Scientific Journal of Impact Factor(SJIF): 4.14

e-ISSN(O): 2348-4470
p-ISSN(P): 2348-6406

International Journal of Advance Engineering and Research Development

Special Issue for ICPCDECT 2016,Volume 3 Issue 1

Sensitive data hiding with the mining of association rules

Gayatri Bagul[1], Pooja Katkar[2], Nupoor Kumbhar[3], Pallavi Patil[4]

[1-4]*Computer, AISSMS IOIT,*

ABSTRACT:*In this Project we are applying a heuristic based algorithm named FADSRRC (Fast Algorithm for Decrease Support of R.H.S item of Rule Clusters) to hide the sensitive association rules from sensitive item set with multiple items in subsequent (R.H.S) and precedent (L.H.S). This algorithm overcomes the limitation of existing rule hiding algorithm DSRRC andMDSRRC. PPDM techniques are helpful to enhance the security of database. FADSRRC algorithm selects the items and transactions according to some circumstances which remodels transactions to prevent the sensitive information from getting leaked and by using FADSRRC algorithm we can hide more sensitive data from sensitive item sets. The proposed FADSRRC algorithm is highly efficient and conserves the virtue of database. And also it gives more accuracy than both the algorithm MDSRRC and DSRRC.*

Association rule hiding problem can be defined as: convert the original database into sanitized database so that data mining techniques will not be able to mine sensitive rules from the database while all non-sensitive rules remain visible. Association rule mining technique is widely used in data mining to find consociation between item sets.

Keywords: *Association rule, sensitive pattern, Privacy preserving data Mining (PPDM), DSRRC (Decrease Support of R.H.S item of Rule Clusters), MDSRRC (Modified Decrease Support of R.H.S. item of Rule Clusters),FADSRRC* (*Fast Algorithm for Decrease Support of R.H.S. item of Rule Clusters).*

I. INTRODUCTION

Association rule mining is data mining technique which finds dependency among the item sets. The issue of privacy plays important role when several organizations share their data for mutual benefit but no one wants to disclose their private data. Therefore before disclosing the database, sensitive patterns must be hidden and to solve this issue PPDM techniques are helpful to enhance the securityof database.

Many approaches are proposed for sensitive pattern in database. In paper [1], author proposed the two new algorithms for problem of exploring association rules between elements in large sales transactions. These two algorithms for solving this problem are fundamentally different from the known algorithms. They combine the algorithm into apriori algorithm which has exquisite properties, increasing performance with respect to the transaction size and number of elements in database. In paper [2], authorproposed algorithm to maintain the privacy of the sensitive items from huge dataset and avoid disclosure of sensitive information.

Let a digital store that purchase electronic items from two companies, A and B, and both can access customers' database of the store. Now A applies data mining techniques and mines association rules related to B's products. A had found that most of the customer who buy computer of the B also buy hard disk. Now A offers some discount on hard disk if customer purchases A's computer. As a result thebusiness of B goes down. So releasing the database with sensitive information cause the problem. This scenario gives the direction to research on sensitive rules (or knowledge) hiding in database.

The proposed algorithm is the improved version of DSRRC and MDSRRC. DSRRC could not mask association rules with more than single item in precedent (L.H.S) and subsequent (R.H.S.) and MDSRRC is less efficient. To overcome this drawback, we have discovered an algorithm FADSRRC which uses count of sensitive items in subsequent of the sensitive rules. It modifies the minimum number of transactions to mask maximum sensitive rules and conserves data quality.

A. PROBLEM DESCRIPTION

Association rule hiding problem is to convert the original database into sanitized database so that data mining techniques will not be able to mine sensitive rules from the database while all non-sensitive rules remain visible.

Association rule hiding techniques can be classified into heuristic based approaches, reconstruction based approaches, border based approaches, exact approaches, and cryptography based approaches. Proposed algorithm use heuristic based approach which is widely used. DSRRC could not hide sensitiveassociation rules with more than single items in precedent (L.H.S) and subsequent (R.H.S.).
FADSRRC algorithm is used. FADSRRC hides the sensitive association rules with more than single items in subsequent (R.H.S) and precedent (L.H.S). FADSRRC overcomes drawback of DSRRC. Proposed algorithm selects the items and transactions according to certain conditions which reforms transactions to hide the confidential knowledge.
FADSRRC algorithm is proposed to hide sensitive rules with multiple items in a faster and efficient manner.

II. LITERATURE REVIEW

In the existing systems, DSRRC and MDSRRC algorithms are used. Proposed algorithm select items and business operations according to their frequency count and sensitivity. Then modifies those business operations so as to hide them. Association rule hiding techniques can be categorized into heuristic based approaches, reconstruction based approaches, cryptography based approaches, exact approaches, and border based approaches. Proposed algorithm use heuristic based approach which is widely used.
Proposed System uses Apriori Algorithm for association rule learning over transactional databases, K-Means Clustering method for modelling grouping of the association rules and Binarization methods such as Data Distortion, Data Blocking.

A. APRIORI ALGORITHM

Apriori algorithm is a technique for mining the commonly appearing patterns and generating association rules over large databases. This algorithm used prior knowledge of frequent itemset properties by identifying commonly appearing individual items and extends them to larger sets till they occur very often in database.Apriori employs an iterative approach known as a level-wise search. The common item sets identified by Apriori algorithm generate association rules which focuses on creating the commonly appearing tendency in the database.
For example, assume an electronics store tracks sales data by material management module (MM module) for each item; each item, such as "computer" or "hard-disk", is identified by a numerical part code. Let the electronics store has a database of transactions that consist of following itemsets:

Table 1.

Items
{p,q,r,s}
{p,q,s}
{p,q}
{q,r,s}
{r,s}
{q,s}

We will use Apriori to determine the frequent item sets of this database. The first step of Apriori is to count up the number of occurrences, called the support, of each member item separately, by scanning the database a first time. Let the support threshold be 2.Now we obtain the following result:

Table 2. First step of Apriori Algorithm

Items	Support
{p}	3
{q}	6
{r}	4
{s}	5

All the itemssets of size 1 have a support of at least 2, so they are all frequent. The next step is to generate a list of all pairs of the frequent items:

Table 3. Second step of Apriori Algorithm

Items	Support
{p,q}	3
{p,r}	1
{p,s}	2
{q,r}	3
{q,s}	4
{r,s}	3

The pairs {p, q}, {q,r}, {q,s}, {p,s} and {r,s} all meet or exceed the minimum support of 2, so they are frequent. The pairs {p,r} is not. Now, because {p,r} is not frequent, any larger set which contains {p,r} cannot be frequent. In this way, we can prune sets: we will now look for frequent triples in the database, but we can already exclude all the triples that contain one of these two pairs:

Table 4. Final result from Apriori Algorithm

Items	Support
{p,q,s}	2
{q,r,s}	2

In the example, there are frequent triplets –{p,q,s} and {q,r,s} are satisfying the minimal threshold, and the other triplets were excluded because they were super sets of pairs that were already below the threshold.
We have thus determined the frequent sets of items in the database, and exemplified that the subset which were below the specified threshold value were not included in the commonly appearing patterns.

B. CLUSTERING

The process of grouping a set of physical or abstract objects into classes of similar objects is called clustering. Proposed System uses K-Means Clustering method.K-means clustering- vector quantization- which allows the modelling of probability density functions by the distribution of prototype vectors used for data compression. It works by partitioning large set of data points (vectors) into constellates having approximately the same number of data points nearest to the centroid where each constellate is identified by its centroid point.
The K-Means algorithm assigns each point to the cluster whose centroid (i.e. center) is nearest. The centroid is the average of all the points in the cluster- that is, its coordinates are the arithmetic mean for each dimension separately over all the points in the cluster.
The following steps are followed in K-Means Clustering

Algorithm:

1. Select k centre in the problem space (it can be random).
2. Partition the data into k clusters by grouping points that are closest to those k centers.
3. Use the mean of these k clusters to find new centers.
4. Repeat steps 2 and 3 until centers do not change.

C. BINARIZATION

Binarization is a technique used to replace the sensitive item with 1 and 0 so that other viewers do not understand about the sensitive data.There are mainly two techniques used for hiding sensitive rule: data distortion which permanently deletes some items from database and data blocking which put '?' instead of deleting items from database. Data Distortion changes the item value by a new value in database matrix. It alter '0' to '1' or '1' to '0' for selected items in selected transactions to decrease the confidence, by decreasing or increasing support of items in sensitive rules. Verykios et al. [3] proposed five different algorithms with five assumptions to hide sensitive rules in database. Three are based on reduced support of item set and two are based on reduced confidence of the sensitive rule below the minimum confidence threshold.

Heuristic algorithms cannot give an optimal solution because of side effects to non-sensitive rules. Y-H Wu et al. [4] provide approach to modify few transactions in the transactional database to decrease the support or confidence without producing more effect to sanitized database. S.L wang et al. [5] proposed two algorithms to maintain privacy .They explain two privacy techniques one is output privacy in it data is minimally altered and mining does not disclose certain privacy. And input privacy it explains that data is manipulated so that mining result do not affect the output data result.

Data Blocking: We deputise the values '1' and '0' with '?'in selected transactions which is an alternative to data distortion technique. Hence, adversary will not know theoriginal value of '?'.

Algorithm

IV. EXAMPLE

To understand FADSRRC following example is illustrated. In Table II transnational database D is shown. With 3 as MST and 40% as MCT, The possible generated association rules by Apriori algorithm:

COMP→HD,HD→COMP,COMP→PD,PD→COMP,COMP→CPU,CPU→COMP,HD→PD, PD→HD, HD→CPU, CPU→HD, PD→CPU, CPU→PD, PD→CD, CD→PD, CPU→CD ,CD→CPU ,COMP→PD CPU, PD→COMP CPU, COMP PD→CPU, CPU→COMP PD, COMP CPU→PD, PD CPU→COMP, PD→CPU CD, CPU→PD CD , PD CPU→CD, CD→PD CPU, PD CD→CPU, CPU CD→PD, COMP→HD CPU, HD→COMP CPU, COMP HD→CPU, COMP CPU→HD and HD CPU→COMP.
Let the database owner specify rule COMP→HD CPU, COMP→PD CPU and CPU→COMP PD as sensitive rules.The sensitivity of COMP=3, HD=1, PD=2, CPU=3.
Transaction with its sensitivity is shown in Table I. Now algorithm finds frequency of each item presents in R.H.S of sensitive rules. Here frequency of CPU=2, PD=2, COMP=1, HD=1.
So IS= {CPU, PD, COMP, HD}. In this example item 'CPU' is selected as IS_0. Then it sorts the transactions which supports IS_0 in descending order of their sensitivity. Then Select transaction with highest sensitivity and delete IS_0 item from that transaction. Update confidence and support of all the sensitive rules. Table III show modified database D1 after first deletion of item from first transaction.
Now update sensitivity of each item. Updated count of each item for IS is PD=2, CPU=1, COMP=1, HD=1. So updated IS={PD, CPU,COMP,HD} and IS_0 ='PD'. Sort transactions which support IS_0 and delete the IS_0 from transaction with highest sensitivity. Now all sensitive rules are hidden. Final sanitize database is shown in table IV.

TABLE 1 to 4 from above example

Table 1. Transnational database

TID	Items	Binary matrix of Items
1	COMP HD PD CPU CD	1 1 1 1 1 0 0 0
2	COMP PD CPU	1 0 1 1 0 0 0 0
3	COMP HD CPU FPY CL	1 1 0 1 0 1 1 0
4	HD PD CPU CD	0 1 1 1 1 0 0 0
5	COMP HD CPU	1 1 0 1 0 0 0 0
6	PD CPU CD FPY CR	0 0 1 1 1 1 0 1
7	COMP HD PD CL	1 1 1 0 0 0 1 0
8	COMP PD CPU CD	1 0 1 1 1 0 0 0
9	COMP PD CPU CR	1 0 1 1 0 0 0 1

Table 2: Transaction with its sensitivity

Items	Sensitivity
1	9
2	8
3	7
4	6
5	7
6	5
7	6
8	8
9	8

Table 3. Sanitized database D1

TID	Items
1	COMP HD PD CD
2	COMP PD CPU
3	COMP HD CPU FPY CL
4	HD PD CPU CD
5	COMP HD CPU
6	PD CPU CD FPY CR
7	COMP HD PD CL
8	COMP PD CPU CD
9	COMP PD CPU CR

Table IV. Final sanitized database

TID	Items
1	COMP HD PD CD
2	COMP CPU
3	COMP HD CPU FPY CL
4	HD PD CPU CD
5	COMP HD CPU
6	PD CPU CD FPY CR
7	COMP HD PD CL
8	COMP PD CPU CD
9	COMP PD CPU CR

V. CONCLUSION

We proposed an algorithm named FADSRRC which hides sensitive association rules from sensitive item sets with fewer modifications on database to maintain data quality and to reduce the side effect of database. Functionality of proposed

algorithm in database can be determined with three sensitive rules. Experimental results show that proposed algorithm works better than DSRRC and MDSRRC. In future, FADSRRC algorithm can be used on big data and can be extended to increase the capability and truncate the side effects by lessening the modifications on database.

REFERENCES

[1] Rakesh Agarwal and RamakishnanSrikant, "Fast algorithm for Mining Association Rules" IBM Almaden Research Center 650 Harry Road, SanJose, CA 95120.

[2] Rakesh Agarwal and RamakishnanSrikant,"Privacy preservation Data Mining" IBM Almaden Research Center 650 Harry Road, SanJose, CA 95120.

[3] V. S. Verykios, A. K. Elmagarmid, E. Bertino, Y. Saygin, and E.Dasseni, "Association rule hiding," IEEE Transactions on Knowledge and Data Engineering, 2004.

[4] Y.-H. Wu, C.-M. Chiang, and A. L. Chen, "Hiding sensitive association rules with limited side effects," IEEE Transactions on Knowledge and Data Engineering, 2007.

[5] S.-L. Wang, B. Parikh, and A. Jafari, "Hiding informative association rule sets, " Expert Systems with Applications, vol. 33, no. 2, pp. 316 – 323, 2007.

[6] Jaideep Vaidya and Chris Clifton Purdue, "Privacy Preservation Association Rule Mining in Vertically Partitioned data". Purdue University West Lafayette ,Indiana 47907.

[7] Murat Kantarcioglu and Chris Clifton, "Privacy-Preserving Distributed Mining of Association Rules on Horizontally Partitioned Data". Purdue University West Lafayette ,Indiana 47907.

