I J C T A, 9(34) 2016, pp. 543-549

A Study on Privacy Preserving Association Rule Mining

J. Sumithra Devi[1] and M. Ramakrishnan[2]

ABSTRACT

In recent years, data creation and storing have seen an enormous growth. Huge volumes of data are created by many business applications and it becomes vital to analyze these data to extract knowledge. Data mining techniques use various tools and provide business knowledge. Among various methods used by data mining, association rule mining is important one and it tries to find correlations among various items of a transaction. During association rule mining process, it is important to maintain the security of the data since organization's sensitive information should not be leaked. This paper provides a review on different privacy preservation association rule mining techniques. Various kinds of privacy preserving methods such as heuristic based, exact based, border based, reconstruction based and cryptography based approaches are discussed. It also presents various performance evaluation metrics used to rank the methods and comparative study of various privacy preserving association rule mining methods in a coherent way.

Keywords: Association Rule Mining, ARM, Data Perturbation, Association Rule Hiding, Security, Data Blocking

1. INTRODUCTION

Mining important knowledge as patterns from huge volumes of data is the main objective of data mining. It discovers interesting patterns that help business people to take vital and strategic decisions. Data mining tools analyses large volumes of data under different perspectives and predicts future trends and behaviour that enables business community to take optimal, proactive and knowledge driven decisions [1]. It is an automated process that works beyond the analysis of past events provided by traditional decision support systems.

Data mining technique is widely used in many fields such as Market Basket Analysis, Future Healthcare, Education, Manufacturing Engineering, Fraud Detection, Intrusion Detection, Financial Sector, Banking, Corporate Surveillance, Research Analysis, Bio-Informatics, etc. Since data mining extracts different kinds of knowledge, different kinds of methods and techniques are used. The dominant methods and techniques of data mining include summarization, association, classification, clustering, prediction, etc[2]. Each method is having its own significance in data mining process and either single or combinations of methods are used during knowledge discovery process.

Among several methods, association is an important one that finds association among different items of a transaction. Popularly known as Association Rule Mining (ARM), it primarily focuses on finding frequent itemsets and strong association rules [3]. The objective is to find associations that occur together more often from random sampling of all possibilities.

The raw data onto which association rule mining process is applied may contain vital and sensitive information. For example, the database may contain patient's personal information that is confidential.

[1] Research Scholar, Department of Computer Science, Bharathiyar University, Coimbatore, Tamil Nadu, India, *E-mail: sumithraphd2015@gmail.com*

[2] Chairperson, School of Information Technology, Madurai Kamaraj University, Madurai, Tamil Nadu, India, *E-mail: ramkrishod@gmail.com*

Revealing private information is threatening to that enterprise since it is the prime responsibility of any enterprise to safeguard its customer's data. Sharing the information is useful but is should be done in such a way that data confidentiality should not be compromised[4]. To overcome this difficulty, private data must be hidden before sharing to keep them safe from any unauthorized access.

Hence this paper presents a literature review on maintain privacy while performing association rule mining process. The paper is organized as follows. Paper starts by giving brief introduction about the data mining process and importance of association rule mining. Need for security preservation in association rule mining is also discussed in section1. Section 2 provides brief related experiments and findings. Association rule mining problem is lucidly defined in section 3. Section 4 provides association rule hiding process thereby providing goals of database sanitization, various rule hiding methods and side effects. Section 5 presents performance evaluation parameters and paper ends by briefing the important points as conclusion in section 6.

2. RELATED WORK

Xun Yi *et al.* [5] presented a paper on preserving privacy in association rule mining under cloud computing. To protect private data from unauthorized view and access, k-anonymity, k-support and k-privacy techniques have been used. By using these three techniques, association rule mining can be performed globally without revealing sensitive information. To perform association rule mining, data owner outsources semi honest servers that cooperate to perform rule mining on encrypted data.

Sayin *et al.* [6] presented a method privacy preserving association rule mining. The proposed method secures both data and the extracted knowledge in virtual business environments. This data centre is open-ended and vulnerable to all kinds of attacks. Frame work uses deterministic algorithm for privacy preservation during association rule mining process.

El-Sisi and Ashraf [7] presented a privacy preserving association rule mining by utilizing fast cryptographic technique. This is applied on distributed homogeneous databases and it is a modification of traditional privacy preserving data mining algorithms. It uses semi-honest model with negligible collision probability. The proposed method is fast, accurate and increase in number of client does not affect the scalability of the algorithm.

Dhyanendra Jain *et al.* [8] proposed a method to hide sensitive association rules without altering the support of sensitive items. It uses a different method of modifying the database transactions so that confidence of the sensitive rules can be reduced without changing the support of a sensitive item. This method is more efficient and needs minimum number of scans to generate association rules.

Poitr Andruszkiewicz [9] presented a work on optimization of MASK scheme in privacy preserving association rule mining. MASK stands for Mining Associations with Secrecy Konstraints, and it is effectively high in time, cost. Modified MASK, called MMASK is proposed which is a new optimization algorithm that breaks the exponential complexity and achieves better results.

3. ASSOCIATION RULE MINING

Association rule mining, abbreviated as ARM, try to locate interrelated transactions of a database, which reveals the implicit relationships among data attributes. Association rule uncover relationships between unrelated data in the dataset. An association rule contains two parts; an antecedent part (if) and a consequent part (then). A consequent is an item that is derived from the combination of antecedents.

The problem of association rule mining can be represented as follows: Let $I=\{I_1,I_2,I_3, \ldots.I_n\}$ be the set of items. Let $T=\{t_1, t_2, t_3, \ldots.t_m\}$ be the set of transactions and each transaction contains a set of items such that $T \hat{I} I$. Each transaction is identified by an identifier called tid. Let A be the set of items of transaction T and

T is said to contain A if and only if A⊆T. An association rule is represented of the form A ⇒B, where A∈I, B∈I and A∩B = f. A and B is said to be itemset.

The rule A ⇒ B contains support s where 's' is the percentage of transactions in dataset D. The rule A ⇒ B is also having confidence 'c' which represents the percentage of transactions in D containing A as well as B. hence support and confidence is calculated as follows

$$\text{Support } (A \rightarrow B) = P(A \cup B)$$

$$= \frac{Total\ Number\ of\ (A \mid B)}{Total\ number\ of\ transactions\ in\ D}$$

$$Confidence\ (A \rightarrow B) = P(A \cup B)$$

$$= \frac{Support\ (A \cup B)}{Support\ (A)}$$

Not all the rules are considered as important ones. The rules which hold minimum support and confidence levels fixed by the DB miner are alone selected and remaining rules are not considered. These minimum threshold limits vary time to time and based on user requirements.

Association rule mining process works in two phases. Finding all the frequent itemsets, itemsets that occur atleast as frequent as defined by minimum support, in the first phase. Strong association rules are generated in the second phase. These rules must satisfy minimum support and minimum confidence.

4. ASSOCIATION RULE HIDING

This is the common technique used in Privacy Preserving Data Mining (PPDM) which sanitizes database for hiding sensitive information. Sanitizing the database should result in achieving atleast one of the following goals.

1. No rule is considered as sensitive from owner's perspective and can be mined from original database at minimum support and confidence. The same information is extracted with same or higher thresholds, if the database is sanitized.
2. All the non-sensitive rules that appear when mining the original dataset with predefined support and confidence thresholds can be successfully mined from the sanitized database at same thresholds or higher.
3. Except non-sensitive% rules, if a new rule is not mined from original dataset, it cannot be mined from sanitized database either at same thresholds or higher.

Most of the association rule hiding algorithms depends on adjustments in two important parameters, support and confidence[10]. Either you can increase or decrease the support keeping confidence unchanged or vice versa. But care must be taken as modification of support and confidence values for hiding association rules may result in few kinds of side effects[10]. The general side effects are lost rules, false rules and ghost rules. Rules that can be generated from the original dataset and not in sanitized dataset are lost rules[11]. Rules that cannot be hidden by hiding algorithm are called false rules. Ghost rules are the one which are not derived from original dataset but derived from sanitized datasets. An accurate hiding can cause least change in original dataset[11]. The rules hiding strategies are divided into five categories and they are heuristic based, border based, exact based, reconstruction based and cryptography based.

4.1. Heuristic Based Approaches

This type of approach selectively sanitizes a set of transactions from the original database to hide sensitive association rules. Heuristic based approaches are further classified as data distortion techniques and data blocking techniques[12].

Data distortion technique replaces 1-value to 0-value for reducing the confidence of rules. This technique also adds items by replacing 0-value to 1-value for reducing the support of rules[13]. To hide sensitive rules, deletion or addition of items are performed which results modification in database. To perform this, many algorithms are available. One algorithm inserts the items in transaction thereby increasing the value of support. This automatically decreases the confidence value. This method is having side effects that false are generated because of insertion of new items. In contrast to this, few algorithms delete items of rules so confidence gets decreased[13].

Data blocking techniques use more confidence and did not reduce the sensitivity of rule. To increase or to decrease the support value of the items, 0's and 1's are replaced by unknown "?". This creates difficulty for a rival to predict the exact value of "?". This method needs less number of database scans to replace values with "?". This set up also informs rival party that the database has been sanitized[14].

4.2. Border Based Approach

It modifies the borders in lattice of frequent and infrequent itemsets of the original database thereby hiding sensitive association rule. Group of itemsets that represents the frequent and infrequent forms the border and these itemsets are modified. Minimum affected itemsets are selected for modification else it leads to side effects. Border based approach is still dependent on heuristic approach while deciding item modification.

4.3. Reconstruction based Approaches

This method alters the data and reconstructs the distribution at an aggregate level. By using knowledge base, it conceals the sensitive rules by sanitizing itemset lattice. Publishing frequent itemset will not cause threat to privacy because reengineering on frequent itemset is very difficult. Mielikainen [15] analyzed the computational complexity of inverse frequent set mining and showed that the problem is computationally difficult. It is also described that problem is NP-complete. Reconstruction based approaches are efficient, secure on number of databases and it is performed on frequent itemset which is a halfway process to association rule mining [15].

A variant of this method based on frequent itemset mining was proposed by Guo [16] which uses FP tree. It works in three phases viz. generating frequent itemsets in the first phase, performing sanitization over frequent itemsets in the second phase. Here sensitive frequent itemsets are selected according to sensitive association rules [16]. By using inverse frequent itemset, sanitized database is generated in the third phase.

4.4. Exact Approaches

This approach is capable of providing better solutions compared to heuristic based approaches and the computational cost of this method is very high. This method views database sanitization process as Constraint Satisfaction Problem (CSP) and it provides solution using linear programming solver[17]. Database sanitization is an atomic process and it avoids local minima.

Exact approach tries to minimize the distance between original database and sanitized database. Menon *et al.* [18] presented a method consisting of exact and heuristic approaches. First part of the method formulates CSF with the objective of identifying minimum number of transactions needed to be sanitized. The optimization work is done by heuristic approach.

4.5. Cryptography based Approaches

This method is based on exchanging keys so that security can be maintained. Instead of distorting or altering the data, data is encrypted before sharing. Generally it is used in multi-party computation over distributed data[19]. This method is further divided into vertically partitioned distributed data and horizontally partitioned distributed data.

In vertically partitioned distributed data, support count of every sub-itemset distributed globally is calculated. An itemset is said to be global if its support is above the prescribed threshold. Vertically partitioned distributed data mining has been extended with applications such as decision trees, SVM classification, Naïve Bayes Classification and k-means clustering[20].

Horizontally partitioned distributed data method works by finding global frequent itemsets without leakage of inter-site information. It calculates secure sum of inter-sites and from this overall itemset support degree is calculated[21].

4.6. Comparison of various Rule Hiding Methods

Table 1
Comparison of various association rule hiding methods

Methods	*Advantages*	*Disadvantages*
Heuristic based Approaches	▪ Efficient, scalable and responses are quick. ▪ Maintains data integrity and provides best solution.	▪ Often creates false and ghost rules. ▪ Reproducing the original dataset from sanitized one is very difficult.
Border based Approaches	▪ Side effects are very minimal. ▪ Selects frequent itemsets using greedy method	▪ Still dependent on heuristic based approach to take decision on item modification
Exact Approaches	▪ Provides optimal solution without any side effect	▪ Computational complexity is very high due to linear programming
Reconstruction based Approaches	▪ Lesser side effect. ▪ Creates privacy aware database using sensitive characteristic from original dataset.	▪ Number of transactions is restricted in new database.
Cryptography based Approaches	▪ Use of cryptography provides additional security to database. ▪ Best for partitioned data	▪ Fails to protect output of ARM.

5. PERFORMANCE EVALUATION PARAMETERS

There are few metrics that are used to calculate the performance of association rule hiding algorithm. In this section, we will discuss few important ones.

Hiding Failure

It represents the number of rules that are not covered in sanitized dataset. It measures sensitive association rules that appears in sanitized dataset. It is calculated using the below formula

$$HF = \frac{|R_s(D')|}{|R_s(D)|}$$

Where $|R_s(D')|$ and $|R_s(D)|$ are the number of sensitive rules appearing in the sanitized dataset and original dataset respectively. The lower the value of HF, higher the accuracy.

Lost Rules Cost

It represents the percentage of non-sensitive patterns which are accidentally hidden in sanitized database. Other name is Misses Cost. It is calculated as follows

$$MC = \frac{|R_{NS}(D)| - |R_{NS}(D')|}{|R_{NS}(D)|}$$

where $|R_{NS}(D)|$ and $|R_{NS}(D')|$ are the number of non-sensitive association rules present in the original database and in sanitized database respectively.

Side Effect Factor

It represents the amount of non-sensitive association rules that are removed during sanitization of database. It is calculated as follows

$$SEF = \frac{|P| - |P'| + |S_R(D)|}{|P| + |S_R(D)|}$$

Where | P | and | P' | are the number of discovered association rules in the original dataset and sanitized dataset respectively. $S_R(D)$ is the size of all non-sensitive rules in original dataset D.

Artifactual Patterns

It gives the number of ghost rules created in the sanitized database. It is calculated as follows

$$AF = \frac{|P'| - |P \cap P'|}{|P'|}$$

Where | P | and | P' | are the set of association rules discovered in the original dataset and sanitized dataset respectively.

Apart from the above, scalability, data quality, privacy level, dissimilarity are some of the other performance metrics used to evaluate the association rule hiding process.

6. CONCLUSION

Maintaining privacy in association rule mining is an important concept in data mining technique. It provides security thereby protecting sensitive rules from unauthorized entities. A survey on various security in association rule mining process is presented in this paper. The paper first presents the basics of association rule mining process, followed by introduction to database sanitization process. The goals of database sanitization process and various methods of association rule hiding methods viz. heuristic based, exact based, border based, reconstruction based, cryptography based methods are discussed elaborately. The paper also presents comparative study of various methods.

REFERENCES

[1] Pal, Jiban K. "Usefulness and applications of data mining in extracting information from different perspectives." Annals of Library and Information Studies 58.1 (2011): 7-16.

[2] Castro, Félix, *et al*. "Applying data mining techniques to e-learning problems." Evolution of teaching and learning paradigms in intelligent environment. Springer Berlin Heidelberg, 2007. 183-221.

[3] Feldman, Ronen, and James Sanger. The text mining handbook: advanced approaches in analyzing unstructured data. Cambridge University Press, 2007.

[4] Pearson, Siani. “Taking account of privacy when designing cloud computing services.” Proceedings of the 2009 ICSE Workshop on Software Engineering Challenges of Cloud Computing. IEEE Computer Society, 2009.

[5] Yung-Feng Lu , Chin-Fu Kuo , Shih-Chun Chou , Rong-Sheng Wang, A secure and efficient multicast protocol for enterprise collaboration systems, Proceedings of the 2015 Conference on research in adaptive and convergent systems, October 09-12, 2015, Prague, Czech Republic.

[6] Saygin, Yucel, Vassilios S. Verykios, and Ahmed K. Elmagarmid. “Privacy preserving association rule mining.” Research Issues in Data Engineering: Engineering E-Commerce/E-Business Systems, 2002. RIDE-2EC 2002. Proceedings. Twelfth International Workshop on. IEEE, 2002.

[7] El-Sisi, Ashraf. “Fast Cryptographic Privacy Preserving Association Rules Mining on Distributed Homogenous Database.” Int. Arab J. Inf. Technol. 7.2 (2010): 152-160.

[8] Jain, Dhyanendra, et al. “Hiding sensitive association rules without altering the support of sensitive item (s).” Advances in Computer Science and Information Technology. Networks and Communications. Springer Berlin Heidelberg, 2012. 500-509.

[9] Andruszkiewicz, Piotr. “Optimization for mask scheme in privacy preserving data mining for association rules.” Rough Sets and Intelligent Systems Paradigms. Springer Berlin Heidelberg, 2007. 465-474.

[10] Alatas, Bilal, Erhan Akin, and Ali Karci. “MODENAR: Multi-objective differential evolution algorithm for mining numeric association rules.” Applied Soft Computing 8.1 (2008): 646-656.

[11] Guo, Yuhong. “Reconstruction-based association rule hiding.” Proceedings of SIGMOD2007 Ph. D. Workshop on Innovative Database Research. Vol. 2007. 2007.

[12] Bertino, Elisa, Igor Nai Fovino, and Loredana Parasiliti Provenza. “A framework for evaluating privacy preserving data mining algorithms*.” Data Mining and Knowledge Discovery 11.2 (2005): 121-154.

[13] Luo, Yongcheng, Yan Zhao, and Jiajin Le. “A survey on the privacy preserving algorithm of association rule mining.” Electronic Commerce and Security, 2009. ISECS’09. Second International Symposium on. Vol. 1. IEEE, 2009.

[14] Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. Unsupervised learning. Springer New York, 2009.

[15] Mielikäinen, Taneli. “On inverse frequent set mining.” Workshop on Privacy Preserving Data Mining. 2003.

[16] Guo, Yuhong, et al. “A FP-tree-based method for inverse frequent set mining.” Flexible and Efficient Information Handling. Springer Berlin Heidelberg, 2006. 152-163.

[17] Verykios, Vassilios S., and Aris Gkoulalas-Divanis. “A survey of association rule hiding methods for privacy.” Privacy-Preserving Data Mining. Springer US, 2008. 267-289.

[18] Paulino, Glaucio H., Govind Menon, and Subrata Mukherjee. “Error estimation using hypersingular integrals in boundary element methods for linear elasticity.” Engineering Analysis with Boundary Elements 25.7 (2001): 523-534.

[19] Ahmadi, Hossein, et al. “Privacy-aware regression modeling of participatory sensing data.” Proceedings of the 8th ACM Conference on Embedded Networked Sensor Systems. ACM, 2010.

[20] Steinberg, Dan, and Phillip Colla. “CART: classification and regression trees.” The top ten algorithms in data mining 9 (2009): 179.

[21] Sathiyapriya, K., and G. Sudha Sadasivam. “A survey on privacy preserving association rule mining.” International Journal of Data Mining & Knowledge Management Process 3.2 (2013): 119.