A Comprehensive Study of Privacy Preservation Techniques in a Distributed Association Rule Mining

Nusrat Jabeen. T
Research Scholar, Department of Computer Science, Bharathiyar University, Coimbatore, Tamil Nadu, India

M. Chidambaram
Assistant Professor, Computer Science Department, Rajah Serfoji Government College, Thanjavur Tamil Nadu, India

ABSTRACT

Association rule mining is a popular technique in data mining process which tries to find interesting associations and correlations among various data items in a transaction. Many business organizations share data with others, outsource their business data for specific business solution. In these situations, the sensitive information leakage is the biggest problem. Strong and efficient privacy preserving techniques are needed to secure organization's data during third party sharing. In this paper, a comprehensive study of various methods for privacy preservation is presented which evaluates and analyzes various techniques for maintaining privacy during association rule mining process. Moreover, paper also prospects the development of privacy preservation for future applications.

General Terms

Association rule mining, ARM, Privacy Preservation, Data Sanitization, Data Distortion, Data Perturbation, Cryptography, Frequent Patterns

Keywords

Association rule mining, ARM, Privacy Preservation, Data Sanitization, Data Distortion, Data Perturbation, Cryptography, Frequent Patterns

1. INTRODUCTION

Association rule mining is the process of finding associations or casual structures, interesting correlations and frequent patterns that exists among the set of items in a transaction database. It is the process of finding association rules that satisfy the predefined minimum support and confidence[1].

Association rule mining is widely used in areas such as telecommunication networks, market-basket analysis, inventory control, risk management, etc. Other domains where association rule mining plays a vital role are finding patterns in biological databases, analyzing library circulation data, protein composition study, study of population and economic census, etc. Hence association rule mining is one of the core concepts in knowledge discovery process[2].

Performing association rule mining is a complex task and the whole process is decomposed into two sub tasks[3]. The first task in association rule mining is to find frequent item sets-item sets whose occurrences exceeds a predefined limit in the transaction database. The second task is to extract association rules from large item sets that satisfy the minimum confidence. The first task can be divided into two sub tasks : generation of candidate large item sets and generation of frequent item sets[3].

Let I = { i1, i2, i3, ……. in} be a set of n items. Let D be the transaction database with 'm' transactions D = {t1, t2, t3, …..tm}. The rule is described in the form $X \Rightarrow Y$, where X and Y are disjoint patterns. Other names for X and Y are antecedent and consequent respectively. Association rules are the rules that satisfies user defined minimum support and confidence. Support is defined as the percentage of transactions in D that contain $X \cup Y$ and confidence is the percentage of transactions in D that contains X that also contain Y. Therefore an association rule $X \Rightarrow Y$ must satisfy

$$\text{Support } (X \cup Y) > \sigma \quad \text{and}$$

$$\text{Confidence}(X \Rightarrow Y) > \delta$$

where σ and δ are user defined minimum support and confidence respectively.

If the data to be mined is distributed across multiple sources, the biggest problem is how data can be mined without either party disclosing its data to others? The simplest solution is to perform association rule mining at each source independently and results are combined[4]. This strategy has failed to give valid results. The problem of performing data mining with multiple parties, with each party wanting to hide its data from others, is said to be Secure Multiparty Computation (SMC) problem[5].

Providing security in association rule mining process is decided by practical application of different privacy protection requirements. The traditional methods include data distortion, data encryption and data released[6]. Many algorithms for privacy preservation were developed mainly based on encryption method. Few algorithms were described on data streams[6]. Hence this paper reviews several algorithms relating to privacy preservation during association rule mining process.

The rest of the paper is organized as follows: Section 1 describes the introduction about association rule mining and need for privacy while performing ARM in a distributed fashion. Section 2 and 3 briefs goals of ARM and performance metrics used for evaluating ARM methods respectively. Section 4 elaborately analyzes various privacy protection techniques in ARM such as data distortion, data perturbation, block-based ARM, reconstruction-based ARM, cryptography-based ARM and FP-tree based ARM. Conclusion is presented in section 5 and paper ends by listing the references in section 6.

2. GOALS OF PRIVACY PRESERVING ARM

- No sensitive association rules mined from original database or mined from sanitized database with predefined support and confidence should be made public. The sensitive rules may be of specific form or from owner's perspective.

- All the non-sensitive rules that can be mined from original database with predefined support and confidence, should also be mined from sanitized database at the same support and confidence level.
- No rule that was not derived from original database with predefined support and confidence, should not be derived from sanitized database at the same support and confidence level.

3. PERFORMANCE METRICS

Performance of any privacy preserving association rule mining is estimated using the following metrics:

3.1 Hiding Failure (HF)

It is the measure of restrictive association rules that appear in the sanitized database. It is the percentage of data that remain exposed in the sanitized dataset. It is calculated by using the below formula:

$$H_F = \frac{|S_R(D^{'})|}{|S_R(D)|}$$

Where D is the original data set, D′ is the sanitized data set, SR is the number of sensitive association rules.

3.2 Misses Cost

It is the measure of amount of legitimate association rules that are hidden by accident after sanitization. It is the percentage of non-sensitive data hidden during sanitization process. It is calculated as follows:

$$M_C = \frac{|S_R(D^{'})| - |S_R^{'}(D^{'})|}{|S_R^{'}(D)|}$$

where $|S_R^{'}(D)|$ is the size of set of all non-sensitive rules.

3.3 Artifactual Patterns

It is the measure of artificial association rules created by adding the noise in the data. It is the measure of discovered artifacts. It is calculated by:

$$A_F = \frac{|P^{'}| - |P \cap P^{'}|}{P^{'}}$$

where P is the set of discovered association rules in the original database D and P′ is the set of association rules in the sanitized database D′

3.4 Difference

It is the measure of difference between original database and sanitized database. It is calculated by:

$$Diff\,(D, D^{'}) = \frac{1}{\sum_{i=1}^{n} f_D(i)} \times \sum_{i=1}^{n} \left[f_D(i) - f_{D^{'}}(i) \right]$$

where $f_D(i)$ represents frequency of ith item in the original database, $f_{D^{'}}(i)$ is the frequency of ith item in the sanitized database and 'n' is the number of distinct items in the original database.

3.5 Side-Effect Factor (SEP)

It is the amount of non-sensitive association rules that are removed during sanitization process. It is calculated by:

$$S_{EF} = \frac{|P| - \left(|P^{'}| + |R_P(D)| \right)}{|P| - |R_P|}$$

4. PRIVACY PROTECTION TECHNIQUES

4.1 Data Distortion

The basic idea behind this technique is to modify original database records before it is passed to other party so that privacy can be maintained. While hiding the sensitive data in the original database, care must be taken that data between hidden and original must have the same characteristics[7]. Data distortion methods include blocking, aggregation, swapping, perturbation and sampling. To handle complex problems, random perturbation, blocking and condensation methods relating to data distortion technique are used[7].

4.2 Perturbation

By using support and confidence, statistical significance is estimated. This is because, few among the generated association rules are sensitive and others are not[8]. Association rule hiding technique uses this method to purify the original data set. All the sensitive rules are kept in original database. The non-sensitive rules can be dug out from original dataset with the same support and confidence of sensitive rules. Perturbation can be done by altering an attribute value by a new value or by adding noise to the actual data[8].

4.3 Block-based Association Rule Mining

Blocking method works by reducing the degree of support and confidence of sensitive association rules and replacing some attribute values of data items with question mark or true value. This method is very popular in medical association rule mining[9]. The minimum support and confidence is modified into minimum support interval and minimum confidence interval. The confidentiality of the data is maintained till the values of support and confidence of a sensitive association rule lies within this interval. One problem with block-based privacy preserving association rule mining is that it is difficult to determine the support and confidence of sensitive association rule since original data is replaced with some data of unknown value[9]. This can be solved by using uncertain symbols which then can be replaced with actual support and confidence[10].

4.4 Reconstruction based Association Rule Mining

These types of privacy preservation schemes first perform perturbation of data and then reconstructing the distributions. There are varieties of algorithms for reconstructing the distributions and data types[11]. For distributed data, Bayesian reconstruction process is used which is based on EM algorithm. EM algorithm is robust and it can estimate the original distribution when large amount of data is obtained. Another way of data reconstruction is to keep original data aside and start from sanitizing knowledge base. The new data is reconstructed from sanitized knowledge base[12].

4.5 Cryptography based Association Rule Mining

Cryptographic protocols can be used to share the information with different parties during association rule mining process. The association rule mining techniques that uses cryptography for privacy preservation can be classified as:

> **Vertically partitioned:** Vertically partitioned data set contains different attributes for each item in different sites. Mining association rules from

vertically partitioned data is done by finding support count of an item set. Support count for every sub-item sets is calculated across different sites. An item set is determined as global frequent item set if its support count is greater than the user defined support count[13].

Horizontally partitioned data: Here the transactions are distributed across multiple sites in horizontally partitioned database. The idea is to find frequent item sets without leakage of inter-site information. The total support count of an item set is calculated which is the sum of all local support counts. The frequent item sets with support count greater than the user specified value are declared as global frequent item sets[14].

4.6 FP- tree based Association Rule Mining

It is one type of reconstruction technique used for performing inverse frequent set mining[15]. Frequent Pattern (FP) is a divide and conquer type methodology that decomposes association rule mining tasks into smaller ones, probably three sub tasks[17]. In the first phase, all the frequent item sets with their supports and support counts are generated from original database. Sanitation algorithm is executed on generated frequent item sets and sanitized frequent item sets are generated. By using inverse frequent set mining algorithm, new database is generated from sanitized frequent item sets. This method is secure and collision-resistant, and even if n-1 dishonest parties collude with a dishonest data miner, they will be usable to success[17].

5. CONCLUSION

In this paper, various privacy preservation techniques in association rule mining process were discussed along with their advantages and disadvantages. The database sanitization process happens to be NP-Hard problem and there should be some balance between privacy and accuracy. Since privacy protection technique involves multiple disciplines, there are many issues to be addressed. A good privacy preserving association rule mining technique must have the following properties:

- It must achieve good privacy with accuracy
- Data sanitization with minimum negative impact
- The computation cost, communication cost and disclose cost should be minimum

Apart from this, user mobility patterns are also available as mobile data due to enormous growth in spatial and geographical applications. Parameters and framework for evaluating various privacy preserving association rule mining techniques should also be designed for calculating efficiency.

6. REFERENCES

[1] Jiawei Han, Hong Cheng, Dong Xin and Xifeng Yan, “Frequent pattern mining: current status and future directions”, Data Min Knowledge Disc (2007) 15:55–86

[2] Razan Paul, Tudor Groza, Jane Hunter and Andreas Zankl, “Semantic interestingness measures for discovering association rules in the skeletal dysplasia domain”, Journal of Biomedical Semantics 2014, 5:8

[3] Frank S.C. Tseng and Pey-Yen Chen, “Parallel Association Rule Mining by Data De-Clustering to Support Grid Computing”

[4] Shashikumar G. Totad, Geeta R. B, Chennupati R Prasanna, N Krishna Santhosh and Prasad Reddy, “Scaling Data Mining Algorithms to Large and Distributed Datasets”, International Journal of Database Management Systems (IJDMS), Vol.2, No.4, November 2010

[5] Yehuda Lindell and Benny Pinkas, “Secure Multiparty Computation for Privacy-Preserving Data Mining”, The Journal of Privacy and Confidentiality (2009) 1, Number 1, pp. 59-98

[6] P.Kamakshi and Dr.A.Vinaya Babu, “Preserving Privacy and Sharing the Data in Distributed Environment using Cryptographic Technique on Perturbed data”, JOURNAL OF COMPUTING, VOLUME 2, ISSUE 4, APRIL 2010, ISSN 2151-9617

[7] Elisa Bertino, Dan Lin and Wei Jiang, “A Survey of Quantification of Privacy Preserving Data Mining Algorithms”

[8] Li Liu, Murat Kantarcioglu and Bhavani Thuraisingham, “The Applicability of the Perturbation Model-based Privacy Preserving Data Mining for Real-world Data”, Sixth IEEE International Conference on Data Mining-Workshops (ICDMW'06) 0-7695-2702-7/06

[9] Animesh Tripathy and Matrubhumi Pradhan. 2012. A novel framework for preserving privacy of data using correlation analysis. In Proceedings of the International Conference on Advances in Computing, Communications and Informatics (ICACCI '12). ACM, New York, NY, USA, 650-655.

[10] Luciano Bononi, Michele Bracuto, Gabriele D'Angelo and Lorenzo Donatiello. 2005. Concurrent Replication of Parallel and Distributed Simulations. In Proceedings of the 19th Workshop on Principles of Advanced and Distributed Simulation (PADS '05). IEEE Computer Society, Washington, DC, USA, 234-243.

[11] Dragos N. Trinca. 2008. Fast and Cost-Effective Algorithms for Information Extraction in some Computational Domains. Ph.D. Dissertation. University of Connecticut, Storrs, CT, USA. AAI3345214.

[12] Yongcheng Luo, Yan Zhao and Jiajin Le. 2009. A Survey on the Privacy Preserving Algorithm of Association Rule Mining. In Proceedings of the 2009 Second International Symposium on Electronic Commerce and Security - Volume 01 (ISECS '09), Vol. 1. IEEE Computer Society, Washington, DC, USA, 241-245

[13] Jaideep Vaidya, Chris Clifton, Murat Kantarcioglu and A. Scott Patterson. 2008. Privacy-preserving decision trees over vertically partitioned data. ACM Trans. Knowl. Discov. Data 2, 3, Article 14 (October 2008)

[14] Shyue-Liang Wang, Ting-Zheng Lai, Tzung-Pei Hong and Yu-Lung Wu. 2010. Hiding collaborative recommendation association rules on horizontally partitioned data. Intell. Data Anal. 14, 1 (January 2010), 47-67.

[15] Stefan Edelkamp, Stefan Schroedl, and Sven Koenig. 2010. Heuristic Search: Theory and Applications. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

[16] Graham Cormode, Minos Garofalakis, Peter J. Haas, and Chris Jermaine. 2012. Synopsis for Massive Data: Samples, Histograms, Wavelets, Sketches. Found. Trends databases 4, 1–3 (January 2012), 1-294

[17] Ralph Hughes. 2012. Agile Data Warehousing Project Management: Business Intelligence Systems Using Scrum. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.