ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 3, March 2015

Privacy Preserving Association Rule Mining in Retail Industries

Mr. Yewale Aniket J.[1], Mr. Rajput Virendra D.[2], Mr. Shirapure Sagar B.[3], Mr. Patel Hardik K[4]

BE Scholar, Department of Computer Engineering, S.R.E.S.'s College of Engineering, Kopargaon, India[1,2,3,4]

Abstract: With data mining techniques one can easily disclose other's sensitive information or knowledge. So, preserving privacy for sensitive knowledge has become an important aspect. Privacy preserving data mining (PPDM) is a novel research direction to preserve privacy for sensitive knowledge from disclosure. DSRRC (Decrease Support of R.H.S. item of Rule Clusters) algorithm is used to preserve privacy for sensitive association rules in database. This algorithm clusters the sensitive association rules based on certain criteria by modifying fewer transactions and hides many rules at a time. Moreover it provides privacy for sensitive rules at certain level while ensuring data quality.

Keywords: Association Rules, Support, Confidence, DSRRC, Privacy preserving, Data Mining, Clustering, Sensitivity, Sanitized Database

I. INTRODUCTION

The concept of data mining is widely used in all sectors like in government sector and in corporate sector. 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 others 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.

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.

II. RELATED WORK

C N Modi [1] proposed a heuristic algorithm named DSRRC in "Maintaining privacy and data quality in privacy preserving association rule mining".

These algorithms maintain privacy and quality of database. These algorithms used to improve the quality of database.

Dr. K. Duraiswamy [5] in "Advanced Approach in Sensitive Rule Hiding" proposed an algorithm ISSRH (Increase Support Sensitive Rule Hiding) to hide the sensitive rules that contain sensitive items, so that sensitive rules containing specified sensitive items on the right hand side of the rule cannot be inferred through association rule mining.

V K S K Sai Vadapalli & G Loshma[2] in their paper had surveyed existing approaches regarding knowledge hiding problem in context of association rule mining by their performance and limitations.

They have analyzed security and privacy of it against involving sites or adversary, which provides certain level of privacy and security under some other security assumption.

Ahmed Haj Yasien [3] in their thesis proposes solution on privacy preserving data mining problems.

The goal of association rule mining is to find all patterns based on some hard thresholds, such as the minimum support and the minimum confidence.

The owners of these databases might need to hide some patterns that are of a sensitive nature. The sensitivity and the degree of sensitivity are decided by experts with help from the data owners.

III. MOTIVATION

Consider, two biscuit companies, A and B, and granting them to access our customer database. Now, suppose company B misuse the database and mines association rules related to company A, saying that most of the customers who buy milk also buys A's biscuit. Company B now runs a coupon scheme that offers some discount on milk with purchase of B's biscuit.

So, the amount of sales on A is down rapidly and business of company A goes down. So, releasing database with sensitive knowledge is bad for us. This scenario leads to the research of sensitive knowledge or rule hiding in database.

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

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 3, March 2015

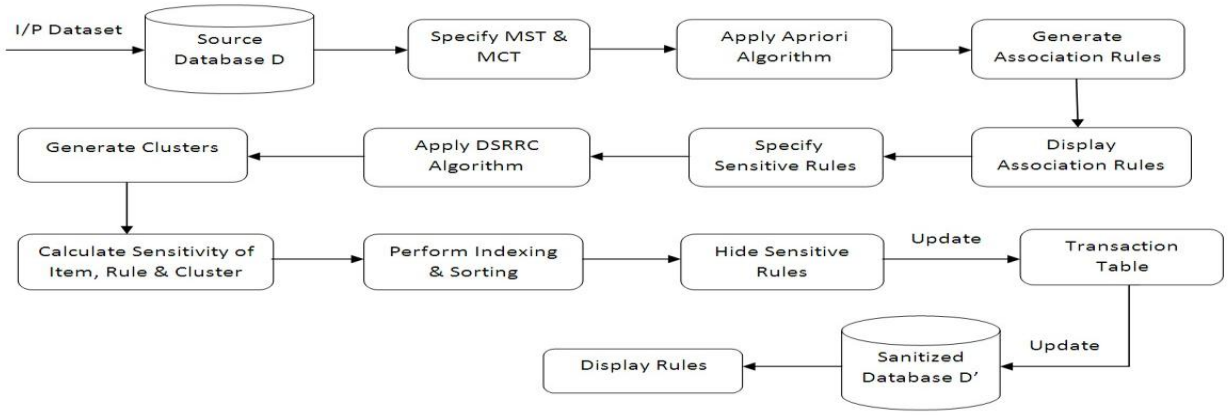

Figure 1. System Flow Diagram

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 produce some other knowledge. Association Rule hiding is the process of hiding strong association rules and creating sanitized database from the original database in order to prevent unauthorized party to
generating frequent sensitive patterns.

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 [1].

DSRRC (Decrease Support of R.H.S. item of Rule Clusters) algorithm is to preserve privacy for sensitive association rules in database. This algorithm modifies fewer transactions and hides many rules at a time. It maintains data quality in sanitized database [1].

IV. PROPOSED WORK

In this paper, we implement four different modules which together satisfy the intention of user. These modules are 1.Registration Module, 2.Rule Mining System, 3.Rule Hiding System, 4.Result Generation. These are explained as follows:

A. Registration Module

This module consists of authentication process of admin. The admin will first register to the system. After registration, he can login into the system. So unauthorized person will not get access to the system and misuse the database. Admin will add input dataset i.e. transaction database as a input to the system. Transaction Database can be from any retail industries, database oriented sales, malls or supermarkets.

B. Rule Mining System

From the given input dataset (source database) i.e. transaction database, association rules are generated (mined) by using association rule mining algorithm i.e. Apriori algorithm.

Apriori Algorithm

Apriori algorithm is one of the first algorithms to evolve for frequent itemset and association rule mining. Apriori is iterative approach that uses level wise search. In each level it uses k frequent item sets to explore k+1 frequent item sets. Two major steps of the Apriori algorithm are the join and prune steps. The join step is used to construct new candidate sets. A candidate itemset is basically an itemset that could either be frequent or infrequent with respect to the support threshold. Higher level candidate itemsets (Ci) are generated by joining previous level frequent itemsets are Li-1 with itself. The prune step helps in filtering out candidate itemsets whose subsets (prior level) are not frequent. Thus a candidate item set which is composed of one or more infrequent item sets of a prior level is filtered(pruned) from the process of frequent itemset and association mining [2].

Algorithmic Steps for Apriori:

1. Scan the entire transaction database to get the support S of each 1-itemset.
2. Compare the support S of each itemset with minimum support and generate frequent 1-itemsets, L1.
3. Use Lk-1, join Lk-1 to generate a set of candidate k-itemsets.
4. Use Apriori property to prune the itemsets, which are not frequent.
5. Scan the transaction database to get the support S of each candidate k-itemset in the final set, compare support S of each item with minimum support and generate a set of frequent k-itemsets, Lk.
6. Check whether Candidate item set is null. If no, goto step 3.
7. For each frequent item set I, generate all nonempty subsets of I.
8. For every nonempty subset s of I, output the rule s $\Rightarrow$ (I-s), if its confidence C > minimum confidence.

C. Rule Hiding System

Now, we have to preserve privacy for sensitive association rules in the database. DSRRC (Decrease Support of R.H.S. item of Rule Clusters) algorithm is used to preserve privacy for sensitive association rules in database. DSRRC

DOI 10.17148/IJARCCE.2015.4320

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 3, March 2015

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 [1].

The framework of DSRRC algorithm is shown in Fig. 2. After generating association rules (AR) from source database D, admin will specify sensitive rules (SR) from generated association rules. Rules with only single R.H.S. item are specified as sensitive. These specified sensitive rules are to be hidden in a sanitized database. Selected rules are clustered based on common R.H.S. item of the rules. Rule-clusters are denoted as RCLs. Sensitivity of each cluster is calculated.

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 Now, the rule hiding (RH) process tries to hide all the sensitive rules by deleting common R.H.S. item of the rules in cluster, from the sensitive transactions. Hiding process starts from highest sensitive transaction and continues until all the sensitive rules in all clusters are not hidden [2]. Finally modified transactions are updated in original database and produced database is called sanitized database D', which ensures certain privacy for specified rules and maintains data quality.

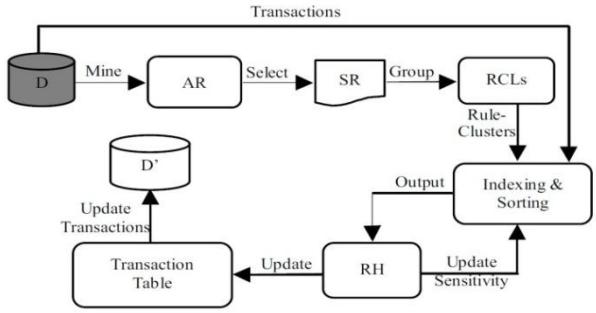

Figure 2. Framework of DSRRC Algorithm

DSRRC Algorithm

In DSRRC (Decrease Support of RHS item of Rule Cluster) algorithm, we specify sensitive rules from the generated association rules of Apriori algorithm. Rules with only single R.H.S item are specified as sensitive. DSRRC algorithm clusters sensitive rules based on common R.H.S item and calculates sensitivity of each item, rule and cluster.

Then it index sensitive transactions for each cluster and then sorts the clusters by decreasing order of their sensitivities. The highest sensitive item is hidden in the sanitized database.

Algorithmic Steps for DSRRC:

1. Begin
2. Generate association rules.
3. Selecting the Sensitive rule set RH with single antecedent and consequent e.g. $x \Rightarrow y$
4. Clustering-based on common item in R.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 ϵ C
12. {
13. While(all the sensitive rules ϵ c are not hidden)
14. {
15. Take first transaction for cluster c.
16. Delete common R.H.S. item from the transaction.
17. Update the sensitivity of deleted item for modified transaction in other cluster and sort it.
18. For i = 1 to no. of rule Rh ϵ c
19. {
20. Update support and confidence of the rule r ϵ 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.

D. Result Generation

Result generation module consists of User Interface (GUI) to navigate and use software functions. It will display the transaction database, association rules generated, hidden rules, transaction tables and rules, sanitized database, etc.

V. EXAMPLE

The following example are the transactions from a supermarket.

Table 1. Sample transaction Database D

TID	List of Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Beer, Diaper, Milk, Coke
4	Diaper, Beer, Bread, Milk
5	Bread, Milk, Coke, Diaper

Suppose, MST = 60% & MCT = 50% . We apply Apriori algorithm to generate frequent itemsets satisfying MST and MCT.

Table 2. Frequent Itemsets satisfying MST

DOI 10.17148/IJARCCE.2015.4320

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 3, March 2015

Frequent Itemsets
Bread
Milk
Diaper
Beer
Bread $\Rightarrow$ Milk
Bread $\Rightarrow$ Diaper
Milk $\Rightarrow$ Diaper
Diaper $\Rightarrow$ Beer

Table 3 . Strong association rules from the frequent itemsets

We generate strong association rules from the frequent itemsets, satisfying both MST and MCT.

Frequent Itemsets	Support	Confidence
Bread,Milk	60%	75%
Bread,Diaper	60%	75%
Milk,Diaper	60%	75%
Diaper,Beer	60%	75%

As, Bread $\Rightarrow$ Milk, Bread $\Rightarrow$Diaper, Milk $\Rightarrow$ Diaper, Diaper $\Rightarrow$ Beer satisfies given MST & MCT, so they are strong association rules. Suppose, Bread $\Rightarrow$ Diaper, Milk $\Rightarrow$ Diaper, and Diaper $\Rightarrow$Beer are specified as as sensitive rules.

Table 3. Clusters generated by DSRRC

Cluster-1 (Diaper)
Bread $\Rightarrow$Diaper,
Milk $\Rightarrow$Diaper

Item	Sensitivity
Bread	1
Diaper	2
Milk	1
Total Sensitivity	4

Cluster-2 (Beer)
Diaper $\Rightarrow$ Beer

Item	Sensitivity
Diaper	1
Beer	1
Total Sensitivity	2

After calculating sensitivity, we will perform indexing and sorting on selected sensitive rules and the item with highest sensitivity is hidden in sanitized database.

Table 4. Sanitized database D1

TID	List of Items
1	Bread, Milk
2	Bread, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Table 5. Final Sanitized database D'

TID	List of Items
1	Bread, Milk
2	Bread, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

VI. EXPERIMENTAL RESULTS AND ANALYSIS

For determining privacy, we compared the number of frequent itemsets generated from the original dataset and the sanitized dataset. We recorded the number of frequent itemsets generated from the original dataset and further applied DSRRC algorithm to generate sanitized dataset. Then we applied the same sanitized dataset as a input to our system.

As shown in Figure 3, our experiments and analysis showed that, the number of frequent itemsets generated in sanitized dataset is less than that which were generated in original dataset. So if any one who mines the sanitized dataset, will not get the proper result in terms of

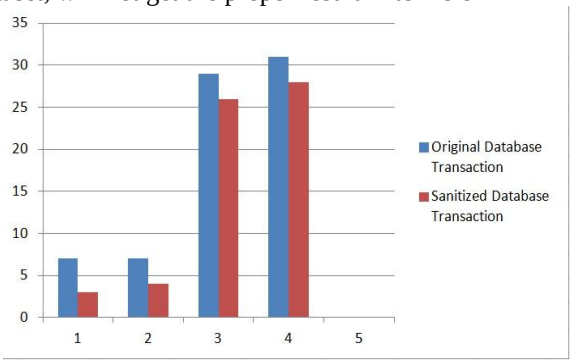

Figure 3. Comparision of frequent itemsets of original and sanitized datasets

frequent itemsets and strong association rules. So our aim is satisfied, and we are successful in maintaining privacy of association rules along with database quality.

VII. CONCLUSION

Privacy preserving association rule mining is a new body of research focusing on the security and privacy implications originating from the applications of data mining algorithms to large public databases. Apriori algorithm is implemented in order to generate frequent itemsets and thus strong association rules from the input transactional database. DSRRC algorithm is implemented for sensitive rule hiding. DSRRC algorithm hides many sensitive association rules at a time while maintaining database quality. Performance of the DSRRC algorithm is better than other existing heuristic approaches.DSRRC algorithm hides only rules that contain single item on R.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 R.H.S. items. The communication and computation cost are also reasonable for small databases which contain less number of items.

DOI 10.17148/IJARCCE.2015.4320

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 3, March 2015

REFERENCES

[1] Chirag N. Modi, Udai Pratap Rao and Dhiren R. Patel "Maintaining Privacy and Data Quality in Privacy Preserving Association Rule Mining".2010 Second International conference on Computing, Communication and Networking Technologies, pp. 1–6, Jul. 2010.

[2] V K S K Sai Vadapalli & G Loshma "Secure Strategy for Privacy Preserving Association Rule Mining". International Conference on Computer Science and Engineering, April 28th, 2012, Vizag.

[3] Adsure Sharad S, Prof. S .Pratap Singh "Preserving Data Privacy By Susceptible Association Rule Hiding Approach". International Journal of Computer Engineering and Applications, Volume VII, Issue I, July 14.

[4] Ahmed HajYasien"Preserving Privacy in Association Rule Mining". June 2007.

.[5] Dr. K. Duraiswamy, Dr. D. Manjula, N. Maheswari "Advanced Approach in Sensitive Rule Hiding". CCSE, Vol. 3, No. 2, February 2009.

DOI 10.17148/IJARCCE.2015.4320