IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-12, Dec 2018 E-ISSN: 2347-2693

Privacy Preservation for Association Rule Mining

N. S. Mrudula Jyothi[1*], A. Suraj Kumar[2]

[1,2]Department of CSE, Sanketika Vidya Parishad Engineering College, Andhra University, Andhra Pradesh, India

[*]*Corresponding Author: nsmrudula.1603@gmail.com,*

Available online at: www.ijcseonline.org

Accepted: 19/Dec/2018, Published: 31/Dec/2018

***Abstract*—** Data mining is the process of extracting hidden patterns of data. Association rule mining is an important data mining task that finds an interesting association among a large set of a data item. Association rule hiding is one of the techniques of privacy-preserving data mining to protect the association rules generated by association rule mining. In this paper, proposed a new data distortion technique for hiding sensitive association rules. Algorithms based on this technique either hide a specific rule using data alteration technique or hide the rules depending on the sensitivity of the items to be hidden. The proposed technique uses the idea of representative rules to prune the rules first and then hides the sensitive rules.

***Keywords*—**Data mining, Association rule mining, Support , Confidence

I. INTRODUCTION

Data mining is the progression of extraction of knowledge from the huge amount of databases to discover useful patterns. It has various applications such as business intelligence, medical analysis, web search, scientific exploration and more. Discovery of knowledge from a database can be expressed in patterns such as decision tree classification, clustering and association rules. Retail stores are interested in associations between different items patron placed in the shopping basket. As a result of discovering interesting association rules, an organization can identify primary patterns valuable in the forecast a new store layout, new product collections and which product to put on promotion. The knowledge discovered can be beneficial to business competitors. In order to preserve their competitive edge, some partners may hesitate to disclose sensitive information and also to prevent data mining techniques from discovering sensitive knowledge which is not even known to the database owners.

Privacy preserving data mining is a broad research area for protecting sensitive knowledge. There have been two types of privacy concerning data mining. The first type of privacy, called output privacy, is that preserving the mining output from malicious inference attacks. The second type of privacy, input privacy, is that sanitizing the raw data itself before performing mining.

In market basket databases, the sensitive association rule must be hidden for ability. Several algorithms are proposed for hiding sensitive association rules. These algorithms are given some set of sensitive items manually. The market basket database contains an outsized variety of items. Choice of sensitive items manually from the market basket database takes more time.

In the proposed method, representative association rules are computed from association rules. Based on the discovered rules sensitive items are selected for privacy preserving.
The contribution of this research paper is to hiding the sensitive association rules. The author developed an algorithm for discovering frequent itemsets from these frequent items generating sensitive rules by proving minimum support and minimum confidence threshold value and efficient distortion technique is developed using python.
The next session of this paper deals with literature study, generation of sensitive rule mining, proposed methodology, Experimental Results and finally ending with a conclusion.

II. RELATED WORK

In [1] the author described discovering the frequent itemset and generating the association rules to detect sensitive items from market basket database. Here, selecting all the rules from X with having same LHS and combining RHS of selected rules are stored. The selected RHS will be considered as sensitive items.

In paper [2] discussed the hybrid algorithm with distortion technique which is based on support and confidence approach for preserving sensitive data. This leads to hiding sensitive rules from the perspective of database owner can maintain helpful rules.

In paper [3] the author described a different approach for hiding association rule. The heuristic method to generate rules which gives privacy for sensitive data while ensuring data quality and hides as many as possible rules at a time modifying some transactions.

In paper [4] discusses the improved apriori algorithm which mines frequent item set without generation of the new candidate. So, it reduces the querying frequencies and storage resources.

In paper [5] author discussed the concise outline of Apriori calculation and late upgrades were done in the range of Apriori calculation With the learning on different enhanced calculations, it is presumed that the real absorption is to produce fewer applicant sets which contains visit things inside a sane measure of time.

III. ASSOCIATION RULE MINING

Association rule mining was first introduced by Agrawal et al in 1993[6].

Assume that set of items I= $\{i_1,i_2,\ldots i_n\}$. Let D be the transactional database. Every transaction T ϵ D is an item set such that t is a proper subset of I. An association rule is an implication expression of the form P→Q, where P and Q are disjoint item sets, P∩Q =Ø. Association rule mining can be computed by using two basic parameters Support and Confidence.

Definition of support and confidence as follows:
Support is a percentage of transactions in database D that contains both P and Q (i.e., PUQ).

$$Support(P \rightarrow Q) = \frac{|PUQ|}{|N|} \quad (1)$$

Where, |PUQ| is the number of transactions containing both P and Q in database D. |N| denotes the number of transactions in the database D.
Confidence is the percentage of transactions in database D, containing P and also contains Q.

$$Confidence\ (P \rightarrow Q) = \frac{|PUQ|}{|P|} \quad (2)$$

Where, |P| denotes the number of transactions P in the database D. |PUQ| denotes the number of the transactions containing both items P and Q in the database D. The predicament of the mining association rule is to discover all rules that are greater than the addict-specified minimum support and minimum confidence.

As an illustration, considering a database having six transactions shown in table 1 and the items in the Transactional data bitmap or data matrix in the form of binary representation shown in table 2.

Table 1. Original data

TID	ITEMS
T_1	{XYZ}
T_2	{XYZ}
T_3	{XYZ}
T_4	{XY}
T_5	{X}
T_6	{XZ}

Table 2. Transactional data bit map or matrix

TID	X	Y	Z
T_1	1	1	1
T_2	1	1	1
T_3	1	1	1
T_4	1	1	0
T_5	1	0	0
T_6	1	0	1

Mining Association rule is used to discover the most frequent items in a transactional database using apriori algorithm based on their support. Minimum support threshold (MST) parameter plays a major role to discover most frequent item sets. There are two ways to assign the MST by user-specified or automatically. Next, to discover the association n between the items sets. This is called an association rule. Many association rules are generated from the transactional database. Out of these, some rules satisfy the condition of minimum support and minimum confidence are called sensitive rules.

By considering, the above table 1 discovering sensitive association rules by user-specified minimum support of 40% and a minimum confidence of 70%, the generated rules as shown in table 3.

Table 3. Sensitive Association Rules

Association Rules	Minimum Support (40 %)	Minimum confidence (70%)
Y→X	66	100
Z→X	66	100
Y→Z	50	75
Z→Y	50	75
XY→Z	50	75
XZ→Y	50	75
YZ→X	50	100
Z→XY	50	75
Y→XZ	50	75

IV. METHODOLOGY

In this paper, we provide confidentiality to protect sensitive data from being accessed by other users. The drawback of the existing system is overcome by hiding a set of association rules that are generated. Initially the database is considered and the transactions are marked in a transactional matrix where each row represents a new transaction and column represents the item sets involved in the transaction. Then threshold values are given by the administrator for confidence and support. The transactions which satisfy this

condition are considered and hidden by modifying the values in the matrix. Then the resultant matrix contains the modified values. Hence by mining, this matrix one cannot get the exact association rules used by the firm, thereby providing privacy to the data. The proposed framework of hiding association rules was shown in figure1.

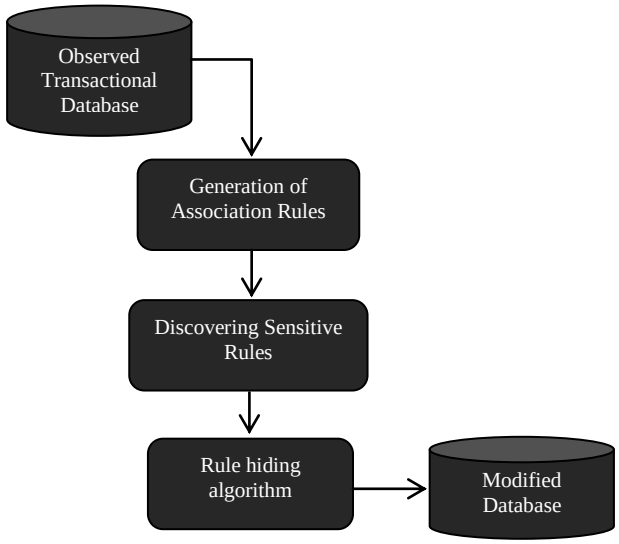

Figure1. Framework for Hiding Sensitive Association Rules

In this technique initially identifies the sensitive transactions, items, number of changes and cost of the items are initialized. Now calculate the overall cost of the individual transactions with respect to items. The sum of costs of the distinct transaction is the overall cost of the transactional database. Select the sensitive items one by one from transforming the value from 1 to 0. After all, changes made to the sensitive items finally calculate the sum of new costs of distinct transactions and the result is a modified transactional database. The flow of the process was shown in figure 2.

Algorithm for hiding sensitive items from association rules

Input:

(1) Database D

(2) Minimum support: min_supp

(3) Minimum confidence: min_conf

Output:

Sensitive item(s) for Association rule hiding.

Algorithm:

1. Find all items sets from D
2. For each sensitive item sϵS {
3. If s is not a large item set then S=S-{s};
4. If S is empty then quit;
5. Select all the rules with min_support and min_confidence and store in A
6. Select a rule r from A
7. Change the RHS of the rule r in the transactional matrix
8. Finally store the modified transactional matrix M.

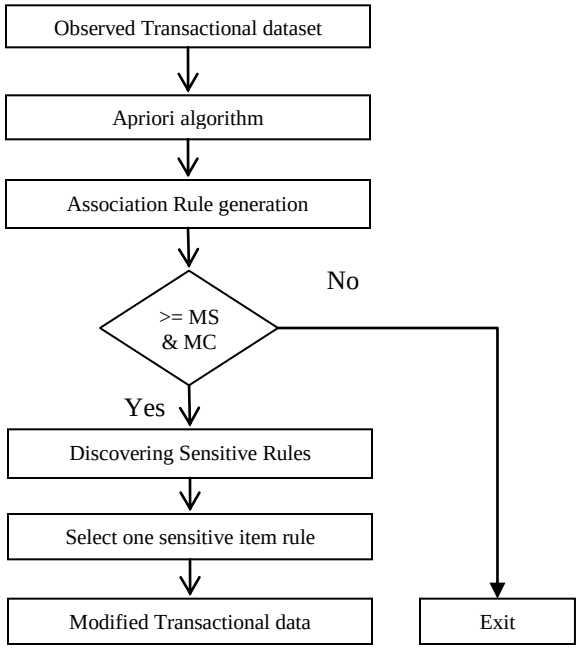

Figure 2. Flow Chart for the System Process

V. RESULTS AND DISCUSSION

Experiments are conducted on Intel Core i3 processor with 2GB RAM using Windows operating system and Python 3.6. Table 4 shows observed transactional database contains 5 items and 8 transactions. The input to the system is transactional database shown in table 5 and it produces sensitive association rules.

Table 4. Observed Transactional database

S.no	Items	Tid
1	Milk	1, 2, 3, 5, 6, 8
2	Vegetables	1, 2
3	Yogurt	1, 2, 4, 7
4	Brown_Bread	1, 2, 3,4,5,7, 8
5	Eggs	2, 4, 5, 6, 7, 8

Table 5. Transactional Data with Bit Map Representation

Transaction	Milk	Vegetables	Yogurt	Brown_Bread	Eggs
T1	1	1	1	1	0
T2	1	1	1	1	1
T3	1	0	0	1	0
T4	0	0	1	1	1
T5	1	0	0	1	1
T6	1	1	0	0	1

T7	0	0	1	1	1
T8	1	0	0	1	1

Frequent items are selected based on the support of individual items. This is shown in Table 6.

Table 6. Frequent Items

Items	Tid
Milk	6
Vegetables	3
Yogurt	4
Brown_Bread	7
Eggs	6

For the dataset given in table 5 frequent items are generated by apriori algorithm with the threshold value is 3. Consider minimum support is taken as 50% and a minimum confidence is taken as 80% then the following rules are as follows in Table 7

Table 7. Sensitive Association Rule

S.No	Association Rules	Min_Support	Min_ Confidence
1	Milk→Brown_Bread	62	83
2	Yogurt→Brown_Bread	50	100
3	Eggs →Brown_bread	62	83

The items identified in the sensitive association rules are Milk, Yogurt and Brown_Bread from Table 7. From this rule, two i.e., {Yogurt→Brown_Bread}, support of yogurt is 4 and the brown_bread is 7. Based on rule 2 modifying the observed transaction matrix of Table 5 with the presence of brown_bread with yogurt, converting of the presence of brown_bread to absence (i.e., changing 1→0). The modified transaction matrix is shown in table 8. Hence rule two is considered as an association rule, a sensitive item of yogurt is detected.

Table 8: Modified Transaction Matrix

Transaction	Milk	Vegetables	Yogurt	Brown_Bread	Eggs
T1	1	1	1	0	0
T2	1	1	1	0	1
T3	1	0	0	1	0
T4	0	0	1	0	1
T5	1	0	0	1	1
T6	1	0	0	0	1
T7	0	0	1	0	1
T8	1	0	0	1	1

Hence, for the given transactional matrix dataset yogurt and brown_bread are detected as sensitive items. A number of sensitive items can be selected by changing minimum support and minimum confidence of frequent items.

The original Transactional database is represented in bitmap representation it can be seen in table 5. Finding frequent items of one item, two items and three items and so on until no frequent items occurred and generate association rules. Now, discover the sensitive items based on the condition of the min_support and min_confidence value. From sensitive rules manually select one item rule and distorted in the item in the bitmap and store in modified transaction bitmap shown in table 8.

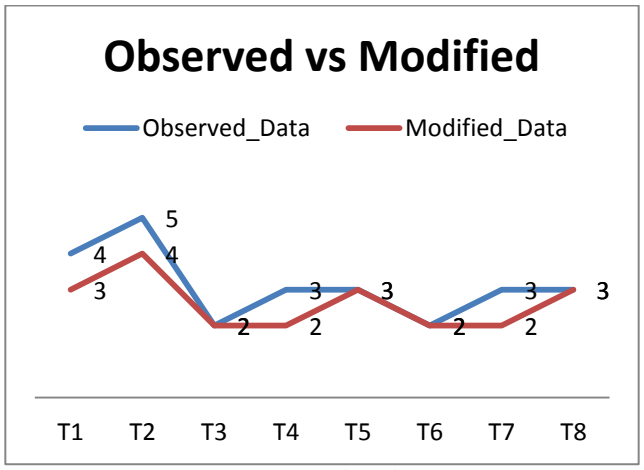

Figure 8. Observed vs Modified of Transaction data

The observed transaction data count and the modified transaction data count is the Figure3. The modified transaction data is stored.

VI. CONCLUSION AND FUTURE SCOPE

In this paper, the transactions are considered in a matrix and by applying threshold values of support and confidence, specific association rules satisfying the condition are chosen for performing association rule hiding. For those rules using distortion technique the values in the transaction matrix are changed accordingly and are stored in the modified transactional matrix. By doing so the original rules are hidden from other users and cannot be accessed by them. Hence if anyone mines the database, one cannot know the exact association rules being used by that firm. Hence in this way by implementing association rule hiding, the confidentiality of the transactions is preserved without being accessed by other users or without losing any data. By implementing association rule hiding one can preserve the pattern of their transactions and hence achieve confidentiality of the data.

REFERENCES

[1] S.Kasthuri and T.Meyyappan, " Detection of Sensitive Items in Market Basket Database using Association Rule Mining for Privacy Preserving", In the Proceeding of 2015, international conference on Pattern Recognition, Informatics adn Mobile Engineering, Feb 21-22, pp.200-203, 2015.

[2] S. Choudhary and A. Upadhyay, " Hiding Sensitive Data Item Using Association Rule Mining", International Journal of Engineering Sciences & Management, Vol.6, Issue.1,pp.13-21, 2016.
[3] D.C. Kalariya, V.Shah and J.Vala, " Association Rule Hiding based on Heuristic Approach by Deleting Item at R.H.S side of Sensitive Rule", International Journal of Computere Application, Vol.122, No 8, pp.25-28, 2015.
[4] P.Madhave, M.Mane adn S. Patil," Data mining using Association rule based on APPIORI algorithm and improved approach with illustration", International Journal of Latest Trends in Engineering and Technology,Vol.3, Issue.2, pp.107-113, 2013.
[5] M.Shridhar, M. Parmar,"Survey on Association Rule Mining and Its Approaches", International Journal of Computer Science and Engineering, Vol.5, Issue.3, pp.129-135, 2017.
[6] R. Solanki, "Principle of Data Mining", McGraw-Hill Publication, India, pp. 386-398, 1998.

Authors Profile

Miss. N.S.Mrudula Jyothi pursed Bachelor of Information Technology from University of JNTU-Kakinada, in 2014. She is currently pursuing Master in Computer Science and Engineering in Sanketika Vidya Parishad Engineering College, Visakhapatnam. Her main research work focuses on data mining.

Mr. A. Suraj Kumar pursued M.Tech(CST) from Andhra University,Visakhapatnam in 2010. He is currently an Associate Professor in Dept Of Computer Science Engineering,SVP Engineering College,Visakhapatnam. He is a Certified Ethical Hacker CEH v10, EC Council, USA. He is a life member of ISTE. He has published 6 research papers in reputed National Journals. His research interests include CyberSecurity,Data Mining,Image Processing and Computer Networking. He has 10 years of teaching experience and 2 years of industry experience. Acting as Remote Center Co-ordinator as part of NMEICT Project, MHRD Govt Of India for SVP Engg College, Visakhapatnam from August 2013 till date.

