International Journal of Computer Science and Network (IJCSN)
Volume 1, Issue 3, June 2012 www.ijcsn.org ISSN 2277-5420

Hiding Sensitive Association Rule Using Clusters of Sensitive Association Rule

1Sanjay keer, 2Prof. Anju Singh

1 MTech, CSE Department, Barkatullah University Institute of Technology Bhopal, M.P., India

2 Asst. Prof., CSE Department, Barkatullah University Institute of Technology Bhopal, M.P., India
Asingh0123 @rediff.com

Abstract

The security of the large database that contains certain crucial information, it will become a serious issue when sharing data to the network against unauthorized access. Association rules hiding algorithms get strong and efficient performance for protecting confidential and crucial data. The objective of the proposed Association rule hiding algorithm for privacy preserving data mining is to hide certain information so that they cannot be discovered through association rule mining algorithm. The main approached of association rule hiding algorithms to hide some generated association rules, by increase or decrease the support or the confidence of the rules. The association rule items whether in Left Hand Side (LHS) or Right Hand Side (RHS) of the generated rule, that cannot be deduced through association rule mining algorithms. The concept of Increase Support of Left Hand Side (ISL) algorithm is decrease the confidence of rule by increase the support value of LHS. It doesn't work for both side of rule. It works only for modification of LHS. In this paper, we propose a heuristic algorithm named ISLRC (Increase Support of L.H.S. item of Rule Clusters) based on ISL approach to preserve privacy for sensitive association rules in database. Proposed algorithm modifies fewer transactions and hides many rules at a time. The efficiency of the proposed algorithm is compared with ISL algorithms.

Keywords: Data mining, ISL, Association Rule Hiding, Sensitivity, Clustering.

I.INTRODUCTION

Successful applications of data mining techniques have been demonstrated in many areas that benefit commercial, social and human activities. Along with the success of these techniques, they pose a threat to privacy. One can easily disclose other's sensitive Information or knowledge by using these techniques. So, before releasing database, sensitive information or knowledge must be hidden from unauthorized access. To solve privacy problem, PPDM has become a hotspot in data mining and database security field. Researchers have proposed several approaches for knowledge hiding, in context of association rule mining. M.Attallah et al. [1] was the first to propose heuristic algorithms for preventing disclosure of sensitive knowledge. Oliveria et al. [3] presented taxonomy of attacks against sensitive knowledge. In following, we show the necessity of sensitive association rule hiding in real life application Successful applications of data mining have been demonstrated in marketing, business, medical analysis, product control, engineering design, bioinformatics and scientific exploration, among others. The current status in data mining research reveals that one of the current technical challenges is the development of techniques that incorporate security and privacy issues. Providing security to sensitive data against unauthorized access has been a long term goal for the database security research community and for the government statistical agencies. Recent advances in data mining technologies have increased the disclosure risks of sensitive data. Hence, the security issue has become, recently, a much more important area of research. In this paper, we propose a heuristic algorithm named ISLRC (Increase Support of L.H.S. item of Rule Clusters) to preserve privacy for sensitive association rules in database. Proposed algorithm modifies fewer transactions and hides many rules at a time. So, it is more efficient than other heuristic approaches. Moreover it maintains data quality in sanitized database. So, sanitized database is as useful as original database. A detailed description of proposed ISLRC algorithm is given in section 3.

Problem Description

The association rule hiding problem is to sanitize database in a way that through association rule mining one will not be able to disclosing the sensitive rules and will be able to mine all the non-sensitive rules. More specifically, the problem statement can be defined as follows: Let given dataset D, a set of association rules R over D is given and also R0□R, R0 is specified as sensitive rules set. Now, the problem is to find sanitized database D0 such that there exist only a set of rules RR0, can be mined. Finding an optimal solution to this problem is NP-hard, proved in [1]. The rest of this paper is organized as follows. In section 2, we discuss related background and existing approaches. In section 3, a detailed description of proposed ISLRC algorithm is given. An example demonstrating ISLRC algorithm is given in section 4. In section 5 we analyze and discuss the performance results of proposed algorithm

II. RELATED WORK

Association rule using support and confidence can be defined as follows. Let I= $\{i_1,\ldots,i_m\}$be a set of items. Database D=$\{T_1,\ldots,T_n\}$is a set of transactions, where T_iI($1\leq i\leq m$). Each transaction T is an itemset such that TI. A transaction T supports X, a set of items in I, if XI. The association rule is an implication formula like XY, where XI, YI and XY=. The rule with support s and confidence c is called, if |XY|/|D| ≥ s and |XY|/|X| ≥ c. Because of interestingness, we consider user specified thresholds for support and confidence, MST (minimum support threshold) and MCT (minimum confidence threshold). A detailed overview of association rule mining algorithms are presented in [2]. Privacy preserving association rule mining should achieve one of 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 related to the side effect of the sanitization process.

Many approaches have been proposed to preserve privacy for sensitive knowledge or sensitive association rules in database. They can be classified in to following classes: heuristic based approaches, border based approaches, exact approaches, reconstruction based approaches, and cryptography based approaches. In following, a detailed overview of these approaches is given.

A. Heuristic Based Approaches

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

Data distortion techniques try to hide association rules by decreasing or increasing support (or confidence). To increase or decrease support (or confidence), they replace 0's by 1's or vice versa in selected transactions. So they can be used to address the complexity issue. But they produce undesirable side effects in the new database, which lead them to suboptimal solution. M.Attallah et al. [1] were the first proposed heuristic algorithms. The proof of NP-hardness of optimal sanitization is also given in [1]. Verykios et al. [11] proposed five assumptions which are used to hide sensitive knowledge in database by reducing support or confidence of sensitive rules.

Y-H Wu et al. [14] proposed method to reduce the side effects in sanitized database, which are produced by other approaches [11]. K.Duraiswamy et al. [19] proposed an efficient clustering based approach to reduce the time complexity of the hiding process. **Data blocking techniques** replace the 0's and 1's by unknowns ("?") in selected transaction instead of inserting or deleting items. So it is difficult for an adversary to know the value behind "?". Y.Saygin et al. [7][15] were the first to introduce blocking based technique for sensitive rule hiding. The safety margin is also introduced in [7] to show how much below the minimum threshold new support and confidence of a sensitive rule should. Wang and Jafari [17] proposed more efficient approaches than other approaches presented in [7][15].

B. Border Based Approaches

Border based approaches use the notion of borders presented in [4]. These approaches preprocess the sensitive rules so that minimum numbers of rules are given as input to hiding process. So, they maintain database quality while minimizing side effects. Sun and Yu [10] were the first to propose the border revision process. Hiding process in [10] greedily selects those modifications that lead to minimal side effects. The authors in [13] presented more efficient algorithms than other similar approaches presented in [10].

C. Exact Approaches

Exact approaches formulate hiding problem to constraint satisfaction problem (CSP) and solve it by using binary integer programming (BIP). They provide an exact (optimal) solution that satisfies all the constraints. However if there is no exact solution exists in database, some of the constraint are relaxed. These approaches provide better solution than other approaches. But they suffer from high time complexity to CSP. Gkoulalas and Verykios [6] proposed an approach to find optimal solution for rule hiding problem. The authors

in [12] proposed a partitioning approach for the scalability of the algorithm.

D. Reconstruction Based Approaches

Reconstruction based approaches generate privacy aware database by extracting sensitive characteristics from the original database. These approaches generate lesser side effects in database than heuristic approaches. Mielikainen [9] was the first analyzed the computational complexity of inverse frequent set mining and showed in many cases the problems are computationally difficult.

Y. Guo [16] proposed a FP tree based algorithm which reconstruct the original database by using non characteristic of database and efficiently generates number of secure databases.

E. 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 [5] proposed a secure approach for sharing association rules when data are vertically partitioned. The authors in [20] addressed the secure mining of association rules over horizontal partitioned data.

We proposed a more efficient heuristic algorithm than other heuristic approaches presented in this section

III. ISLRC- PROPOSED HEURISTIC BASED ALGORITHM

To hide an association rule like X Y, we decrease its confidence (|XY|/|X|) to smaller than specified minimum confidence threshold (MCT). We increase the support of X (L.H.S. of the rule) in the most sensitive transactions. To increase support count of an item, we put one item from selected transaction by changing from 0 to 1.

A. Framework of ISLRCAlgorithm

Some important concepts used in proposed framework of ISLRC algorithm are as follows:-

1) Item Sensitivity: is the frequency of data item exists in the number of the sensitive association rule containing this item. It is used to measure rule sensitivity.

2) Rule Sensitivity: is the sum of the sensitivities of all items containing that association rule.

3) Cluster Sensitivity: is the sum of the sensitivities of all association rules in cluster. Cluster sensitivity defines the rule cluster which is most affecting to the privacy.

4) Sensitive Transaction: is the transaction in given database which contains sensitive item.

5) Transaction sensitivity: is the sum of sensitivities of sensitive items contained in the transaction.

in decreasing order of their sensitivity and sensitive transactions supporting first rule-cluster are sorted in decreasing order of their sensitivity. Transaction change continues until all the sensitive rules in all clusters are not hidden. Finally modified transactions are Detailed overview of sensitivities is given in [21]. The proposed framework of ISLRC algorithm is shown in Figure.1. Initially association rules (AR) are mined from the source database D by using association rule mining algorithms e.g. Apriori algorithm in [2]. Then sensitive rules (SR) are specified from mined rules. Selected rules are clustered based on common L.H.S. item of the rules. Rule-clusters are denoted as RCLs. Then for each Rule-cluster sensitive transactions are indexed. Sensitivity of each item (and each rule) in each Rule-cluster is calculated. Rule-Clusters are sorted

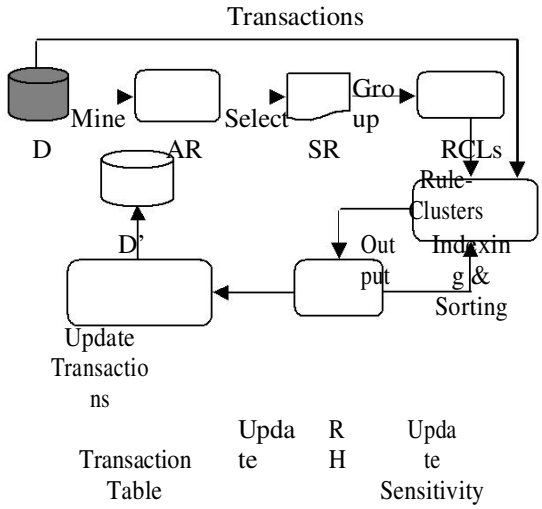

Figure 1. Framework of proposed ISLRC algorithm.

After sorting process, rule hiding (RH) process hides all the sensitive rules in sorted transactions for each cluster by using strategy mentioned in this section and updates the sensitivity of sensitive transactions in other cluster. Hiding process starts from lowest sensitive updated in original database and produced database is called sanitized database D' which ensures certain privacy for specified rules and maintains data quality.

B. ISLRC Algorithm

According to above presented framework for hiding association rules in database, the proposed ISLRC algorithm is shown in Figure 2. By using given minimum support threshold (MST) and minimum confidence threshold (MCT), algorithm first generates the possible number of association rules from source database D. Now some of the generated association rules are selected as sensitive rule set (set RH) by database owner. Rules with only single L.H.S. item are specified as sensitive. Then algorithm finds C clusters based on common L.H.S. item in sensitive rule set RH and calculates the sensitivity of each cluster. After that it index sensitive transactions for each cluster and sorts all the clusters by decreasing order of their sensitivities. For the highest sensitive cluster, algorithm sorts sensitive transaction in decreasing order of their sensitivities.

INPUT: Source database D, Minimum Confidence Threshold (MCT), Minimum support threshold (MST).

TID	Items	Items (Binary Form)
1	abce	11101
2	ace	10101
3	abc	11100
4	cd	00110
5	ab	11000
6	abc	11100
7	de	00011

OUTPUT: The sanitized database D'.

1. **Begin**
2. Generate association rules.
3. Selecting the Sensitive rule set RH with single antecedent and consequent e.g. x□y.
4. Clustering-based on common item in L.H.S. of the selected rules
5. Find sensitivity of each item in each cluster.
6. Find the sensitivity of each rule in each cluster.
7. Find the sensitivity of each cluster
8. Index the sensitive transactions for each cluster.
9. Sort generated clusters in decreasing order of their sensitivity.
10. For the first cluster, sort selected transaction in decreasing order of their sensitivity
11. For each cluster c□ C
12. {
13. While(all the sensitive rules c are not hidden)
14. {
15. Take first transaction for cluster c.
16. put common L.H.S. item into the transaction.
17. Update the sensitivity of new item for modified transaction in other cluster and sort it.
18. For i = 1 to no. of rule Rh□c
19. {
20. Update support and confidence of the rule r□ c.
21. If(support of r < MST or confidence of r < MCT)
22. {
23. Remove Rule r from Rh
24. }
25. }
26. Take next transaction.
27. }
28. End while
29. }
30. End for
31. Update the modified transactions in D.
32. **End**

Frequent Itemsets with Support Count
a:5,b:4,c:5, ab:4,ac:4,bc:3, abc:3

Figure 2.

Now, the hiding Process tries to hide all the sensitive rules by putting common L.H.S. item of the rules in cluster, into the sensitive transactions. While loop continues until all the rules are not hidden in cluster c. Every time in while loop it updates the sensitivity of new item for modified transaction in other cluster and sorts it. Finally algorithm updates all the modified transactions in original database. Proposed ISLRC algorithm produces sanitized database D, in which most of the sensitive rules are hidden. This algorithm hides many rules in an iteration of hiding process and it modifies less transaction in database.

IV EXAMPLE

The following example illustrates proposed ISLRC algorithm. A sample transaction database D is shown in Table 1. TID shows unique transaction number. Binary valued item shows whether an item is present or absent in that transaction. Suppose MST and MCT are selected 40% and 75% respectively. Table 2 shows frequent itemsets satisfying MST, generated from sample database D.

Cluster-2(b) (b□a, b□c)	
TID	**Sensitivity**
1	4
2	2
3	4
4	1
5	3
6	4
7	0

Table 1. Sample Transaction Database D.

In following, the possible number of association rules Satisfying MST and MCT, generated by Apriori algorithm [2]: a□b, b□a, a□c, c□a, b□c, ab□c, b□ac, ac□b, bc□a, Suppose the rules a□b, a□c, b□a and b□c specified as sensitive and should be hidden in sanitized database. There are two different L.H.S. items in selected rules, named "a" and "b". As shown in Table 3, Algorithm generates two clusters based on common L.H.S. item of the selected rules.

Cluster-1(a) (a⇒b, a⇒c)	
TID	**Sensitivity**
1	4
2	3
3	4
4	1
5	3
6	4
7	0

Table 2. Frequent Item sets with Support Count.

Item	Sensitivity
a	2
b	1
c	1
Total sensitivity	4

Table 3. Clusters generated by ISLRC.

Cluster-1(a) (a□b, a□c)

Cluster1 includes sensitive rules namely b□d, c□d, Where cluster 2 includes b□a and b□c. For cluster 1 sensitivities of items a, b, and c have 2,1and 1 respectively, where Sensitivities for items b, a and c in cluster 2 have 2,1 and 1 Respectively. Total sensitivity for cluster-1 and cluster-2 is 4 and 4 respectively. For each cluster, sensitive transactions are indexed. Indexed transactions with their sensitivity are shown in table 4. Clusters are sorted based on their sensitivity. For the first cluster (here cluster-1), algorithm sorts transactions in decreasing order of their sensitivity.

Cluster-2(b) (b⇒a, b⇒c)

Table 4. Clusters generated by ISLRC algorithm.

Item	Sensitivity
b	2
a	1
c	1
Total sensitivity	4

Table 5. Sanitized Databases.

TID	Items
1	abce
2	ace
3	abc
4	cd
5	ab
6	abc
7	ade

TID	Items
1	abce
2	ace
3	abc
4	cd
5	ab
6	abc
7	abde

(a)Sanitized Database D1. (b) Final Sanitized Database

Hiding process of algorithm modifies seventh transaction by putting item a (common L.H.S. of rules in cluster-1). Table 5(a) shows sanitized database after first iteration. Now, the support or confidence for all the rules in cluster-1 is decreased below the minimum thresholds. Then next cluster is taken. After one iteration, final sanitized database as shown in Table 5(b) is generated. Now, if we mine association rules from final sanitized database, we can see that most of the specified sensitive rules are hidden and very few side effects produced. But using only two iterations and modifying only one transaction, algorithm successfully hides many sensitive rules. So, ISLRC provides database quality while preserving privacy.

V. Result

We can see that simple by ISL algorithm if we want to hide b and a, we check it by modifying the transaction T7 of Table1 from de to bde (i.e. from 00011 to 01011) in Table6, we can hide only two rules b□c, b□ac, and remaining seven rules are not hidden.

TID	Items
1	abce
2	ace
3	abc
4	cd
5	ab
6	abc
7	bde

Table 6. Transaction changed by ISL.

But by using ISLRC algorithm we hide the four rules a□c, b□c, ab□c and b□ac in first iteration. And only five rules are left. That we can also hide by next iteration of ISLRC algorithm.

.

VI. CONCLUSION AND FUTURE SCOPE

In this paper, we proposed a heuristic algorithm named ISLRC which hides many sensitive association rules at a time while maintaining database quality. Several existing approaches regarding sensitive rule hiding problem are also discussed. Our proposed algorithm hides only rules that contain single item on L.H.S. of the rule. But it is more efficient than other heuristic approaches. Proposed algorithm can be modified to hide sensitive rules which contain different number of L.H.S. items.

REFERENCES

[1] M. Atallah, E. Bertino, A. Elmagarmid, M. Ibrahim, and V. S.
Verykios "Disclosure limitation of sensitive rules,".In Proc. of the 1999 IEEE Knowledge and Data Engineering Exchange Workshop (KDEX'99), pp. 45–52, 1999.
[2] J. Han and M. Kamber, Data Mining: Concepts and Techniuqes. Morgan Kaufmann Publishers, San Francisco, CA, 2001, pp. 227–245.
[3] S.R.M. Oliveira, M., O.R. Zaiane, and Y. Saygin, "Secure Association Rule Sharing," In Proc. of the 8th Pacific-Asia Conf. PAKDD2004, Sydney, Australia, pp. 74–85, May 2004.
[4] H. Mannila and H. Toivonen, "Levelwise search and borders of theories in knowledge discovery," Data Mining and Knowledge Discovery, vol.1(3), pp. 241–258, Sep. 1997.
[5] J. Vaidya and C. Clifton, "Privacy preserving association rule mining in vertically partitioned data," In
proc. Int'l Conf. Knowledge Discovery and Data Mining, pp. 639–644, July 2002.
[6] A. 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.
[7] Y.Saygin, V. S. Verykios, and C. Clifton, "Using Unknowns to Prevent Discovery of Association Rules," ACM SIGMOD, vol.30(4), pp. 45–54, Dec. 2001.
[8] I.N. Fovino, and A. Trombetta, "Information Driven Association Rule Hiding Algorithms," In Proc. 1st Int'l Conf. on Information Technology, pp.1–4, May 2008.
[9] T. Mielikainen, "On inverse frequent set mining," In Proc. 3rd IEEE ICDM Workshop on Privacy Preserving Data Mining. IEEE Computer Society, pp.18–23, 2003.
[10] X. Sun and P.S. Yu, "A Border-Based Approach for Hiding Sensitive Frequent Itemsets," In Proc. Fifth IEEE

Int'l Conf. Data Mining (ICDM '05), pp. 426–433, Nov. 2005.
[11] V.S. Verykios, A.K. Elmagarmid, E. Bertino, Y. Saygin, and E. Dasseni, "Association rule hiding," IEEE Transactions on Knowledge and Data Engineering, vol. 16(4), pp. 434–447, April 2004.
[12] A. Gkoulalas-Divanis and V.S. Verykios, "Exact Knowledge Hiding through Database Extension," IEEE Transactions on Knowledge and Data Engineering, vol. 21(5), pp. 699–713, May 2009.
[13] Moustakides and V.S. Verykios, "A Max-Min Approach for Hiding Frequent Itemsets," In Proc. Sixth IEEE Int'l Conf. Data Mining (ICDM '06), pp. 502–506, April 2006.

[14] Y. H. Wu, C.M. Chiang and A.L.P. Chen, "Hiding Sensitive Association Rules with Limited Side Effects," IEEE Transactions on Knowledge and Data Engineering, vol.19(1), pp. 29–42, Jan. 2007.
[15] Y. Saygin, V. S. Verykios, and A. K. Elmagarmid, "Privacy preserving association rule mining," In Proc. Int'l Workshop on Research Issues in Data Engineering (RIDE 2002), 2002,pp. 151–163.
[16] Y. Guo, "Reconstruction-Based Association Rule Hiding," In Proc. Of SIGMOD2007 Ph.D. Workshop on Innovative Database Research 2007(IDAR2007), June 2007.
[17] S.L.Wang and A. Jafari, "Using unknowns for hiding sensitive predictive association rules," In Proc. IEEE Int'l Conf. Information Reuse and Integration (IRI 2005), pp. 223–228, Aug. 2005.
[18] Charu C. Aggarwal, Philip S. Yu, Privacy-Preserving Data Mining: Models and Algorithms. Springer Publishing Company Incorporated, 2008, pp. 267-286.
[19] K. Duraiswamy, and D. Manjula, "Advanced Approach in Sensitive Rule Hiding" Modern Applied Science, vol. 3(2), Feb. 2009.
[20] 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), pp. 1026-1037, Sept. 2004.
[21] S. Wu, H. Wang, "Research On The Privacy Preserving Algorithm Of Association Rule Mining In Centralized Database," Int'l Symposiums on
Information Processing (ISIP), pp. 131 – 134, May 2008