A Survey on Privacy Preserving Association Rule Mining

Lili Zhang [1,2]
1.National Key Laboratory of Integrated Services Networks
1.Xidian University,Xi'an, China
2.Information Engineering College
2.Henan University of Science and Technology
Lillyzh@126.com

Danmei Niu ,Yuxiang Li, Zhiyong Zhang
Information Engineering College Henan University of Science and Technology,Luoyang,China
niudanmei@163.com
liyuxiang@haust.edu.cn
xidianzzy@126.com

***Abstract*—In recent years, privacy preserving association rule mining (PPARM) has emerged as a new research area and captured the attention of many researchers who are interested in preventing the privacy violations occurring in during association rule mining. A detailed survey of the present methodologies for the privacy preserving association rule data mining and a review of the state of art method for privacy preserving association rule mining is presented in this paper. An analysis is provided based on the association rule mining algorithm techniques. Finally, the authors obtain some conclusions and come with up future direction.**

Keywords- association rule mining; privacy preserving; data mining; association rule mining

1. INTRODUCTION

Data Mining (DM) can be considered as a particular type of knowledge discovery process. It can be defined as the analysis of observational data sets to discover relevant information and to summarize the data in novel ways, understandable and useful to the owner.

Lots of work has been done on the DM area. DM techniques can be divided into two classes: predictive techniques and descriptive techniques [1]. Predictive DM focuses on making predications by historical data. Descriptive DM focus on mining potential rules hidden in the big data set without having any predefined target. DM techniques classification is shown in Fig. 1.

Among a large number of DM algorithms, ARM is one of most common algorithms. It is descriptive DM and aims to discover interesting relationships among sets of items in the transaction databases. Association rules are widely used in various areas such as telecommunication networks, market basket analysis，risk management, health care, web usage mining etc.

Recent advances in ARM have generated controversial impact in both scientific and technological arenas. On the one hand, ARM is capable of uncovering the potential useful rules in large volume of data. On the other hand, the excessive processing of mining the association rules puts the sensitive data and the user's confidential information at risk. Therefore, on the premise of ensuring data security and user privacy, developing privacy preserving association rule mining algorithms becomes an especially requested issue. In recent years, a great many PPARM [2-3] algorithms have been developed. In this paper, we aim to make a comprehensive survey of PPARM algorithm.

The remainder of this paper is organized as follows. In Section II, we formalize the definition of ARM. A detailed survey of the present methodologies for PPARM is provided in Section III. In Section IV, we make a review of PPARM algorithms. In Section V, we conclude the paper and give the future research direction.

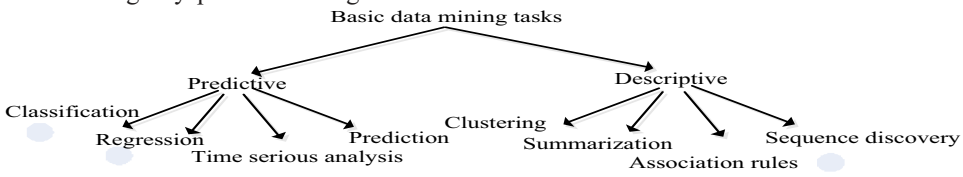

Figure 1 Data mining techniques classification

I. PROBLEM DEFINITION OF ARM

ARM algorithms are designed to discover relevant relationships between the variables of a dataset. The problem definition for the association rule mining is stated as follows: Given $I = \{i_1 , i_2 , \dots , i_m \}$ is a set of items, $T = \{t_1, t_2 , \dots , t_n\}$ is a set of transactions. Each transaction t_i is a set of items such that $t_i \subseteq I$. An association rule is an implication of the form: $X \Rightarrow Y$, where $X \subset I$, $Y \subset I$ and $X \cap Y = \emptyset$. X (or Y) is a set of items, called itemset. In the rule $X \Rightarrow Y$, X and Y are called the antecedent and consequent of the rule respectively. It is obvious that the antecedent's

978-1-5386-5500-9/18/$31.00 ©2018 IEEE
DOI 10.1109/ICISCE.2018.00029

occurrence implies the consequent's occurrence in the same transaction with a certain confidence. Support and confidence are the two most important quality measures for evaluating the interestingness of a rule. Support of an association rule is defined as the percentage of records that contain $X \cup Y$ to the total number of records in the database. The support of a rule is represented by formula (1)

$$Support(X \Rightarrow Y) = \frac{\sigma(X \cup Y)}{n} \quad (1)$$

Where $\sigma(X \cup Y)$ is the number of transactions that contain all the items of the rule and n is the total number of transactions. The confidence of an association rule is defined as the percentage of the number of transactions that contain X and Y to the total number of records that contain X. Confidence of a rule is represented by the formula (2).

$$Confidence(X \Rightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)} \quad (2)$$

ARM is decomposed into two different sub problems.

(1) Frequent Itemset Mining Phase: Find all combinations of items whose support is greater than a user-defined Minimum Support Threshold.

(2)Association Rule Generation Phase: Use the frequent itemset to generate the association rules whose confidence is greater than a user-defined Minimum Confidence Threshold.

II. CLASSIFICATION OF PPARM TECHNIQUES

Many methods have been developed to solve different aspects of the PPARM problem. We make a reference to the classification method of privacy preserving data mining by Verykios [4] and try to classify them by the following four dimensions: (1) data distribution; (2) data modification technologies; (3) data content that needs to be protected; (4) privacy preservation technologies.

The first dimension refers to the data distribution. Some of the approaches have been developed for centralized data. In this model, there is only one site whose data are published to the excited site. While others are designed for distributed data scenarios, which also be classified as horizontal data distribution and vertical data distribution. Horizontal data distribution refers to those cases where different records are collected at different sites, but each record contains all of the attributes for the object. Vertically data distribution where different attributes of the same set of records are collected at different sites. Each site collects the values of one or more attributes for each record.

The second dimension refers to the data modification method. Data modification is used to modify the original data to ensure privacy protection. Three types of data modification methods are as follows.

1) data randomization and anonymization, which are used to obscure data.
2) hiding techniques, which include perturbation, blocking, aggregation, etc. Perturbation refers to the alteration of an attribute value by a new value (i.e., changing a 1-value to a 0-value, or adding noise). Blocking is the replacement of an existing attribute value with a "?". Aggregation is the combination of several values into a coarser category.
3) encryption, which changes the original data into ciphertext .

The third dimension refers to the data content that needs to be protected from disclosure. Some PPARM algorithms are developed to protect raw data, while others are developed to protect the association rules.

The fourth dimension refers to the privacy preservation technique used for the modification of the original data. The developed techniques can be divided into:

- data obscure-based techniques, which are used in data sanitization to protect the some sensitive items.
- heuristic-based techniques like adaptive modification that modifies only selected values that minimize the utility loss rather than all available values.
- cryptography-based techniques like secure multiparty computation where a computation is secure.
- reconstruction-based techniques where the original distribution of the data is reconstructed from the randomized data.

Fig. 2 shows the taxonomy of the existing PPARM.

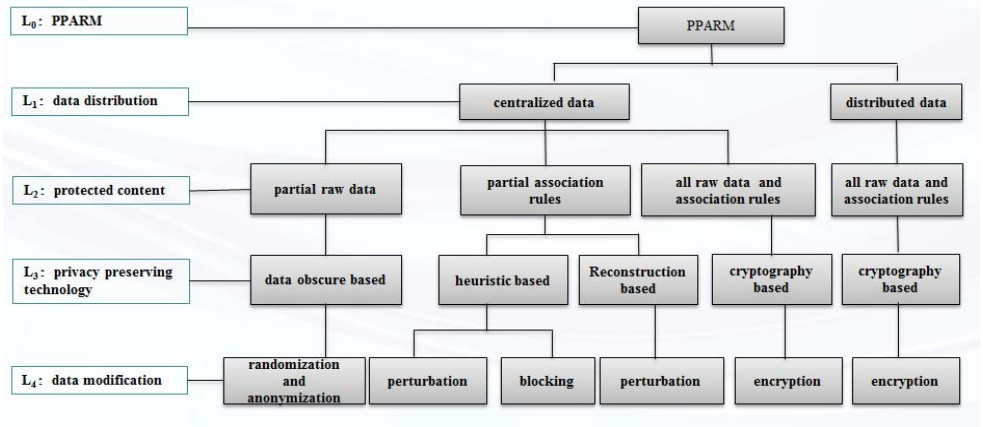

Figure 2. A taxonomy of the developed PPARM algorithms

III. REVIEW OF PPARM ALGORITHMS

A. Data Obscure-based raw data protection techniques

In this type of model, the owner site obscures or removes its data before he publishes its data to the external site. Two well-known methods are data randomization and data anonymization. Rizvi et al. [5] proposed randomization-based mask scheme. In addition, Rizvi et al. presented optimization method to decrease high computational overhead in the frequent itemset mining. However, data randomization methods have the risk that public records may be utilized to discover an identity in the sanitized dataset records. Data anonymization is a raw data protection method which avoids the weakness of randomization methods. Data anonymization aims to prevent the identification of individual records. In general, data anonymization functions as follows: (1) remove the record identifier; (2) anonymize the quasi identifier attributes which can identify the record owner. K-anonymity proposed by Sweeney [6] is the most popular anonymization privacy model and aims to generate a K-anonymous table such as each record is indistinguishable from at least K-1 other records of the quasi identifier.

B. Heuristic-based Association Rule hiding (ARH)Techniques

1) Perturbation-based ARH over centralized data

ARH [7] focuses on transforming the raw data such that certain sensitive association rules cannot be discovered from the published dataset while other rules can.

Atallah et al. [8] firstly proposed the heuristic-based ARH scheme, which is an itemset-based algorithm which hides all the sensitive association rules by reducing their supports. In [8] the optimal sanitization was shown to be an NP-Hard problem.

Elena et al. [9] extended the sanitization of sensitive large itemsets to the sanitization of sensitive rules. In this algorithm, either reduction of support or reduction of confidence is used to hide the sensitive rules. This algorithm has two disadvantages: hiding only one rule at a time and generating ghost rules, which reduces the utility of the released database.

On the premise of ensuring security, later heuristics work is expected to consider the utility issues. Based on this idea, a complete work is done by Verykios [10].

Amiri et al. [11] proposed an itemset-based algorithm with aggregate approach and hides multiple rules at one time. The aggregate approach works as follows. The transaction that supports the most sensitive itemsets and the least non-sensitive itemsets is removed.

Wang et al. [12] proposed a rule-based ARH algorithms with sanitization method to protect informative association rules. In this paper, the sanitization method is to increase support of LHS and decrease support of RHS. By decreasing support and confidence (DSC), Wang et al. [13] proposed another ARH algorithm, which requires less running time and fewer transaction modifications than the algorithm in [12].

Based on MATRIX APRIORI algorithm, Yildiz et al. [14] proposed the itemset-hiding algorithm, which integrated the post-mining with the hiding algorithm. Thus, the output of this algorithm is different from all of the previously mentioned algorithms. A sanitized dataset and a frequent itemset of a sanitized dataset are the output of this algorithm.

2) Blocking-based ARH over centralized Data

Blocking refers to replacing certain attributes of some data items with a question mark. It is sometimes more desirable for specific applications (i.e., medical applications) to substitute an unknown value for a real value rather than placing a false value.

The first blocking-based approach was proposed in [15]. Introducing a question mark in the dataset, changes the definition of the support and confidence of an association rule to some extent. In this respect, the minimum support and minimum confidence will be altered into a minimum support interval and a minimum confidence interval correspondingly.

Later, Wang et al [16] proposed a blocking-based sanitization algorithm. Compared with the algorithm in [15], this approach achieves further efficiency; however, it must hide all rules containing the hidden items on the left hand side, where Saigon's approach can hide any specific rule.

C. Reconstruction-Based Techniques over centralized data

Many proposed techniques address privacy preservation by reconstructing the distributions at an aggregate level in order to perform the mining. The reconstruction approach was firstly proposed in the context of ARM by Chen et al. [17]. In order to hide sensitive frequent itemsets, they give a coarse Constraint-based Inverse Itemset Lattice Mining. Inspired by the idea of [15], Guo et al. [18] proposed a FP tree based algorithm for inverse frequent set mining to reconstruct the original database by using non characteristic of database. This algorithm can work more efficiently than the algorithm in [17] and other inverse frequent set mining algorithms.

D. Cryptography-Based Techniques

Cryptography-based approaches are mainly used to solve secure association rule mining outsourcing over the centralized data and Secure Multiparty Computation (SMC) problem over distributed data.

1) Secure ARM outsourcing over centralized data.

Privacy-preserving outsourced frequent itemset mining and association rule mining have been studied in the setting of a single data owner [19-21].

Wong et al. [19] proposed a solution to counter frequency analysis attack based on a one-to-n item mapping. However, it was lacking a formal theoretical analysis of privacy Guarantees. Molloy et al. [20] demonstrated that the

solution can't withstand known frequency analysis attacks and proposed alternatives. The success of the attacks in [20] mainly relies on the existence of fake items, defined in [19]. [20] showed that the random "fake" items can be removed by detecting the low correlations between items, and that top frequent items can be re-identified by attackers; Giannotti et al. [21] proposed similar methods k anonymity frequency for both raw data and mining results against a knowledgeable adversary with certain background knowledge. To achieve k-privacy, the data owner need sends the encrypted database of both the real and fictitious transactions to the cloud. To cancel out fictitious transactions, the data owner in [21] is required to count itemset occurrences in fictitious transactions. However, none of these works assume that the adversary has the capability of launching chosen plaintext attacks; i.e., they do not provide semantic security.

Based on predication encryption, Lai et al. [22] proposed the first semantically secure solution. It is resilient to chosen-plaintext attacks on encrypted items, but it is vulnerable to frequency analysis attacks.

2) Secure ARM over vertically distributed data

A great many cryptography-based approaches have been developed in the context of PPARM algorithms, to solve Secure Multiparty Computation. In secure multiparty mining over distributed datasets, vertical data distribution is one of the distributed data scenarios.

For the first time, Vaidya et al. [23] address privacy issues in vertically partitioned databases. In this paper, secure scalar product protocol was presented to build a privacy-preserving frequent itemset mining solution. Association rules can then be found given frequent itemset and their supports. Since the publication of this seminal work, many PPARM or frequent itemset mining solutions have been published.

Similar to [23], Kharat et al. [24] used the scalar product to enable association rules mining. Asymmetric homomorphic encryption [25] was used to compute the supports of itemsets. However, other solutions use a set intersection cardinality protocol or a secret sharing scheme [26] to perform these computations.

All existing solutions do not utilize a third-party server to compute the mining result except for [27-28]. In the scheme in [27], the data owner (a.k.a. the master) is responsible for the mining and the other data owners (a.k.a. slaves) insert fictitious transactions to their respective datasets, and send the datasets to the master. Each data owner will also send his set of real transactions' IDs to a semi-trusted third-party server. However, in this solution, the master does the majority of the computational. Though fictitious data are added in datasets to lower data usability, the master is able to learn significant information about other data owners' raw data from the received datasets. Based on symmetric homomorphic encryption, Li et al. [28] proposed a cloud-aided frequent itemset mining solution.

3) Secure ARM over horizontally distributed data

Kantarcioglu et al [29] firstly address privacy issues in horizontally distributed data. Also Schuster et al. [30] considered this problem in the horizontal setting, but he considered large-scale systems in which, on top of the parties that hold the data records (resources) there are also managers which are computers that assist the resources to decrypt messages; Moreover, [30] assumes that no collusions occur between the different resources or managers. Tassa et al. [31] improved the protocol in [29] in terms of efficiency and privacy. Zhang et al. [32] merged the secure multiparty computation and differential privacy to preserve the privacy of the statistical operations. However, it is not clear how this approach can be applied to handle ARM given that division operations must be performed between the parties in a secure way in order to validate the minimum support and confidence. Wahab et al. [33] found the global strong association rules confidentially, and satisfied ε-differential privacy.

The common problem with these schemes is that they assume that communication channel for data exchange is secure. In practice, this assumption is strong. Chirag N et al. [34] proposed a privacy preserving association rule mining in horizontally partitioned databases without trusted third party, even communication channel is unsecured between involving sites.

IV. CONCLUSION AND FUTURE DIRECTION

In this paper, we present a classification of PPARM algorithm and survey major algorithms in each class. After a review of many existing PPARM techniques, we obtain some conclusions and come with up future direction.

●A single technique does not exceed all the parameters such as performance, data utility etc., rather it can perform better than other algorithms on certain parameters. So it is necessary to develop the combined algorithms, in order to achieve comprehensive privacy protection requirements.

●In most existing literature, rule interestingness measures in ARM algorithms are support and confidence. On the basis of the nature of application, different measures can be used to measure the interestingness of quantitative rules.

●Since each user may have different concern and requirement over privacy, therefore, user-oriented and more refined privacy preserving techniques can be developed.

●Parallel algorithms could be developed to improve the performance of the algorithm for large datasets.

ACKNOWLEDGMENT

The work was sponsored by National Natural Science Fo undation of China Grant No.61772174, Plan For Scientific I nnovation Talent of Henan Province Grant No.1742005100 11, Program for Innovative Research Team (in Science and Technology) in University of Henan Province Grant No.15I RTSTHN010, Program for Henan Province Science and Tec hnology Grant No.142102210425, Natural Science Foundati on of Henan Province Grant No.162300410094, Project of t

he Cultivation Fund of Science and Technology Achievements of Henan University of Science and Technology Grant No.2015BZCG01.

REFERENCES

[1] S Sharma, "A Study on Data Mining Horizons," International Journal of Recent Trends in Engineering & Research (IJRTER) Vol. 02, Apr. 2016, pp.322-326.

[2] Geeta S. Navale and Suresh N. Mali, "Survey on Privacy Preserving Association Rule Data Mining," International Journal of Rough Sets and Data Analysis, Vol.4, Apr.2017, pp.63-80.

[3] Ibrahim S. Alwatban andAhmed Z. Emam, "Comprehensive Survey on Privacy Preserving Association Rule Mining:Models, Approaches, Techniques and Algorithms," International Journal on Artificial Intelligence Tools, vol. 23, May. 2014.

[4] V. S. Verykios, E. Bertino, I. Nai Fovino, L. Parasiliti Y. Saygin and Y. Theodoridis, "State-of-the-art in Privacy Preserving Data Mining." ACM Sigmod Record, ACM Press , Vol.33, 2004, pp. 50-57.

[5]S. J. Rizvi and J. R. Haritsa, Maintaining data privacy in association rule mining, Proc. 28thInt. Conf. Very Large Data Bases (VLDB Endowment, Hong Kong, China, 2002), pp. 682–693.

[6]L. Sweeney. k-anonymity: a model for protecting privacy. International Journal on Uncertainty Fuzziness and Knowledge-based Systems, 10(5), 2002.

[7] V.S. Verykios and A. Gkoulalas-Divanis, "A survey of association rule hiding methods for privacy," in: Privacy-Preserving Data Mining, 2008, pp. 267–289.

[8] Mike J. Atallah, Elisa Bertino, Ahmed K. Elmagarmid, Mohamed Ibrahim, and Vassilios S. Verykios, "Disclosure Limitation of Sensitive Rules," In Proceedings of the IEEE Knolwedge and Data Engineering Workshop, 1999, pp. 45–52.

[9] Elena Dasseni, Vassilios S. Verykios, Ahmed K. Elmagarmid, and Elisa Bertino, "Hiding Association Rules by using Confidence and Support," In Proceedings of the 4th Information Hiding Workshop , 2001, pp. 369–383.

[10] V. S. Verykios, A. K. Elmagarmid, E. Bertino, Y. Saygin and E. Dasseni, Association rule hiding, *IEEE Trans. Knowl. Data Eng. Vol.*16 ,2004, pp. 434–447.

[11] A. Amiri, "Dare to share: Protecting sensitive knowledge with data sanitization," Decis. Support Syst. Vol.43, 2007, pp.181–191.

[12] S.-L. Wang, B. Parikh and A. Jafari, "Hiding informative association rule sets, " Expert Syst.Appl. Vol.33, 2007, pp.316–323.

[13]SL. Wang, R. Maskey, A. Jafari and T.-P. Hong, "Efficient sanitization of informative association rules, Expert Systems with Applications An International Journal, Vol.35, 2008, pp. 442-450.

[14] B. Yıldız and B. Ergen?, Integrated Approach for Privacy Preserving Itemset Mining Intelligent Control and Innovative Computing, eds. S. I. Ao, O. Castillo and X. Huang(Springer, New York, 2012), pp. 247–260.

[15]Yucel Saygin, Vassilios S. Verykios, and Ahmed K. Elmagarmid, Privacy preserving association rule mining, In Proceedings of the 12th International Workshop on Research Issues in Data Engineering, 2002, pp.151–158.

[16] S. L. Wang and A. Jafari, "Using unknowns for hiding sensitive predictive association rules," IEEE Int. Conf. Information Reuse and Integration, IRI-2005, 2005, pp. 223–228.

[17] X. Chen, M. Orlowska and X. Li, "A new framework of privacy preserving data sharing," Proc, IEEE 4th Int. Conf. Data Min.ICDM04 Workshop,2004, pp. 47–56.

[18] Yuhong Guo, Reconstruction-Based Association Rule Hiding, in Proceedings of SIGMOD2007 Ph.D. Workshop on IDAR2007, June 10, 2007.

[19] W. K. Wong, D. W. Cheung, E. Hung, B. Kao, and N. Mamoulis. Security in outsourcing of association rule mining. In Proceedings of he 33rd International Conferenceon Very Large Data Bases (VLDB), 2007, pp.111–122.

[20] I. Molloy, N. Li, and T. Li, "On the (in) security and (im) practicality of outsourcing precise association rule mining," in ICDM 2009, 2009, pp. 872-877

[21] F. Giannotti, L. Lakshmanan, A. Monreale, D. Pedreschi, and H. Wang, "Privacy-preserving mining of association rules from outsourced transaction databases," IEEE Systems Journal, vol. 7, no. 3, pp. 385–395,2013.

[22] 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," Information Sciences, vol. 267, pp. 267–286, 2014.

[23] Jaideep Vaidya and Chris Clifton, "Privacy preserving association rule mining in vertically partitioned data," In the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining ,2002, pp639–644.

[24] R. Kharat, M. Kumbhar, and P. Bhamre, "Efficient privacy preserving distributed association rule mining protocol based on random number," in Intelligent Computing, Networking, and Informatics. Springer, 2014, pp. 827–836.

[25] S. Zhong, "Privacy-preserving algorithms for distributed mining of frequent itemsets," Information Sciences, vol. 177, no. 2, 2007, pp. 490–503.

[26] X. Ge, L. Yan, J. Zhu, and W. Shi, "Privacy-preserving distributed association rule mining based on the secret sharing technique," in SEDM 2010, 2010, pp.345-350.

[27] B. Rozenberg and E. Gudes, "Association rules mining in vertically partitioned databases," Data & Knowledge Engineering, vol. 59, no. 2, 2006, pp. 378–396.

[28] Lichun Li, Rongxing Lu, Kim-Kwang Raymond Choo, Anwitaman Datta, and Jun Shao, Privacy-Preserving Outsourced Association Rule Mining on Vertically Partitioned Databases Transactions on Information Forensics and Security , Vol.11, 2016 , pp1847-1861.

[29]M. Kantarcioglu and C. Clifton, "Privacy-preserving distributed mining of association rules on horizontally partitioned data," IEEE Transactions on Knowledge and Data Engineering, Vol.16, 2004, pp.1026–1037.

[30]A. Schuster, R. Wolff, and B. Gilburd, "Privacy-preserving association rule mining in large-scale distributed systems," In CCGRID, 2004, pp. 411–418.

[31] Tamir Tassa, "Secure Mining of Association Rules in Horizontally Distributed Databases, transactions on Knowledge and Data Engineer", Vol.26, 2014, pp970-983.

[32]N Zhang, M li, W Lou, Distributed Data Mining with Differential Privacy, IEEE International Conference on Communication, 2011 , 57 (4) , pp.1-5.

[33] Wahab, O.A., Hachami, M.O., Zaffari, A., Vivas, M., Dagher, G.G.: "DARM: a privacy-preserving approach for distributed association rules mining on horizontally-partitioned data," In: 18th International Database Engineering and Applications Symposium, 2014, pp. 1–8.

[34] Chirag N. Modi and Ashwini R. Patil, "Privacy Preserving Association Rule Mining in Horizontally Partitioned Databases Without Involving Trusted Third Party (TTP)," Proceedings of 3rd International Conferenceon Advanced Computing, Networking and Informatics, 2015, pp.549-555.