Hiding Sensitive Association Rule Constructed from Table

[1] Ruchi . P. Kanekar, [2]Professor Rachel Dhanaraj
[1][2] Goa College Of Engineering

Abstract— Data mining is a technique which is used by every fields and also gains lots of progress in organization. Data Mining secure the sensitive data in table by using data mining techniques. This paper proposes a methodology which hides the sensitive rules by using association rule hiding techniques so that they cannot be discovered through association rule mining algorithm but at the same time not losing the great benefit of association rule mining and try to hide more and more rules in tables. Data mining provides the opportunity to extract useful information from enormous amount of data.

Keywords— Association rule, Data Mining, Frequent Itemset, Knowledge Discovery in Database(KDD), Sensitive Rule.

I. INTRODUCTION

Data mining is being put into use and studied for databases, including relational databases, object-relational databases and object oriented databases, data warehouses, transactional databases, unstructured and semi- structured repositories such as the World Wide Web, advanced databases such as spatial databases, multimedia databases, time-series databases and even fact file.

Association Rules have proven to be beneficial in inducing knowledge from the dataset and helping in the crucial decision making in all the fields. However, it also possess threat to the privacy of data.

Various data mining techniques such as, decision trees, association rules, and neural networks are already proposed and become the point of attention for several years.

Consider a scenario where two companies are collaborating for business relation and in future certain misunderstanding occur between the company and they break the business relation and don't want to collaborate again. At this time one company has the database of other company it could easily retrieve the association rule and the transaction of customer purchase. To avoid the risks in future one company should hide the sensitive rule and modify the database and handover the database to other company now this time it becomes unavailable for other company to find out the sensitive rule and the items which is more frequently purchased.

Data Mining technique are used to discover hidden information from large databases. Data Mining, also popularly known as Knowledge Discovery in Databases

(KDD), refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases.

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.

II. MOTIVATION

Now a days there is high scope in machine learning and data mining and also used in ever field as well as in every business.

Association rule can be used on centralized and decentralized data. Centralized means the data is been present at one location and Decentralized means the data is been present at several different location. For business gain association rule mining is done so that the organization can get success and also avoid the leakage of sensitive rule, here a hiding approach is used . Data mining is consist of several data, which performs operation on different kinds of data it does not required a particular form of data to perform an operation ,machine learning and data mining both performs classification, clustering and association techniques on data.

III. LITERATURE SURVEY

Ila Chandrakar et al. created a Hybrid Algorithm which combines the Increase support of left hand side (ISL) algorithm and Decrease Support of Right hand side (DSR) algorithm. The main idea of author is to extract unknown rules or extract the hidden patterns and rules from the database . Increase Support of Left Hand side and Decrease Support of Right Hand Side are two different Algorithm which are used to hide the patterns or the most important rules which reveals the information . Hence therefore the combination of algorithm is known as Hybrid Algorithm . This Hybrid Algorithm which hides the item for hiding the rules therefore item which is present in right hand side and then tries to hide the rules in which item is in left hand side . The algorithm first generate all the rules then it calculate the confidence for all the rules and also select the hidden item. Then it finds out which rules that contain the sensitive item .Then it will check that the hidden item is present at Right hand side .If the confidence of rule is less then the minimum confidence than it will generate the 2-itemset else it will Decrease Support of sensitive item which is present in Right Hand Side. Then modify the transaction by replacing 0 with 1 at Right Hand Side and then remove the transaction and save the transaction .If the rules confidence which is in Left Hand Side then increase the support of Left Hand Side and then select the first transaction from various transaction and then replace 1 with 0 for the item at Left Hand Side and then remove the transaction and save the transaction . Then calculate the confidence of the rule . If the transaction is empty then items cannot be hidden and at the last modify the Database. In proposed Algorithm it shows that the efficiency with comparing with ISL and also it shows that it does less database scan and prunes more and more number of hidden rules.[5]

In[6]Khalid Iqbal et al. has hide the sensitive XML Rule by using a Bayesian Network .In this Fast Hiding Sensitive Association Rules Algorithm is consider a database .The author implemented the algorithm in two step . In first step Fast Hiding Sensitive Association Rule scan the database once and it gathers correlated information of the transaction and sensitive rules. Correlation is denoted by the Graph G . in second step transaction is been modified one after another until the entire set of sensitive rules is hidden . Very first step there is the XML document that is the original database , in this

the XML documents are in the XML formats. In step two it forms a binary table as well as the transactional item set with the help of input file or documents . Transactional Symbolized Items are a group of symbolized items that forms a transactional based on XML document items . Binary Table is the table which contains 1 and 0 which shows the presents or absences of items in the transaction . At step three , Apriori algorithm is been performed on the transactional item set to generate the association rules from Database. Apriori Algorithm is the algorithm which is most commonly used which helps to generate the XML Association rules. At step four Bayesian Network is generated with the help of K2 Algorithm and Binary table transactional Itemset, Graphical Model is generated by using the Bayesian Network and K2 Algorithm . Graphical Model also gives an idea and displays the interesting relationship among nodes in a probabilistic manner. At step five form the Conditional Probability Table which include the items and there Conditional Probability. In step six Modify the transactional itemset based on identified Mode. In last step Apply the Apriori Algorithm on the modified database and the result would be XML Association Rules .

Shan-Tai Chen et al. created an Algorithm which is known as ADSSI . Advance Decrease Support of Sensitive Items which reduce the risks of sensitive information leakage by modifying transactions in database .Advance Decrease Support of Sensitive Item Algorithm that overcomes the disadvantage of Decrease Support of Sensitive Item and also reduce the lost rules which is generated by DSSI Algorithm . It take the input as the database , minimum support , minimum confidence and a sensitive item two non-increasing data structures, the support count table and the Item table are created for efficiency and completeness of the algorithm . It also shows the table which shows the relationship and also contain the sensitive itemset , Sensitive-related frequent Itemset and Non-sensitive Frequent itemset . There are various stage which is used to performed the algorithm .In first Step it check the database and select the useful information and then create a support count table , all itemset a table and WT and then check and remove Sensitive frequent Itemset . At step two it modifies the database by using Decrease Support of Sensitive item Algorithm and then calculate weight for each transaction in WT and then modify the database to decrease Support Count. At step three Modify the whole database to release the lost rule side effects [7].

IV. PROBLEM STATEMENT

Association Rule mining is one of the most important data mining tools used in many real life applications The problem of mining association rules is to find all rules that are greater than the user-specified minimum support and minimum confidence and also the problem of efficiently hiding predictive association rules.

Make use of association hiding technique in such a way that it won't affect database and also do not loss the information.

V. RELATED THEORY

Data mining is used to build predictive and descriptive models. A predictive model is used to explicitly predict values. As an example, based on the customers who have responded to an offer, the model can predict what other customers are most likely to respond to the same offer. Descriptive models on the other hand describe patterns in existing data. It can provide valuable information, such as identifying different customer segments

ASSOCIATION RULE MINING

Association rule mining is important field to be considered under privacy preserving data mining, Agrawal et al. was first who proposed the concept of association rule mining, basically association rule is the if-then relationship among the data.[8] Consider example for better understanding the concept of association rule "If a customer buys a dozen eggs, then he is 85% likely to also purchase milk". By analysing above example it can be concluded that milk is somewhat related to egg because every time a customer buy egg, 85% of the time he/she also likely to buy milk. Initially association rule is used for market basket analysis in order to determine the relationships among the items bought by customers.

Association Rule

For a given set of transactions, an association rule is a pattern that will predict the occurrence of a data item based on the occurrences of other data items in the transactions. Association Rules have proven to be beneficial in inducing knowledge from the dataset and helping in the crucial decision making in all the fields.
Example : {Milk, Bread, Diaper}

Frequent Itemsets

An itemset is called a frequent itemset is the minimum support threshold given by users. An itemset whose support is greater than or equal to a minsup threshold.

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

Figure : Table with 5 transaction

Itemset

A collection of one or more items
Example: {Milk, Bread, Diaper}

k-itemset

An itemset that contains k items

Support (s)

Fraction of transactions that contain both X and Y
Example:

$$s = \frac{\sigma(\text{Milk, Diaper, Beer})}{|T|} = \frac{2}{5} = 0.4$$

Confidence (c)

Measures how often items in Y appear in transactions that contain X
Example:

$$c = \frac{\sigma(\text{Milk, Diaper, Beer})}{\sigma(\text{Milk, Diaper})} = \frac{2}{3} = 0.67$$

Sensitive Association Rule
In order to hide the knowledge from a database, the rules that help in discovering this knowledge have to be hided These rules that help in the discovering the knowledge and help in taking decisions are called as Sensitive Association Rule.[1]

ASSOCIATION RULE HIDING

An association rule is nothing but some frequent patterns of data items of database, which may consist of one or more than one data item on L.H.S or R.H.S. It has its associated support and confidence. If an association rule is to be hidden its support /confidence has to be reduced so that it becomes less than threshold values. In order to reduce the support or confidence of the rule its data items have to undergo changes. These data items can be either deleted, replaced or modified in the transactional database.

Association Rule Hiding is an emerging area of data mining that aims to transform an original database into a released database such that the Sensitive Association Rules, which are used to make decisions, cannot be discovered, whereas the non sensitive association rules can still be mined.
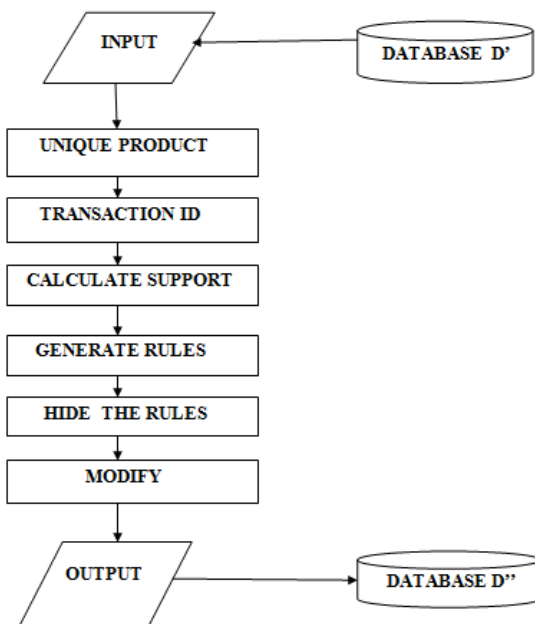
VI. FLOW CHART

Figure : Flow chart of Proposed Model

VII. DESCRIPTION

The above diagram shows the flow of the proposed system .

INPUT : Here the market basket dataset is taken as the document text is first taken as input. The input is Database D'
.

UNIQUE PRODUCT : Then it finds out the unique products from the datasets and gives the product id to all the unique product and save it into database.

TRANSACTION ID : Then it compare the product in the dataset with the product in the table in database and returns the unique product id and give the transaction id to all the transaction and display the product id instead of

product name. Here a star schema is created by assigning the transaction id to the item present in first transaction , second transaction and so on .

CALCULATE SUPPORT : Here support shows that how many time particular item has occur . Then it calculate the support for all the items, which is selected as the unique product in the table.

GENERATE RULE : Then it generate the rules , to generate rules first it find out the frequent item set and based on that association rule is been generated . To Identify the sensitive Rule minimum support and minimum confidence is been taken . The rule which has support or confidence greater is set as sensitive rules and the rules which as less support and confidence is discarded.

HIDE THE RULE : Then it hide the dummy items in the sensitive rule . Here the dummy items are the items which is not there in the dataset but present in the database. Dummy item is selected randomly from the dummy table and assign at the first position in the transaction and then the transaction is written to output file as new or modified Database D'' .

VIII. EXPERIMENTAL RESULT

Here the implementation is based on hiding sensitive rules constructed from table . In this where tables are constructed out of which the dummy table is very important and it should not be reveled to anybody. The implementation make use of market basket dataset which include multiple transaction and each transaction which includes multiple items which is purchased by the users.
1.At first the user select the dataset from the system which acts as the input files.
2. Then check the items in the transaction and find out the unique items
3.After getting the unique items the transaction in the dataset is been converted into the product id by comparing the name of product items and returns the product id of each item.
4. Then calculate the support of each items here support means the count of item which occur in the transaction.
5. Then it takes the minimum support threshold and minimum confidence threshold from the user based on this support and confidences the rules is been generated and also it takes the frequent items to generate the rules.

6. Then it hide the rules by using the constructed table , here the constructed table is known as Dummy table and the items presented in dummy table is used as hiding the Transaction rule.

7. If suppose the user want to retrieve the original dataset so it can use the same dummy table and remove the dummy table by comparing the modified database with the dummy table.

And the final result for 50 transaction will look like

Transaction	Purchased Item
1.	semi-finished, bread ,margarine ,ready soups
2.	tropical fruit ,yogurt ,coffee
3.	whole milk
4.	pip fruit ,yogurt ,cream cheese ,meat spreads
5.	other vegetables ,whole milk, condensed milk, long life bakery product
6.	whole milk ,butter ,yogurt ,rice ,abrasive cleaner
7.	rolls/buns
8.	other vegetables ,UHT-milk ,rolls/buns ,bottled beer ,liquor (appetizer)
9.	pot plants
10.	whole milk ,cereals
11.	tropical fruit ,other vegetables ,white bread ,bottled water ,chocolate
12.	**Potato** ,tropical fruit ,whole milk ,butter ,curd ,yogurt ,flour ,bottled water, dishes
13.	**Beef**
14.	Frankfurter ,rolls/buns ,soda
15.	Chicken ,tropical fruit

16.	Butter ,sugar, fruit/vegetable juice ,newspapers
17.	fruit/vegetable juice
18.	packaged fruit/vegetables
19.	Chocolate
20.	specialty bar
21.	other vegetables
22.	butter milk, pastry
23.	whole milk
24.	tropical fruit ,cream cheese ,processed cheese ,detergent ,newspapers
25.	tropical fruit ,root vegetables, other vegetables, frozen dessert ,rolls/buns, flour, sweet spreads, salty snack, waffles, candy, bathroom cleaner
26.	bottled water, canned beer
27.	Yogurt
28.	sausage, rolls/buns, soda, chocolate
29.	other vegetables
30	brown bread, soda, fruit/vegetable juice, canned beer, newspapers, shopping bags
31.	Yogurt ,beverages ,bottled water ,specialty bar
32.	hamburger meat ,other vegetables ,rolls/buns ,spices ,bottled water, hygiene articles ,napkins
33.	root vegetables ,other vegetables ,whole milk ,beverages ,sugar
34.	pork, berries, other vegetables, whole milk ,whipped/sour cream , artif.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 8, August 2017

	sweetener, soda, abrasive cleaner
35.	beef, grapes ,detergent
36.	Pastry ,soda
37.	fruit/vegetable juice
38.	canned beer
39.	root vegetables, other vegetables ,whole milk ,dessert
40.	Zwieback ,newspapers
41.	sausage, rolls/buns, soda, canned beer ,specialty bar, shopping bags
42.	**Strawberry**, tropical fruit, root vegetables ,whole milk ,yogurt, domestic eggs ,brown bread, pastry, sugar, cereals, coffee ,soda ,waffles ,candy
43.	berries, yogurt
44.	canned beer
45.	**Banana** ,butter milk ,yogurt ,cream cheese ,spread cheese ,rolls/buns ,bottled water ,soda ,newspapers
46.	Coffee
47.	pastry, bottled water
48.	rolls/buns
49.	misc. beverages
50.	root vegetables, other vegetables, butter, curd, whipped/sour cream, UHT-milk, hard cheese, rolls/buns, bottled water long life bakery product

IX. CONCLUSION

It uses the transaction of customer purchase that is market basket dataset , which comprise the items which is purchased most frequently. Generating rules gives the thought that which article is purchased collectively, so with the aid of association rule an society can keep the product mutually . Apriori is the method which is used to locate out the frequent item set and generate the association rules. This method is easy to understand and also easy to execute .By using this approach organizations and folks can effortlessly distribute their database with others without panic of sensitive information getting exposed and also the database ruins protected. With the aid of Dummy items a sensitive rule is unseen . Dummy items are the article which is kept private and cannot be distribute to others .

X. REFERENCES

1. Ms. Chaitrali Chaudhari, Ms.Speril Machado, Association Rule Hiding for Multi-Relational Database, International Journal of Computer Trends and Technology (IJCTT) – volume 30 Number 4 – December 2015.

2. Anton Flank , Multirelational Association Rule Mining, 7th September 2004.

3. Siddhrajsinh Solanki, Neha Soni, A Survey on Frequent Pattern Mining Methods Apriori, Eclat, FP growth International Journal of Computer Techniques- ISSN :2394-2231

4. Tapan Sirole, Jaytrilok Choudhary, A Survey of Various Methodologies for Hiding Sensitive Association Rules, International Journal of Computer Applications (0975 – 8887) Volume 96– No.18, June 2014

5. Ila Chandrakar, Yelipe Usha Rani, Mortha Manasa and Kondabala Renuka, Hybrid Algorithm for Privacy Preserving Association Rule Mining , Journal of Computer Science 6 (12): 1494-1498, 2010 ISSN 1549-3636.

6. Khalid Iqbal , Sohail Asghar , Simon Fong , A PPDM Model Using Bayesian Network for Hiding Sensitive XML Association Rules , 2011 IEEE .