Secure Collaborative Supply Chain using Privacy Preserving Association Rule Mining: A Survey

Hiren R. kamani
M.Tech Scholar, Computer Engineering
School of Engineering, R K University
Gujarat, India

Supriya Byreddy
Assistant Professor, Computer Engineering
School of Engineering, R. K. University
Gujarat, India

***Abstract*—in this paper I identify a major area of research as a topic for privacy preservation data mining. The numerous algorithm developed in privacy preserving data mining since last decade, but there is no research work applied to particular application level, hence it is unclear that available algorithm in privacy preserving data mining can directly applicable to any specific problem context or not , in this paper I identify an application such that it require sensitive data protection and sophisticated data analysis, the application area is supply chain management ,this is important area in production and operations management. I examine that in upcoming years there are lots of challenges and opportunities available for Privacy preserving data mining in the area of supply chain management.**

Index Terms*—Privacy Preservation Rule Mining, Collaborative Supply Chain, Data Mining, Sensitive Data.

I. INTRODUCTION

Data mining is the knowledge discovery process of finding the useful information and patterns out of large database.[1] In recent times data mining has gained immense importance as it paves way for The management to obtain hidden information and use them in decision-making. In this competitive environment organizations are dependent on data mining for the betterment in service providing, smart decision-making and gain handsome profit[2][3]. To do so organization needs to collect huge amount of data, for example, logistics and supply chains , one of the richest data domain where data doubling every 18 month [4] .

The rise of complex and global business networks means that a majority of the data will be generated outside a company‘s firewall. The data of record will no longer be in a single company‘s database. Traditional ERP systems, IT infrastructure and systems will no longer have all the answers. The new economics of global commerce will force companies to move beyond their mission-critical systems of record towards global collaboration and innovation platforms, which enable business networks and partner ecosystems.

As we enter into this brave new world, businesses will not compete company to company but value chain to value chain. The best companies in the future will need to learn how to be a business network. Unfortunately, the current IT infrastructure and applications are not completely ready for this task. In this paper, I examine the challenges companies face as they transition to a business network in the context of the supply chain and how and why collaborative supply chain based solutions are the only way to deliver on that promise.

Here before collaborating/releasing the dataset to the other party, each party is willing to hide sensitive association rules of its own sensitive products/data. So, the sensitive information (or knowledge) will be protected. In 1999 first time Atallah *et al.* Proposed association rule hiding problem in the area of privacy preserving data mining [5].

Privacy preserving data mining (PPDM) is considered to maintain the privacy of data and knowledge extracted from data mining. It allows the extraction of relevant knowledge and information from large amount of data, while protecting sensitive data or information. To preserve data privacy in terms of knowledge, one can modify the original database in such a way that the sensitive knowledge is excluded from the mining result and non sensitive knowledge will be extracted. In order to protect the sensitive association rules (derived by association rule mining techniques), privacy preserving data mining include the area called “association rule hiding”. The main aim of association rule hiding algorithms is to reduce the modification on original database in order to hide sensitive knowledge, deriving non sensitive knowledge and do not producing some other knowledge.

Rest of this paper is organized as follows: - In Section 2, discusses the association rule mining strategy. The concept of Privacy Preservation rule mining is given in section3. Section 4 presents the existing association rule hiding approaches by identifying open challenges. The metrics used for evaluating sensitive rule hiding approaches are given in section 5. Section 6 concludes my study by identifying future work with references at the end.

II. ASSOCIATION RULE MINING

Let define item set $\boldsymbol{I} = \{\boldsymbol{i_1}, \boldsymbol{i_2}\ldots., \boldsymbol{i_m}\}$ where I is a set of **m** distinct literals, Given a set of transactions **D,** where each transaction $\boldsymbol{T}$ is a set of items such that $\boldsymbol{T} \subseteq \boldsymbol{I}$. we can define an association rule in the form of $\mathbf{X}\boldsymbol{->Y}$ where $\boldsymbol{X} \subset \boldsymbol{I}$, $\boldsymbol{Y}$

$\subset \boldsymbol{I}$ and $X \cap \boldsymbol{Y} = \phi$, where X and Y are called body and head of rule respectively [6].

Strength of a rule whether it is strong or not is measured by two parameters called support and confidence of the rule. These two parameters help in deciding the interestingness of a rule [7], [8].

For a given rule $\boldsymbol{X \Rightarrow Y}$

Support is the percentage of transaction that contains ***X*** both ***Y*** and $(\boldsymbol{X} \cup \boldsymbol{Y})$ or is the proportion of transactions jointly covered by the LHS and RHS and is calculated as:

$$S = |X \cup Y| / |N|$$

Where, ***N*** is the number of transactions.
Confidence is the percentage for a transaction that contains ***X*** also contains ***Y*** or is the proportion of transactions covered by the LHS that are also covered by the RHS and is calculated as

$$C = |X \cup Y| / |X|$$

For the database given in Table1, with a minimum support of 33% and minimum confidence 70% following nine association rules could be found:

C=> A (66.667%, 100%), A, B => C (50%, 75%),
B=>C, A (50%, 75%), C, B => A (50%, 100%)
C=> A, B (50%, 75%), C, A => B (50%, 75%)
B=> C (50%, 75%), C => B (50%, 75%)
B=> A (66.667%, 100%)

Table 1. set of transactional data

TID	ITEMS
T1	ABC
T2	ABC
T3	ABC
T4	AB
T5	A
T6	AC

As shown in Fig. 1, association rule mining works in two-step process:
i) First of all find all frequent item sets; a frequent item set can define as an item set which occur at least as frequently as a pre-determined minimum support count.

ii) Generate strong association rules: this rule is generated based upon user defined minimum support and minimum confidence.

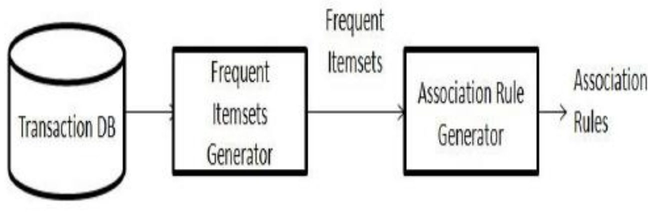

Figure 1: Association rule mining process [9]

III. PRIVACY PRESERVATION RULE MINING

Privacy preserving association rule mining should achieve the following goals: (1) all the sensitive association rules must be hidden in sanitized database. (2) All the rules that are not specified as sensitive can be mined from sanitized database. (3) No new rule that was not previously found in original database can be mined from sanitized database. First goal considers privacy issue. Second goal is related to the usefulness of sanitized dataset. Third goal is the side effect of the sanitization process.

The objective of the association rule hiding problem is to minimally sanitize database in such a way that association rule mining algorithm will not be able to discover sensitive rules and will be able to mine all the non-sensitive rules. The association rule hiding problem can be stated as follows: Given a transactional database D, a set R of relevant rules that are mined from D and a subset R_H of R ,where R_H is the set of sensitive rules, how can we transform D into a database D“ in such a way that the every rule in R can still be mined, except for the rules in R_H.

Thus, the association rule hiding algorithm should transform D to D“ that maximizes the number of rules in R - R_H, that can still be mined. There are two main association rule hiding can be adopted to hide a set R_H of rules (i) either prevent the rules in R_H from being generated, by hiding the frequent sets from which they are derived, or (ii) reduce the confidence of the sensitive rules, by bringing it below a user specified threshold. In [10] the authors demonstrate that solving this problem by reducing the support of the large item sets by removing items from transactions is an NP-hard problem.

IV. ASSOCIATION RULE HIDING APPROACH

Sensitive association rule hiding is a subfield of Privacy Preserving Data Mining (PPDM). Privacy preserving data mining has been recently introduced to cope with privacy issues related to the data subjects in the course of mining of the data. Association Rule Hiding approaches can be classified into five classes: heuristic based approaches, border based approaches, exact approaches, reconstruction based approaches and cryptography based approaches.

A. Heuristic based approaches

These approaches can be further divided in to two groups based on data modification techniques: data distortion techniques and data blocking techniques.

- *Data-Distortion technique*

This technique is based on data transformation. it changes a selected set of 1-values to 0- values (delete items) or 0-values to 1- values (add items), if we assume two-dimensional transaction database matrix then aim of this technique is to reduce the support as well as confidence of the sensitive rule as much as possible then the user predefined threshold. Verykios et al. [11] proposed five assumptions for hiding sensitive knowledge in database by reducing support or confidence of sensitive rules.

- *Data Blocking Technique*

Y. Saygin et al.[12][13] were the first to propose blocking technique in order to increase or decrease the support of the items by replacing 0's or 1's by unknowns "?", so that it become difficult for an adversary to know the value behind "?". This technique is effective and provides certain privacy. Wang and Jafari [10] proposed more efficient approaches then other approaches as in [12][13]. While hiding many rules at a time, they require less number of database scans and prune more number of rules. Now, consider the table shown in Table 2. For rule A=>C, Support (A=>C) = 80% and Confidence (A=>C) = 100%. After fuzzifying the values, support and confidence becomes marginal. So in new database: 60% ≤ Confidence (A=>C) ≤ 100% and 60% ≤ Support (A=>C) ≤ 80% [14].

Table 2. Hiding A=>C by blocking [14]

A	B	C	D
1	1	1	0
1	0	1	0
0	1	0	1
1	1	1	0
1	0	1	1

⇨

A	B	C	D
1	1	1	0
1	0	?	0
?	1	0	1
1	1	1	0
1	0	1	1

B. *Cryptography based approaches*

Cryptography based approaches used in multiparty computation. If the database of one organization is distributed Among several sites, then secure computation is needed between them. These approaches encrypt original database instead of distorting it for sharing. So they provide input privacy. Vaidya and Clifton [15] proposed a secure approach for sharing association rules when data are vertically partitioned. The authors in [16] addressed the secure mining of association rules over horizontal partitioned data.

C. *Border based approaches*

Border based approach uses the theory of borders presented in [17]. These approaches pre-process the sensitive rules so that minimum numbers of rules are given as input to hiding process. The sensitive association rules are hidden by modifying the borders in the lattice of the frequent and the infrequent item set of the original database. The item sets which are at the position of the borderline separating the frequent and infrequent item sets forms the borders. So, they maintain database quality while minimizing side effects.

D. *Exact approaches*

This approach formulates the hiding process as a constraints satisfaction problem (CSP) or an optimization problem which is solved by binary integer programming (BIP). These approaches provide better solution than other approaches. But they suffer from high time complexity to CSP. Gkoulalas and Verykios [18] proposed an approach to find optimal solution for rule hiding problem which tries to minimize the distance between the original database and its sanitized version.

The authors in [19] proposed a novel, exact border-based approach that provides an optimal solution for the hiding of sensitive frequent item sets by minimally extending the original database by a synthetically generated database part - the database extension. Extending the original database for sensitive item set hiding is proved to provide optimal solutions to an extended set of hiding problems compared to previous approaches and to provide solutions of higher quality.

E. *Reconstruction Based Approach*

This approach is implemented by perturbing the data first and reconstructing the distributions at an aggregate level in order to perform the association rules mining. Mielikainen [20] analyzed that inverse frequent set mining computational complexity is very high and it causes more problems. In this approach it first places the original data aside and start from knowledge base. To sanitize, it conceals the sensitive rules by sanitizing item set lattice rather than sanitizing original dataset. Later Y. Guo[21] proposed a FP tree approach which is based on inverse frequent set mining algorithm. The proposed model has three phases, first phase generates frequent item sets from the original database, second phase performs sanitization algorithm over frequent item sets by selecting hiding strategy and identifying sensitive frequent items sets according to sensitive association rules. The third phase generates sanitized database by using inverse frequent item set mining algorithm and then releases this database see fig. 2.

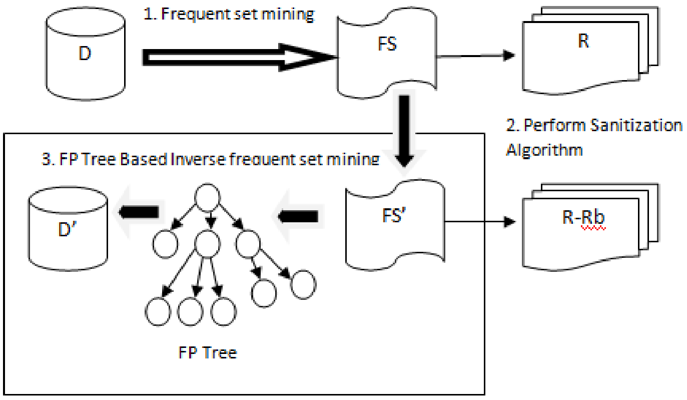
1. Frequent set mining
D
FS
R
2. Perform Sanitization Algorithm
3. FP Tree Based Inverse frequent set mining
D'
FS'
R-Rb
FP Tree

Figure 1: Framework of Reconstruction based approach [21]

V. SUMMERY OF ASSOCIATION RULE HIDING APPROACHES

The advantages and limitations of the above presented association rule hiding approaches are given in table 3.

Table 3. Summary of association rule hiding approaches

Advantages	Limitations
Heuristic Based Approaches (Distortion technique)	
Efficiency, scalability and quick responses due to which it is getting focus by majority of the researchers. Totally takes best decision	Produce undesirable side effects in new database (i.e. Lost rules and new rules).
Heuristic Based Approaches (Blocking technique)	
Maintains truthfulness of the underlying data. Minimizes side effects.	Difficult to reproduce original dataset.
Border Based Approaches	
Maintains data quality by greedily selecting the modification with minimal side effects. Improvement over pure heuristic approach.	Unable to identify optimal hiding solution But still dependent on heuristic to decide upon the item modification.
Exact Approaches	
Guarantees quality for hiding sensitive information than other approaches.	But requires very high time complexity due to integer programming
Reconstruction Approaches	
Create privacy aware database by exacting sensitive characteristic from the original database. Lesser side effects in database than heuristic approach.	The open problem is to restrict the number of trans-actions in the new database.
Cryptographic Approaches	
Secure mining of association rule over partitioned database.	Communication and computation cost is higher.

VI. EVALUATION METRICS

Following metrics are used to evaluating association rule hiding algorithms [22][23].

1) ***Efficiency***- It is measured in terms of CPU-time, space requirements and communication required for hiding. In short, good performance in terms of resources allocated.

2) ***Scalability***- It is measured in terms of good performance for increasing sizes of input datasets.

3) ***Data quality***- Data quality parameters are accuracy measure, completeness, consistency which is in relationship to preservation of original data values and of data mining results.

4) ***Hiding failure***- It is the percentage of the portion of information that fails to be hidden. It is derived by, HF = |Rs(D')| / |Rs(D)| where, |Rs(D')| are the number of sensitive rules appearing in the sanitized database and |Rs(D)| are the number of sensitive rules in the original database.

5) ***Privacy level***- It measures the degree of uncertainty according to which the protected information can still be predicted.

6) ***Lost Rules cost***- It measures the number of no sensitive association rules found in the original database but not in sanitized database.

7) ***Ghost Rules***- It measures the percentages of rules that are not present in the original database but can be derived from sanitized database.

8) ***Dissimilarity***- It quantifies difference between original database and sanitized database.

VII. CONCLUSION

The ever increasing ability to identify and collect large amounts of data, analyzing the data using data mining process and decision on the results gives prospective benefits to organizations .But, such repositories also contains private and sensitive information and releasing the personal information can cause significant damage to data owner. Hence there is increased need to discover and distribute the databases, without compromising the privacy of the individual's data.

Association rule hiding is an important concept in the area of privacy preserving data mining. It protects the privacy of sensitive information in databases against the association rule mining approaches. In this paper, we surveyed methods of hiding sensitive association rules by identifying some open challenges that will be useful to research community in this area. Here existing approaches provide only the approximate solution to hide sensitive knowledge. There is need of finding exact solution to the privacy problem in database disclosure.

In future, hybrid technique can be found to reduce the side effects and increase the efficiency by reducing the modifications on database, while hiding the association rules. An algorithm for incremental environment can also be developed, as most of the current frequent hiding algorithms are designed for static database.

REFERENCES

[1] H. Kargupta, S. Datta, Q. Wang, and K. Sivakumar, "On the privacy preserving properties of random data perturbation techniques" In Proc. of 3rd IEEE Int. Conf. on Data Mining, Washington, DC, USA, pages99– 106, 2003.

[2] R.Agarwal and R.Srikant, "Privacy preserving data mining", In Procseedings of the 19th ACM SIGMOD onference on Management of Data ,Dallas,Texas,USA, May2000

[3] J. Canny, "Collaborative filtering with privacy". In IEEE Symposium on security and privacy , pages 45-57 Oakland, May 2002.

[4] Kamesh Pemmaraju, ―Five Challenges of Managing Big Data in Supply Chains‖ http://sandhill.com/article/five-challenges-of-managing-big-data-in-supply-chains/ (2011).

[5] M. Atallah, E. Bertino, A. Elmagamind, M. Ibrahim, and V. S. Verykios "Disclosure limitation of sensitive rules," .In Proc. of the 1999 IEEE Knowledge and Data Engineering Exchange Workshop(KDEX 1999), pp. 45-52.

[6] Marzena Kryszkiewicz. "Representative Association Rules", In proceedings of PAKDD'98, Melbourne,Australia(Lecture notes in artificial Intelligence,LANI 1394, Springer-Verleg, pp 198-209, (1998)

[7] A. Jafari and S-L. Wang, "Using unknowns for hiding sensitive predictive association rules," In IEEE International Conference on Information Reuse and Integration, pp. 223 – 228, (2005)

[8] Yucel Saygin, Vassilios S. Verykios, Chris Clifton. "Using unknowns to prevent discovery of association rules", ACM SIGMOD Record Volume 30 Issue 4, pp. 45 - 54 , (2001)

[9] H. Q. Le, "Association rule hiding in risk management for retail supply chain collaboration" 2013 Elsevier on Computers in Industry 64, pp. 776-784, 2013.

[10] Atallah**,** M., Bertino, E., Elmagarmid, A., Ibrahim, M., and Verykios, V.S,"Disclosure limitation of sensitive rules", In: Scheuermann P, ed. Proc. of the *IEEE Knowledge and Data Exchange Workshop (KDEX'99)*. IEEE Computer society, 1999. pp. 45-52.

[11] Verykios, V.S., Elmagarmid, A., Bertino, E., Saygin, Y., and Dasseni, E. "Association rule hiding", *IEEE Transactions on Knowledge and Data Engineering*, 2004, 16(4): pp. 434-447.

[12] Y. Saygin, V. Verykios, and C. Clifton, "Using Unknowns to Prevent Discovery of Association Rules" ACM SIGMOD, Vol. 30, No. 4, pp. 45–54, 2001.

[13] Y. Saygin, V. Verykios, and A. Elmagarmid, "Privacy preserving association rule mining," In: Proc. Int'l. Workshop on Research Issues in Data Engineering (RIDE 2002), pp.151–163, 2002.

[14] K. Shah, A. Thakkar and A. Ganatra, "A Study on Association Rule Hiding Approaches". (IJEAT)International Journal of Engineering and Advanced Technology, vol 3, issue-3, February 2012, pp. 72-76.

[15] J. Vaidya and C. Clifton, "Privacy preserving association rule mining in vertically partitioned data" ,In proc. Int‟l Conf. *Knowledge Discovery and Data Mining*, July 2002, pp. 639- 644,.

[16] 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(9), Sept. 2004, pp. 1026-1037.

[17] H. Mannila and H. Toivonen, "Levelwise search and borders of theories in knowledge discovery" *Data Mining and Knowledge Discovery*, vol.1 (3), Sep. 1997, pp. 241-258.

[18] Gkoulalas-Divanis and V.S. Verykios, "An Integer Programming Approach for Frequent Itemset Hiding", In Proc. *ACM Conf. Information and Knowledge Management (CIKM '06)*, Nov. 2006.

[19] Gkoulalas-Divanis and V.S. Verykios, "Exact Knowledge Hiding through Database Extension," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21(5), May 2009, pp. 699-713.

[20] T. Mielikainen, "On inverse frequent set mining", In Proc. of 3rd IEEE ICDM Workshop on Privacy Preserving Data Mining. IEEE Computer Society, 2003, pp.18-23.

[21] Y. Guo, "Reconstruction-Based Association Rule Hiding" In Proc. of SIGMOD2007 Ph.D. Workshop on Innovative Database Research 2007(IDAR2007), June 2007, pp.51-56.

[22] V. S. Verkios, "Association rule hiding methods" 2013 John Wiley & Sons, Inc, Vol. 3, January/February 2013, pp. 28-38.

[23] C. Modi, U.P. Rao and D.R.Patel, "A Survey on Preserving Privacy for Sensitive Association Rules in Databases" Springer-Verlag Berlin Heidelberg 2010, pp. 538-544.