \Rightarrow Y$ is meaningful and useful if the confidence is high and $X \cup Y$ is frequent**.**

Support and confidence of a rule are expressed by Equations:

$$\text{Support}(X \Rightarrow Y) = \frac{\text{Support}\ (X \cup Y)}{N}$$

$$\text{Confidence}(X \Rightarrow Y) = \frac{\text{Support}\ (X \cup Y)}{\text{Support}\ (X)}$$

Where,
X, Y are itemsets,
N is a total number of transactions in the database.

Data owners outsource their data to the joint database to the server, but they may not wish to disclose their private data and they are wishing to learn association rules or frequent itemsets from their joint data. If each data owner has one or more columns in the joint database, the database is considered as vertically partitioned, which is used for secure and storage purpose. If each data owner has one or more rows in the joint database, the database is considered as horizontally partitioned.

In data mining applications, association rule mining discovers item sets with co-occurrence of frequently performed transactions on the database. Privacy concerns are the major important scenario in recent approaches because each owner may not wish to disclose in their own partitioned database that exists relative data efficiency in commercial services. Privacy preserving data mining helps to achieve data mining goals without sacrificing privacy of the owner.

Therefore, the data owners can outsource their encrypted data and mining task to a server in privacy preserving manner.

2 Significance of Survey

A number of methods have been proposed for the privacy-preserving association rule mining in the distributed database. Methods such as vertical partitioning, horizontal partitioning, random data perturbation, cryptography are designed for preserve private information.

M. Zaki [4] designed classic frequent itemset mining and association rule mining algorithms for a centralized database setting where the raw data is stored in the central site for mining, but privacy concerns are not considered in the setting.

J. Vaidya and C. Clifton [5], has proposed the first work to identify and address privacy issues in vertically partitioned databases, a secure scalar product protocol is presented and is used to build a privacy-preserving frequent itemset mining solution. Association rules can then be found from the given frequent itemsets and their supports.

M. Kantarcioglu and C. Clifton [6] proposed various works to modify the statistical distribution of data by implementing substitution ciphers and adding fake transactions. In this solution following encoding steps are proposed:
(i) '1-to-n' mapping of individual original items.
(ii) Additional unique and common items are added to the Transaction-level mappings.
(iii) The addition of fake items to each transaction.

However, the processing, and especially storage overheads are quite considerable. The '1-to-n'mappings mean that the data storage required is increased by quite a number of folds.

B. Rozenberg and E. Gudes [7] proposed solution for the association rule mining which follows collaborative Apriori algorithm for the vertically partitioned database. In this solution, a data owner which is responsible for the mining is known as master and other data owners known as slaves.

This solution utilizes semi-trusted third party server to compute the results. Each side (slaves) sends his set of real transaction ID's to the server. The server calculates the size of the intersection of the sets and if the size is greater than or equal to the minimum support, then sends a message to each side that mining is possible. Then each side adds fake transactions to their individual datasets in the preprocessing stage and sends the datasets to the master. The master generates association rule candidates from the joint database. For each generated rule candidate $X \Rightarrow Y$, the master sends transaction ID list of $X \cup Y$ and X to the server. Then the server verifies whether the rule is qualified or not and send the answer to the master. But this solution results in the leakage of sensitive information as the private dataset is not encrypted.

S. Zhong [8] proposed solution for both horizontally and vertically partitioned database. Vertically partitioned data means each party has one or more attributes in the joint database and horizontally partitioned data means each party has one or more records in the joint database with the same number of attributes. Private data of the data owner is encrypted using asymmetric encryption. Also, asymmetric homomorphic encryption is used to compute support. There are two privacy-preserving solutions for frequent itemset mining is proposed. One of the solution does not give exact support. Using this solution to find out frequent itemsets, there is need to test each candidate itemset individually. This solution is applied only for frequent itemset mining. Association rules cannot be mined based on the result of that solution because confidence cannot be computed without exact support.

N. Muthu Lakshmi and K. Sandhya Rani [9] proposed a model to find association rules for vertically partitioned databases considering the privacy constraints with 'n' number of sites along with data miner. This model compromises different cryptography techniques such as encryption, decryption and scalar product technique to find association rules efficiently and securely for vertically partitioned databases.

F. Giannotti *et al*. [10] proposed a solution which is based on k-anonymity frequency. To counter frequency analysis attack, the data owner inserts fake transactions in the database to hide the item frequency. Items in the database are encrypted with the 1-1 substitution cipher. After inserting the fake transactions, any item in the encrypted database will share the same frequency with at least $k - 1$ other items. Then dada owners outsource their database to the server for the mining task. The server runs frequent itemset mining algorithm and returns the resulted frequent itemsets and their supports to the data owner. The data owner revises these itemsets' supports by subtracting them with itemsets' corresponding occurrence count in the fake transactions respectively. Then, the data owner decrypts the received itemsets with the revised supports higher than the frequency threshold and generates association rules based on the frequent itemsets. In these setting, data owner requires counting itemset occurrences in fake transactions to cancel out fake transactions. Using this technique for the vertically partitioned database, data owners are unable to perform such calculations.

J. Lai *et al.* [11] proposed a privacy preserving outsourced association rule mining solution. This solution is vulnerable to frequency analysis attacks. Applying this solution to vertically partitioned databases will result in the leakage of the exact supports to data owners.

T. Tassa [12] proposed for secure mining of association rules in horizontally distributed databases. The proposed protocol is based on the fast distributed algorithm, which is an unsecured distributed version of Apriori algorithm. The protocol computes the union (or intersection) of private subsets that each of the interesting site hold. Also, the protocol tests the inclusion of an element hold by one site in subset held by another. But this solution is only applicable for horizontal partitioning, not for vertical partitioning.

Lichun Li *et al*. [13] proposed a privacy-preserving association rule mining solution for outsourced vertically partitioned databases. In such a scenario, data owners wish to learn the association rules or frequent itemsets from a collective data set and disclose as little information about their (sensitive) raw data as possible to other data owners and third parties. Symmetric homomorphic encryption technique is used for computation of support and confidence which ensures the privacy of the data and mining result also.

In vertical partitioning, splitting the database into different sites in such a way that each site contains some

number of columns.

For example, consider a database with 'n' rows and 'm' columns and partition the database into 3 sites with vertically then,

Partition 1: 'n' rows, 'a' columns
Partition 2: 'n' rows, 'b' columns
Partition 3: 'n' rows, 'c' columns
Where a+b+c=m [14]

While in horizontal partitioning, split the rows, keeping the same number of columns.

With the above example, horizontal partitioning may look like,

Partition 1: 'n/3' rows, 'm' columns
Partition 2: 'n/3' rows, 'm' columns
Partition 3: 'n/3' rows, 'm' columns [14]

In the horizontal partitioning, one can gain entire information of the particular row, but in vertical partition entire information cannot be obtained as columns are split across the sites. Therefore, vertical partitioning is better for security purpose.

Table I compares various methods used in earlier work for privacy preserving in data mining.

Considering distributed database, most of the work is done on the vertically partitioned databases. In some recent work, cryptography and random data perturbation are used together for improving the privacy. However, few solutions do not arrive at association rule result based on the frequent itemset.

TABLE 1
Methods used for privacy preserving mining

METHODS	REFERENCES								
	5	6	7	8	9	10	11	12	13
Random data perturbation		√	√			√			√
Cryptography		√		√	√	√	√		√
Horizontally partitioned distribution		√		√			√	√	
Vertically partitioned distribution	√		√	√	√	√			√
Frequent itemset mining	√		√	√		√		√	√
Association rule mining	√	√	√		√	√	√	√	√

3 REMARKS

Privacy is the major concern in the mining to protect the sensitive/private data. In distributed database, the database is partitioned in a horizontal or vertical manner for secure storage purpose. So that an attacker cannot get entire data at one site. But in the partitioned database also there is a need of protection of data. This paper focuses on the comparative study of privacy preserving data mining techniques used in the horizontally or vertically partitioned databases. From the comparative study, it is proved that cryptography and random data perturbation methods outperform in preserving privacy. Cryptography is the best technique for encryption of sensitive data. Data perturbation is used to disturb the data or records in the original database so it helps to preserve data and hence privacy is maintained. In some literature, a homomorphic encryption technique is used to compute support and confidence which is required for association rule mining along with encryption of private data and data perturbation technique. Homomorphic encryption improves the privacy of the data and mining result in terms of accuracy. Literature survey techniques for encryption of dataset uses 1-1 substitution cipher. But 1-1 substitution cipher suffered from frequency analysis attack so, research work will be a focus on Advanced Encryption Technique (AES) implementation to improve the security of data.

REFERENCES

[1] S. E. Brossette, A. P. Sprague, J. M. Hardin, K. B. Waites, W. T. Jones, S. A. Moser, "Association Rules and Data Mining in Hospital Infection Control and Public Health Surveillance," *J. Amer. Med Inform. Assoc.*, vol. 5, no. 4, pp. 373-381, 1998.

[2] T. Brijs, G. Swinnen, K. Vanhoof, and G. Wets "Using Association Rules for Product Assortment Decisions: A case study," in *Proc SIGKDD*, pp. 254-260, 1999.

[3] X. Yin and J. Han, "CPAR : Classification based on Predictive Association Rules," in *Proc. SIAM SDM*, pp. 1-5, 2003.

[4] M. J. Zaki, "Scalable Algorithms for Association Mining," *IEEE Trans.Knowl. Data Eng.*, vol. 12, no. 3, pp. 372-390, May/Jun. 2000.

[5] J. Vaidya and C. Clifton, "Privacy preserving association rule mining in vertically partitioned data," in *Proc. SIGKDD*, pp. 639-644, 2002.

[6] M. Kantarcioglu and C. Clifton, "Privacy-Preserving Distributed Mining of Association Rules on Horizontally Partitioned Data," *IEEE Trans. Knowl. Data Eng.*, vol. 16, no. 9, pp. 1026-1037, Sep. 2004.

[7] B. Rozenberg, E. Gudes , "Association Rules Mining in Vertically Partitioned Databases," *Data Knowl. Eng.*, vol. 59, no. 2, pp. 378-396, 2006.

[8] S. Zhong, "Privacy-Preserving Algorithms for Distributed Mining of Frequent Itemsets," *Inf. Sci.,* vol. 177, no. 2, pp. 490–503, 2007.

[9] N. V. Muthu Lakshmi1 & K. Sandhya Rani, "Privacy Preserving Association Rule Mining in Vertically Partitioned Databases," In *IJCSA*, vol. 39, no. 13, pp. 29-35, Feb. 2012.

[10] F. Giannotti, L. V. S.Lakshmanan, A. Monreale, D. Pedreschi, and H. Wang, "Privacy-Preserving Mining of Association Rules from Outsourced Transaction Databases," *IEEE Syst. J.*, vol. 7, no. 3, pp. 385-395, Sep. 2013.

[11] J. Lai, Y. Li, R. H. Deng, J. Weng, C. Guan, and Q. Yan, "Towards Semantically Secure Outsourcing of Association Rule Mining on Categorical Data," *Inf. Sci..*, vol. 267, pp. 267-286, May 2014.
[12] T. Tassa, "Secure Mining of Association Rules in Horizontally Distributed Databases Scalable Algorithms for Association Mining," *IEEE Trans.Knowl. Data Eng.,* vol. 26, no. 4, Apr. 2014.
[13] L. Li, R. Lu, S. Member, K. R. Choo, and S. Member, "Privacy-Preserving-Outsourced Association Rule Mining on Vertically Partitioned Databases," *IEEE Trans. Info. Foren. Secur.,* vol. 11, no. 8, pp. 1847–1861, Aug. 2016.
[14] Partitioned database. [Online] Available: https://en.wikipedia.org/wiki/Partition_(database)