Privacy Preservation in Data Mining By Clustering of Sensitive Association Rule

Neelkamal Upadhyay, Kuldeep Tripathi, Ashish Mishra

***Abstract*— In recent years, data mining is a popular analysis tool to extract knowledge from collection of large amount of data. One of the great challenges of data mining is finding hidden patterns without revealing sensitive information. For protection of confidential and crucial data Association rules hiding algorithms get strong and efficient performance. The objective of the proposed Association rule hiding algorithm is to hide certain information for privacy preserving data mining so that they cannot be discovered through association rule mining algorithm. The process for association rule hiding algorithms is to hide some generated association rules, through decrease or increase the confidence or support of the of the association rules. In this paper we present a new approach that necessarily changes few transactions in the transaction database by decreasing support or confidence of sensitive rules without any side effect. At a time proposed algorithm changes fewer transactions and hides many rules at a time. At the end we compared the efficiency of the proposed algorithm with the existed one.**

***Index Terms*— Data mining, ISL, Association Rule Hiding, Sensitivity, clustering.**

I. INTRODUCTION

Data mining is the non-trivial process of identifying valid, novel, potentially useful and ultimately understandable pattern in data. Data mining represents the integration of several fields, including machine learning, database systems, data visualization, statistics and information theory. Several major data mining techniques have been developed and used in data mining projects recently including association, classification, clustering, prediction and sequential patterns.

Today is the generation of data, there is huge amount of data being produced every day from different resources. It is estimated that the amount of data stored in different database is almost doubled in every two years, this data store in storage devices in form of raw data. Thus, there is a need of some techniques in order to extract useful pattern or information from stored data. The Data mining is the technique that extracts the knowledge from the large volume of data. Basically data mining is a process for analyzing data from different prospective and generating some useful information; the extracted information may be used to grow up the business by different organization, for example by extracting the knowledge form market basket database. The market owners may increase their revenue by offering many exciting offer for customer, the extracted information from the data may contain sensitive information like purchasing habits of customer, confidential data of some organization etc.

Besides extracting information or knowledge from raw data, there is also need for some technique or scheme that deal with security of that information, privacy preserving in data mining (PPDM) is the technique that deal with the security of the information that extracted by data mining techniques, PPDM allow to mine the information from large amount of data while protecting sensitive information defined by the data base owner, or the information that database owner do not want to disclose. The main aim of PPDM is to minimized the risk of misuse of data while does not affect the data mining techniques.

In this paper, we propose a heuristic algorithm named MISL (Modified ISL) 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 MISL algorithm is given in section 3.

Problem Description

In the case of privacy preserving association rule mining, we do not concentrate on privacy of individuals. Rather, we concentrate on the problem of protecting sensitive knowledge mined from databases. The sensitive knowledge is represented by a special group of association rules called sensitive association rules. These rules are most important for strategic decision and must remain private (i.e., the rules are private in the company or organization owning the data). The problem of protecting sensitive knowledge in transactional databases draw the assumption that Data owners have to know in advance some knowledge (rules) that they want to protect. Such rules are fundamental in decision making, so they must not be discovered.

More specifically, the problem statement can be defined as follows: Suppose the Dataset D ,set of association rule R and set of sensitive rules SR over D are given .Now the main purpose is to find new sanitized database SD , in which only a set of sensitive rules SR can be mined. The rest of this paper is organized as follows. In section 2, we discuss related background and existing approaches. In section 3, a detail description of proposed MISL algorithm is given. An example demonstrating ISL algorithm is given in section 4. In

Manuscript received November 09, 2014.

Neelkamal Upadhyay, pursuing M.Tech in Computer Science Engineering, Computer Science Engineering Department, SISTec, Bhopal, M.P., India, 09424747687,

KuldeepTripathi, Assistant Professor, Computer Science Engineering Department, SISTec, Bhopal, M.P., India, 08889951188,

Ashish Mishra, Head of Department, Computer Science Engineering Department, SISTec, Bhopal, M.P., India, 09425376828,

section 5 we analyze and discuss the performance results of proposed algorithm.

II. RELATED WORK

Let I = {i1,…., in} be a set of items. Let D be a set of transactions or database. Each transaction t Є D is an item set such that t is a proper subset of I. A transaction t supports X, a set of items in I, if X is a proper subset of t. Assume that the items in a transaction or an item set are sorted in lexicographic order. An association rule is an implication of the form X□Y, where X and Y are subsets of I and X∩Y= Ø. The support of rule X□Y can be computed by the following equation: Support (X□Y) = |X□Y| / |D|, where |X□Y| denotes the number of transactions in the database that contains the item set XY, and |D| denotes the number of the transactions in the database D. The confidence of rule is calculated by following equation: Confidence (X□Y) = |X□Y|/|X|, where |X| is number of transactions in database D that contains itemset X. A rule X□Y is strong if support (X□Y) ≥ min_support and confidence(X□Y) ≥ min_confidence, where min_support and min_confidence are two given minimum thresholds.

Association rule mining algorithms scan the database of transactions and calculate the support and confidence of the rules and retrieve only those rules having support and confidence higher than the user specified minimum support and confidence threshold. Association rule hiding algorithms prevents the sensitive rules from being disclosed. The problem can be stated as follows: "Given a transactional database D, minimum confidence, minimum support and a set R of rules mined from database D. A subset RH of R is denoted as set of sensitive association rules which are to be hidden. The objective is to transform D into a database D‟ in such a way that no association rule in RH will be mined and all non sensitive rules in R could still be mined from D‟.

Privacy preserving association rule mining should achieve one of the following goals:
(1) In sanitized database, all the sensitive association rules must be hidden
(2) The rules which are not specified as sensitive can be mined from sanitized database.
First goal considers privacy issue. Second goal is related to the usefulness of sanitized data set .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.

The strategies of Association rule hiding are classified as follows:

A. *Heuristic based approach*

This approach hides sensitive association rules by using two methods.

a) *Data Distortion based technique*

M. Attallah et al. [3] were first use this technique for hiding association rules, they also gave the proof of NP hardness of optimal sanitization. In this technique rules are hiding by modifying database matrix by changing the value of some items in database matrix by 0 to 1 or vice versa. The data distortion technique contains two basic methods for hiding rules. In first method, rules are hide by decreasing the support of the rule up to an acceptable level and in second method, the confidence of the rule is reduced up to certain threshold.
Verykios et al. [4] proposed five different algorithms for hiding association rules. Three of them based on reduce support and remaining two are based on reducing confidence up to an acceptable level.

b) *Data blocking based techniques*

Y. Saygin et al. [5] and [6] have proposed algorithm for hiding sensitive association rules based on data blocking technique. In this technique, rules are hide by changing the value of some item in database matrix from 0 or 1 to ?(unknown).So, the support of certain items goes down to certain level and rule mining algorithm not able to mine the sensitive rules.

B. Border based Approach

The border based approach hide sensitive association rules by modifying the border in the lattice of frequent and infrequent item sets of the original database. The item set between frequent and infrequent items make the border. The border consist the item sets which separate the frequent item set from infrequent item set. Sun and Yu [7] were first who introduce the concept of border.

C. Reconstruction based Approach

In this approach first frequent item set is extracted from non frequent item set and privacy protected data is released. The new released data is then reconstructed from the sanitized knowledge base. This approach, first perform data perturbing and then reconstruct the database. Basically this approach reconstructs the database in a manner that all sensitive information has been hidden. This method cannot guarantee to find a consistent one within a polynomial time.

D. Exact Approach

This is a non heuristic algorithm which formulates the rule hiding problem in to constraint satisfaction problem or optimal problem which is solved by integer programming. Divanis and Verykios [8] were first who used the exact approach for hiding rules. It provides an optimal solution for the problem of association rule hiding.

E. Cryptographic approach

This technique used in multi-party computation where data is distributed in different location. The database owner may want to share their data, but at the same time they want to ensure their privacy at their end. Cryptographic approach can be categorized in two categories vertical partitioned distributed data and horizontal partitioned distributed data.
In horizontal partitioning different rows are placed in different tables that are distributed in different locations. In vertical partitioning some columns kept in one table and remaining column kept in other tables.

III. MISL- PROPOSED ALGORITHM

For hiding an association rule like P ->Q, we decrease its confidence to smaller than specified minimum confidence (MCT).In the most sensitive transactions, we increase the support of P (L.H.S. of the rule). We keep one item from selected transaction by changing from 0 to 1 for increasing the support count of an item.

A. Framework of MISL Algorithm

Some important concepts used in proposed framework of MISL algorithm is 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 MISL algorithm is shown in Figure.1

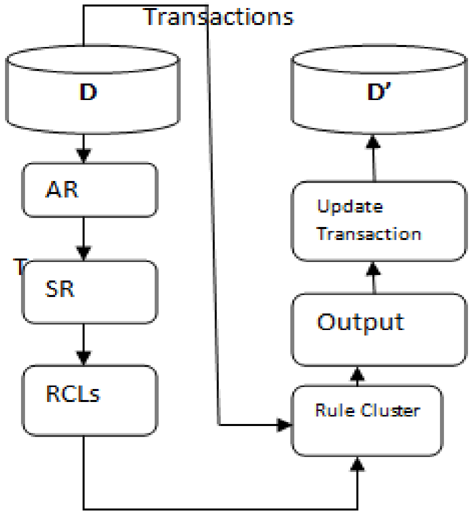

Fig 1 : Framework of proposed MISL Algorithm

By using association rule mining algorithms association rules (AR) are mined from the source database D e.g. Apriori algorithm in [2]. Based on common L.H.S. item of the rules, the sensitive rules (SR) specified and clustered selected rules are clustered based on common L.H.S. item of the rules. Rule-clusters are denoted as RC. 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.

F. MISL Algorithm

According to above presented framework for hiding association rules in database, the proposed MISL 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).
OUTPUT: The sanitized database D'.
Begin
Find Association rules.
Select the sensitive set RH.
Based on common item in L.H.S. of the selected association rule clustering is done.
Find sensitivity of each item in each cluster.
Find the sensitivity of each rule in each cluster.
Find the sensitivity of each cluster
Index the sensitive transactions for each cluster.
Sort generated clusters in decreasing order of their sensitivity.
For the first cluster, sort selected transaction in decreasing order of their sensitivity
For each cluster $c_i$ C
{
    While (all the sensitive rules c are not hidden)
    {
        Take first transaction for cluster c.
        put common L.H.S. item into the transaction.
        Update the sensitivity of new item for modified transaction in other cluster and sort it.
        For i = 1 to no. of rule $Rh_{ic}$
        {
            Update support and confidence of the rule $r_{ic}$.
            If(support of r < MST or confidence of r <MCT)
            {
                Remove Rule r from Rh
            }
        }
        Take next transaction.
    }
    End while
}
End for
Update the modified transactions in D.
End
```

Fig 2. Algorithm

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 DISL 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. XAMPLE

The following example illustrates proposed ISLRC algorithm. A sample transaction database D is shown in Table 1.

Table I : sample transaction database D

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

Table II : Frequent Item sets with Support Count

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

In Table 1TID 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 item sets satisfying MST generated from sample database D.

Table III : Item Sensitivity

Item	Sensitivity
B	2
A	1
C	1
Total sensitivity	4

Item	Sensitivity
A	2
B	1
C	1
Total sensitivity	4

Table V : Clusters generated by MISL

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

Cluster-2(b) **(bıa, bıc)**	
TID	**Sensitivity**
1	4
2	2
3	4
4	1
5	3
6	4
7	0

In following, the possible number of association rules Satisfying MST and MCT, generated by Apriori algorithm [2]: aıb, bıa, aıc, cıa, bıc, abıc, bıac, acıb, bcıa, Suppose the rules aıb, aıc, bıa and bıc 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 V, Algorithm generates two clusters based on common L.H.S. item of the selected rules.

Table VII : 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). Table5 (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, MISL provides database quality while preserving privacy.

V. RESULT

We can see that simple by MISL 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 bıc, bıac, and remaining seven rules are not hidden. Table 6 Transaction changed by ISL. But by using MISL algorithm we hide the four rules aıc, bıc, abıc and bıac in first iteration. And only five rules are left. That we can also hide by next iteration of MISL algorithm.

.

VI. CONCLUSION AND FUTURE SCOPE

Association rule hiding is a technique for hiding sensitive information in database. It is one of the techniques used in PPDM. In this paper, the various techniques of association rule hiding have been discussed. The comparative study, including advantages and limitations, of each technique also has been reviewed. we proposed a heuristic algorithm named MISL which hides many sensitive association rules at a time while maintaining database quality. 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

[11] V.S. Verykios, A.K. Elmagarmid, E. Bertino, Y.Saygin, and E. Dasseni, "Association rule hiding," IEEE Transactions on Knowledge and Data Engineering, vol16(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

Author Details

Neelkamal Upadhyay, pursuing M.Tech in Computer Science Engineering, Computer Science Engineering Department, SISTec, Bhopal, M.P., India, 09424747687.

KuldeepTripathi, Assistant Professor, Computer Science Engineering Department, SISTec, Bhopal, M.P., India, 08889951188.

Ashish Mishra, Head of Department, Computer Science Engineering Department, SISTec, Bhopal, M.P., India, 09425376828,. he has published various research papers and attended international and national conferences.