ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 1, January 2016

A Survey of Various Hiding Approaches

Rachana Naik[1], Pradeep Rupayla[2]

M.Tech Student, Department of Information Technology, MIT, Ujjain, India[1]

Assistant Professor, Department of Computer Science, MIT, Ujjain, India[2]

ABSTRACT: One of the great challenges of data hiding is to finding hidden patterns without revealing sensitive information. Sharing data among organizations often leads to mutual benefit. In recent years, data mining is a popular analysis tool to extract knowledge from the collection of large amount of data. Privacy preservation is a major research area for protecting sensitive data. It is an answer to such challenges. Association rule mining is one of the techniques of PPDM to protect the association rules generated by association rule mining. In this paper we describe a survey of hiding methods for privacy preservation. There are various algorithm proposed in recent years. Here, we summarize them and survey the existing techniques for association rule mining.

KEYWORDS: Data Mining, association rule mining, association rule hiding approaches, border based algorithm.

I. INTRODUCTION

There is huge amount of data being produced every data from different resources. An organization may require to release its data to public or to allow another party to access it. Some sensitive information, which is secret to the organization need to be hidden before the data is released. Data mining is the techniques that extracts knowledge from the large amount of data and enables people to efficiently extract unknown knowledge. While extracting information or knowledge from raw data, there is need for some technique that deals with security of that information. Privacy preservation is the technique that deals with the security of the information.
In [1] the problem of hiding sensitive knowledge has been considered as an important issue of PPDM. Association rule hiding is one of the privacy preservation techniques to hide sensitive association rules [2]. All hiding algorithm aims to minimally modify the original database such that no sensitive association rule is derived from it.

II. RELATED WORK

Association rule mining
Let I= {i1,……in} be a set of items. Let D be a set of transaction. Each transaction tЄD is an item set such that t is a proper subset of I. Transaction t supports X, a set of items in I, if X is a proper subset of t.
An association rule X→Y is an implication form, where X and Y are subsets of I and X∩Y=Ø. The support of rule X→Y can be computed as: Support (X→Y)= $|X \rightarrow Y| / |D|$, where $|X \rightarrow Y|$ denotes the number of transactions that contains the number of transactions that contains the item set XY, and $|D|$ denotes the number of transaction in database. The confidence of rule is calculated by: Confidence (X→Y) = $|X \rightarrow Y| / |X|$, where $|X|$ is number of transactions in D that contain item set X. A rule X→Y is strong if support (X→Y) ≥ min_support and confidence (X→Y) ≥ min_confidence, where min_support & min_confidence are two given minimum thresholds.
Association rule mining algorithm scan the database of transaction and calculate 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 algorithm prevents the sensitive rules from being disclosed. Various association rule hiding approaches are proposed to hide sensitive rules.

1. Revealing Approaches of association rule hiding
A. Heuristic Approach: This involves efficient fast & scalable algorithm that sanitize a set of transactions from the original database to hide the sensitive association rules. For this two methods are used:

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 1, January 2016

a) Data distortion technique: - This technique [3] were first use by M.Attallah rt. al for hiding association rules. In this rules are hiding by modifying the matrix in database by changing the values of some items by 0 to 1 or vice versa.

TABLE 1: Hiding A→C by distortion

A	B	C	D
1	1	0	1
1	0	1	1
0	1	0	0
0	1	0	1
1	0	1	0
1	0	1	1

→

A	B	C	D
1	1	0	1
1	0	1	1
0	1	1	0
0	1	0	1
0	0	1	0
1	0	1	1

b) Data blocking technique: - In this technique, rules are hide by changing the values from 0 or 1 to (?) unknown of some item in data base matrix. Y. Say gin et al.[4] and [5] have proposed algorithm for hiding sensitive association rules based on data blocking technique.

TABLE 2: Hiding A→D by blocking

A	B	C	D
1	0	0	1
1	1	0	1
0	0	1	1
1	1	1	0
0	0	0	0
1	1	1	1

→

A	B	C	D
1	0	0	1
1	1	0	1
0	0	?	1
?	1	1	0
0	0	0	0
1	1	1	1

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. Exact Approach: - By non heuristic algorithm which is contain by exact approach hiding process is formulate as a constraints satisfaction problem which is solved by integer programming. An optimal hiding solution got by this algorithm with no side effects. An exact algorithm is proposed in [6] by which the distance between original database and it's sanitize version. An optimal solution is achieved as compared to previous approaches in [7].

C. Border based Approaches: - Sun and Yu [8] were first to introduce the border based approach. It hides sensitive association rules by modifying the border in the lattice of frequent and infrequent item set of the original database. This approach consist the item sets which separates the frequent and infrequent item set.

D. Cryptographic Approaches: - This approach is used in multiparty computation where data is distributed on different location. The owner wants to share their data, but at the same time they want to ensure about the privacy at their end. This approach is of two type vertical partitioned distributed data and horizontal partitioned distributed data.

2. Hiding sensitive frequent item sets

The paper is based on the concept of frequent item set. Let I = {i1,……..,im} be a set of items. An item set is a subset of I. A transaction T is a pair (tid, X), where tid is a unique identifier of a transaction and X is an item set. A transaction (tid, X) is said to contain an item set Y iff X is a superset of Y. A database D is a set of transaction. The support of an item set X, denoted as Supp(X). For a given threshold σ, X is said to be σ – frequent if Supp (X) ≥ σ.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 1, January 2016

Here an example, is taken

Tid	item
1	abcde
2	acd
3	abdfg
4	bcde
5	abd
6	bcdfh
7	abcg
8	acde
9	acdh

Fig 1. Database

Frequent item set: Support
abd: 3, acd: 4, bcd : 3, cde : 3
ab:4, ac:5, ad:6, bc:4, bd:5, cd:6, ce:3, de:3
a:7, b:6, c:7, d:8, e:3

Fig 2. All Frequent item set

Fig 2. All frequent item sets

Expected frequent item sets on D’
acd, cde
ab, ac, ad, bd, cd, ce, de
a,b,c,d,e

Fig 3. Non-sensitive frequent item set

A transaction database D is given in Fig 1. Let the support threshold σ be 3.
Fig 3. shows all σ – frequent item sets in D. Among those frequent item sets, abd, bcd, and bc are sensitive item sets. Now how to transform D into the result database D’ in a sensible way such that the sensitive frequent item sets become infrequent in D’ and the quality of D’ is maintained.

3. **Description of border based approach: -** Here a border based approach is proposed to address the hiding problem. Here, the non-sensitive frequent item sets (Lr) is used to track the impact on the result database and also maintain the quality of the database by selecting the modification with minimal impact at each step.

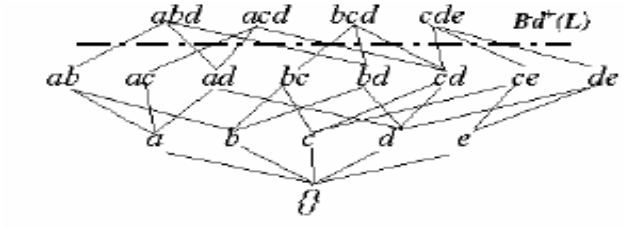

Fig 4. Lattice and border 1

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 1, January 2016

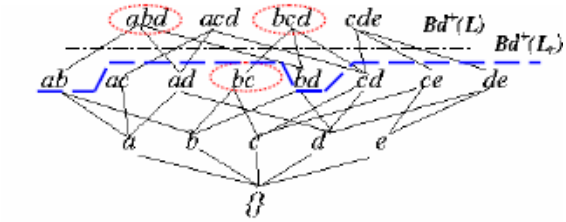

Fig 5. Lattice and border 2

The concept of border is introduced in [9] and it is well applied in the research of maintaining the frequent item sets (e.g., in [10]). Given a set of item sets U called the upper border denoted by Bd+ (U) and lower border as Bd− (U),which is a subset of U such that 1) Bd+ (U) (respectively, Bd− (U)) is an antichain collection of sets1 and 2) ∀X ∈ U, there exists at least one item set Y ∈ Bd+ (U). An item set in the upper border or lower border is called a border element. In this example, we have Bd+ (L) = {abd, acd, bcd, cde} and Bd− (_ΔL) = {abd, bc} . The item set bc is a border element of Bd− (ΔL) .Due to the Apriori property, there is only need to hide the lower border Bd− (ΔL). The graphical representation of Bd+ (L) in the item set lattice is shown in Fig 4. Fig 5 shows Bd+ (Lr) in our example (note that the sensitive frequent item sets are circled). The support of border elements is relatively low and the relative frequency among them is sensitive to the sanitization. Focusing on the most sensitive part of the frequent item sets can effectively avoid the significant change on the relative frequency. Due to Apriori property, the support of the border could also reflect the support of the other frequent item sets to some degree.

III. ALGORITHM

Main Algorithm
Input: A database *D*, the set *L* of *σ*-frequent item set in *D* and the set of sensitive item sets Δ*L*
Output: *D'* so that the quality is maintained
Method:
Compute *Bd*-and *Bd*+;
Sort itemsets in *Bd*- in descending order of length and ascending order of support;
for each *X* ∈*Bd*- do
Compute *Bd*+|*X* and *w* (*Bj*) where *Bj* ∈*Bd*+|*X*;
Initialize *C* (*C* is the set of hiding candidates of *X*);
for(*i* = 0;*i* < *Supp*(*X*) − *σ* + 1; *i*++) do
/* Candidate selection algorithm*/
Find *ui* = (*Ti*, *xi*) such that *I* (*ui*) = *Min*{*I* (*u*) |*u* ∈*C*} ;
Update *C* = *C* − {(*T*, *x*) |*T* = *Ti*};
Update *w* (*Bj*) where *Bj* ∈*Bd*+|*X*;
Update database *D*;
Output *D'*= *D*;

Here the main algorithm of border based is shown of hiding sensitive frequent item sets, which is a summary of the approach. The key step of the algorithm is to efficiently find a hiding candidate with minimal impact on the border. After selecting a candidate, need to update the hiding candidate set and the weights of the border elements (for each sensitive frequent item set, the database is update once after selecting all hiding candidate).

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 1, January 2016

IV. ADVANTAGES AND LIMITATIONS OF HIDING APPORACHES

In Heuristic approach data distortion is scalable and more efficient but difficult to revert the changes made in database. Data Blocking maintains veracity of database, since instead of inserting false value it just block original value but suffer from various side effects like ghost rule, lost rule etc.
In Border based approach database quality is maintained by selecting the transaction that produces minimal side effect. Based on heuristic approach theory of border is difficult to understand.
In Exact approach without any side effects it provides an optimal solution but high complexity due to linear integer programming.
In Reconstruction based approach lesser side effect than heuristic based approaches but in new released database number of transaction is restricted.
Cryptography based approach provide security in multi party computation or where data distributed in different locations but it does not provide security for the output of the computation.

V. SIMULATION RESULTS

The simulation studies involves the techniques of association rule hiding as shown in Table 1 and Table 2. By this distortion and blocking techniques are described. An example is taken in which database is shown as fig [1] in which an item is entered with transaction Id. In fig [2] all frequent item are shown and in fig [3] non-sensitive frequent set are shown. Here a border based approach is proposed to address the hiding problem. By analysis the approach there is problem in maintaining the database with minimal impact at each step as shown in fig [4] and fig [5]. Our results shows that the algorithm which we designed reduce the side effect of resultant database.

VI. CONCLUSION AND FUTURE WORK

In this paper, we have studied about the problem of hiding frequent item sets. Various approaches of association rule mining are described in this paper. A border based approach was proposed to efficiently select the modification with minimal side effect. In future we are tried to combine the various approaches with the border based algorithm to reduce the side effect of result database.

REFERENCES

[1] V. S. Verykios, E. Bertino, I. N. Fovino, L. P. Provenza, Y. Saygin, and Y. Theodoridis. State-of-the-art in privacy preserving data mining. *ACM SIGMOD Record*, 33:50–57, 2004.
[2] Aris Gkoulalas–Divanis;Vassilios S. Verykios "Association Rule Hiding For Data Mining" Springer, DOI 10.1007/978-1-4419-6569-1, Springer Science + Business Media, LLC 2010.
[3] 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.
[4] V.S. Verykios, A. Elmagarmid, E. Bertino, Y. Saygin, and E. Dasseni, "Association rule hiding," IEEE Transactions on Knowledge and Data Engineering, Vol. 16, No.4, 434–447, 2004 .
[5] Y. Saygin, V. Verykios, and C. Clifton, "Using Unknowns to Prevent Discovery of Association Rules" ACM SIGMOD, Vol. 30, No. 4, pp. 45–54, 2001.
[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]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.
[8]X. Sun, and P. Yu, "A Border-Based Approach for Hiding Sensitive Frequent Itemsets," In: Proc. Fifth IEEE Int'l. Conf. Data Mining (ICDM 2005), pp. 426–433, 2005.
[9]H. Mannila and H. Toivonen. Levelwise search and borders of theories in knowledge discovery. *Data Mining and Knowledge Discovery*, 1(3):241–258, 1997
[10] A. Veloso, W. Meira, Jr., M. de Carvalho, B. Possas, S. Parthasarathy, and M. J. Zaki. Mining frequent itemsets in evolving databases. In *Proc. of the 2nd SDM*, 2002.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 1, January 2016

BIOGRAPHY

Rachana Naik received BE degree in department of Information Technology from Mahakal Institute of Technology, Ujjain, India in 2012 and pursuing ME in department of Information Technology from MIT, Ujjain, India.

Prof. Pradeep Rupayla has received his ME (Software Engineering) degree from Jawahar Institute of Technology Borawan, Khargone, India in 2015. Presently, He is Assistant Professor at Computer Science department MIT, Ujjain, India. His research interest include as frequent item sets. He has 7 years of teaching experience. He has publish paper in IJIRCCE.