HIDING SENSITIVE ASSOCIATION RULE USING HEURISTIC APPROACH

Kasthuri S[1] and Meyyappan T[2]

[1]Research Scholar, Alagappa University, Karaikudi.
Kasthu.s@gmail.com
[2]Professor, Alagappa University, Karaikudi.
Meyslotus@yahoo.com

ABSTRACT

Data mining is the process of identifying patterns from large amount of data. Association rule mining aims to discover dependency relationships across attributes. It may also disclose sensitive information. With extensive application of data mining techniques to various domains, privacy preservation becomes mandatory. Association rule hiding is one of the techniques of privacy preserving data mining to protect the sensitive association rules generated by association rule mining. This paper adopts heuristic approach for hiding sensitive association rules. The proposed technique makes the representative rules and hides the sensitive rules.

KEYWORDS

Association Rule Hiding, Data Mining, Privacy Preserving Data Mining, Distortion, Representative Rule.

1. INTRODUCTION

Data mining is a knowledge extracting process of analyzing data from different point of views. This knowledge is expressed in decision trees, clusters or association rules. Association rule mining is a common task in data mining. By discovering interesting association rules in datasets, an organization can identify underlying patterns useful in strategic decision making. This technique has been successfully applied to many application domains such as the analysis of market baskets, website linkages etc. The knowledge discovered can be beneficial to cooperating organizations. In order to preserve their competitive edge some partners may hesitate to disclose sensitive information.

Privacy preserving data mining is a major research area for protecting sensitive data or knowledge. Association rule hiding is one of the privacy preserving techniques to hide sensitive association rules. The main aim of all association rule hiding algorithm is to minimally modify the original database and see that no sensitive association rule is derived from it.

Next section describes the association rule mining. Section 3 explains approaches of association rule hiding algorithms. Section 4 presents the statement of the problem. Section 5 analyses the related work on heuristic approaches based on data distortion technique. Section 6 presents the proposed algorithm for sensitive rule hiding. Section 7 shows the examples demonstrating the algorithm. Concluding remarks and future works are described in section 8.

DOI : 10.5121/ijdkp.2013.3105

2. ASSOCIATION RULE MINING

Let I = {i_1,…., i_n} be a set of items. Let D be a database which contains set of transactions. Each transaction t Є D is an item set such that t is a proper subset of I. As transaction t supports X, a set of items in I, if X is a proper subset of t. Assume that the items in a transaction or an item set are sorted in lexicographic order. An association rule is an implication of the form X→Y, where X and Y are subsets of I and $X \cap Y = \emptyset$. The support of rule X→Y can be calculated by the following equation: Support(X→Y) = |X→Y| / |D|, where |X→Y| denotes the number of transactions ceontaining the itemset XY in the database, |D| denotes the number of the transactions in the database D. The confidence of rule is computed by Confidence(X→Y) = |X→Y|/|X|, where |X| is number of transactions in database D that contains itemset X. A rule X→Y is strong if support(X→Y) ≥ min_support and confidence(X→Y) ≥ min_confidence, where min_support and min_confidence are two given minimum thresholds.

Association rule mining algorithms calculate the support and confidence of the rules. The rules having support and confidence higher than the user specified minimum support and confidence are retrieved. Association rule hiding algorithms prevents the sensitive rules from being revealed out. The problem can be declared as follows "Database D, minimum confidence, minimum support are given and a set R of rules are mined from database D. A subset SR of R is denoted as set of sensitive association rules.SR is to be hidden. The objective is to modify D into a database D' from which no association rule in SR will be mined and all non sensitive rules in R could still be mined from D'.

3. APPROACHES OF ASSOCIATION RULE HIDING ALGORITHMS

Association rule hiding algorithms can be divided into three distinct approaches. They are *heuristic* approaches, *border-revision* approaches and *exact* approaches.

3.1 Heuristic Approach

Heuristic approaches can be further categorized into distortion based schemes and blocking based schemes. To hide sensitive item sets, distortion based scheme changes certain items in selected transactions from 1's to 0's and vice versa. Blocking based scheme replaces certain items in selected transactions with unknowns. These approaches have been getting focus of attention for majority of the researchers due to their efficiency, scalability and quick responses.

3.2 Border Revision Approach

Border revision approach modifies borders in the lattice of the frequent and infrequent item sets to hide sensitive association rules. This approach tracks the border of the non sensitive frequent item sets and greedily applies data modification that may have minimal impact on the quality to accommodate the hiding sensitive rules. Researchers proposed many border revision approach algorithms such as BBA (Border Based Approach), Max– Min1 and Max-Min2 to hide sensitive association rules. The algorithms uses different techniques such as deleting specific sensitive items and also attempt to minimize the number of non sensitive item sets that may be lost while sanitization is performed over the original database in order to protect sensitive rules.

3.3 Exact Approach

Third class of approach is non heuristic algorithm called exact, which conceive hiding process as constraint satisfaction problem. These problems are solved by integer programming. This approach can be concerned as descendant of border based methodology.

4. PROBLEM STATEMENT

Data mining represents a wide range of tools and techniques to extract useful information which can contain sensitive information from a large collection of data. Data should be manipulated or distorted in such a way that sensitive information cannot be discovered through data mining techniques. Sensitive information has to be protected against unauthorized access. The major challenge faced is better balancing the confidentiality of the disclosed with the legitimate needs of the data user. The proposed approach is based on modification of database transactions.

5. ANALYSIS OF EXISTING TECHNIQUES

In Distortion Based Technique (Proposed By Veryki- os *Et Al*, Etc.) [1], authors propose strategies and a suite of algorithms for hiding sensitive knowledge. In order to achieve this, transactions are modified by removing few items, or inserting new items depending on the hiding strategy.

The distortion based Technique (Proposed by shyue-liang wang et al.) [2] hides certain specific items that are sensitive. In this technique, two algorithms are proposed to modify data in the Dataset. If the sensitive item is on the LHS of the rule then the first algorithm increases its support. If the sensitive item is on the right of the rule then the second algorithm decreases its support.

In [1], author tries to hide every rule without checking if rules can be pruned after some transactions have been changed. In [2] the author hides all the rules which contain sensitive items either in the left or in the right. Two different algorithms are applied over the data. The first algorithm hides association rules with sensitive items on the LHS and the second one for sensitive items on the RHS. It takes more number of passes to prune all the rules containing sensitive items.

6. PROPOSED APPROACH

The proposed approach selects all the association rules containing sensitive items either in the left or in the right from the set of all association rules generated from a dataset. These rules are represented in representative rules (RR) format with sensitive item on the left hand side or right hand side of the rules. Select a rule from the set of RR's which contains sensitive item. Select a transaction which completely supports RR i.e. it contains all the sensitive items in the RR. The proposed approach hides the sensitive item by modifying the database without changing the support of the sensitive item.

6.1 Representative Association Rule

The number of association rules discovered in a given database is very large. A considerable percentage of these rules are redundant and useless. A user should be presented with original, novel, and interesting. To address this problem, [6] introduced a representation of association rules, called representative rules (RR). RR is a set of rules that allow deducing all association rules without accessing a database. The cover operator was introduced for driving a set of association rules from a given association rule. The cover of the rule A=>B, $A \neq \phi$, is defined as follows:

$$C(A \Rightarrow B) = \{A \cup B \Rightarrow V \mid Z , V \subseteq B \text{ and } Z \cap V = \phi \text{ and } V \neq \phi\}$$

Each rule in C ($A \Rightarrow B$) consists of a subset of items occurring in the rule $A \Rightarrow B$. The number of different rules in the cover of the association $A \Rightarrow B$ is equal to $3^m - 2^m$, $m = |B|$.

The process of generating representative rules is decomposed in to two sub processes: frequent item-sets generations and generation of RR from frequent item-sets. Let be a frequent itemset and $\phi \neq A \subset B$. The association rule $A \Rightarrow Z / B$ is representative rule if there is no association rule $(A \Rightarrow Z'/ A)'$ where $Z \subset Z'$, and there is no association rule $(A' \Rightarrow Z / A')$ such that $A \supset A'$. Formally, a set of representative rules (RR) for a given association rules (AR) can be defined as follows:

$$RR = \{ r \in AR \mid \neg \exists\, r' \in AR,\ r' \neq r \text{ and } r \in C(r') \}$$

Each rule in RR is called representative association rule. No representative rule may belong in the cover of another association rule [8], [9].

The proposed algorithm gives a modified dataset after distorting the database. A Database, value of min_support, min_confidence, and a set of sensitive items are given as input this algorithm. Association rules are generated using association rule mining algorithm. Select the rules containing sensitive item from the association rules. The selected rules are represented in the representative rules format. The sensitive item is deleted from a transaction, which fully supports the selected RR and added to a transaction, which partially supports RR. The algorithm is given below

Algorithm: Hiding Of Sensitive Association Rules
Input:

(1) D: A source database
(2) min_supp : A min_support.
(3) min_conf : A min_confidence.
(4) H: A set of sensitive items.

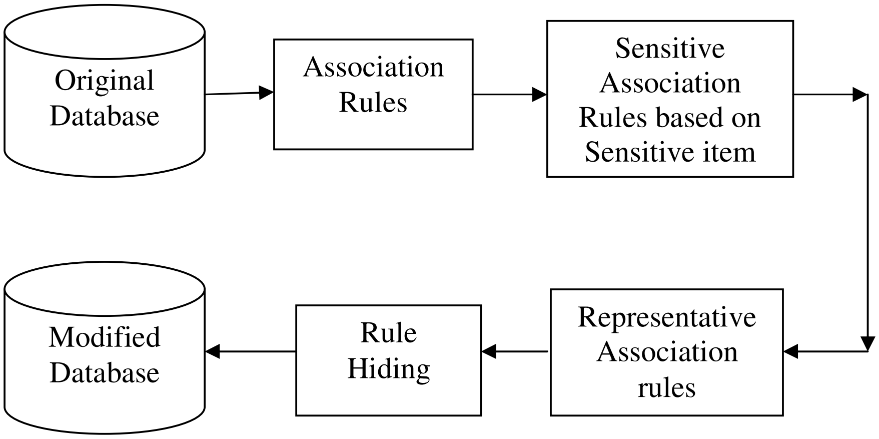
Output:
A transformed database D' where rules containing H on LHS/RHS will be hidden

```
Find item sets from D;
For each sensitive item h∈ H {
If h is a small item set then H=H- {h};
If H is null then EXIT;
Select all the rules with min_supp containing h and store in SR//h can either be on LHS or RHS
Repeat {
Select rule from SR with same LHS
Combine RHS of selected rule and store in R; // representative rules
} Until (U is empty);
Sort R in descending order by supported items;
Select a rule r from SR
Compute confidence of rule r;
If conf>min_conf then {//change the position of sensitive item h.
Find $T_i$ = {t in D |t completely supports R;
If t contains x and h then
Remove h from t
Find $T_i$ = {t in D|t does not support and partially supports x;
Add h to t
Repeat
{
Choose the first rule from U;
22. Compute confidence of r;
23. } Until(SR is empty);
24. }//end of if conf>min_conf
25. Else
26. Go to step 11;
27. Update D with new transaction t;
28. Remove h from H;
29. Go to step 2;
30.}//end of for each h∈ H
```

1. All the rules containing sensitive item(s) either in the left or in the right are selected.
2. Rules are converted in representative rules (RRs) format.
3. A rule from the set of RR's, which has sensitive item on the left of the RR is selected.
4. The sensitive item(s) from the transaction that completely supports the RR is removed and add the same sensitive item to a transaction which partially supports RR.
5. The confidence of the rules in U is recomputed.

The framework of the proposed approach is shown in figure 1.

Figure 1 – Framework of the proposed Approach


Original Database
Association Rules
Sensitive Association Rules based on Sensitive item
Representative Association rules
Rule Hiding
Modified Database

7. Implementation Of The Proposed Algorithm

The proposed algorithm can be illustrated with the following example for a given set of transactional data in Table -1

Table -1: Transactional Dataset1

TID	ITEMS
T1	bread, butter, milk
T2	bread, butter, milk, cheese
T3	butter, milk, fruits
T4	bread, milk, cheese, fruits
T5	cheese, fruits
T6	bread, butter

For the Dataset given in Table - 1 at a min_supp of 33% and a min_conf of 70 % and sensitive item H= {milk} we choose all the rules containing 'butter' either in RHS or LHS and represent them in representative rule format. Out of the 8 association rules the rules containing sensitive items are 6 as shown in Table 2

Table - 2: Sensitive association rules (w.r.t sensitive item 'milk')

AR	SUPP	CONF
bread => milk	50	75
milk => bread	50	75
bread, cheese => butter	33.333	100
milk, cheese => bread	33.333	100
butter => milk	50	75
milk => butter	50	75

From this rules set select the rules that can be represented in the form of representative rules Like milk=> butter and milk=>bread can be represented as milk=> bread, butter Now delete milk from a transaction where bread, butter, milk all the three are present and add milk to a transaction where bread and butter both are absent or only one of them is present. For this we change transaction T2 to bread, butter, cheese and transaction T5 to milk, cheese, and fruits. This results in changing the position of the sensitive item without changing its support. This is shown in Table 3.

Table-3: Modified Dataset1 for the proposed Approach (Sensitive Item – 'milk')

TID	ITEMS
T1	bread, butter, milk
T2	bread, butter, cheese
T3	butter, milk, fruits
T4	bread, milk, cheese, fruits
T5	milk, cheese, fruits
T6	bread, butter

The new set of association rules generated from this modified dataset is shown in Table-4.

Table-4: Association rules remaining unhidden after modifying the Dataset1

AR	SUPP	CONF
butter=> bread	50	75
bread=>butter	50	75

i.e. all the rules of the original association rules set containing sensitive items on the LHS or on the RHS are hidden.

8. CONCLUSIONS

In this paper, the database privacy problems are addressed and a new technique for privacy preservation is proposed. Association rule hiding techniques are used to hide sensitive association rules. A new heuristic method to hide the sensitive association rules is proposed. Data distortion technique is applied so that sensitive information cannot be discovered through data mining techniques. Confidence of the rules is represented as representative rules. Confidence of the rule is recomputed and compared with threshold level. The confidence of the sensitive rules might be reduced while maintaining the support. From the experimental results, it is observed that all the rules containing sensitive items are hidden. The algorithm is implemented and numerical example is shown. Further research is in progress to evolve a method which can avoid the computational overhead associated with confidence of the rules.

REFERENCES

[1] Vassilios S. Verykios,, Ahmed K. Elmagarmid , Elina Bertino, Yucel Saygin, Elena Dasseni. “Association Rule Hiding”, IEEE Transactions on knowledge and data engineering, Vol.6, NO.4, April 2004

[2] Shyue-Liang Wang, Yu-Huei Lee, Billis S., Jafari, A. “Hiding sensitive items in privacy preserving association rule mining”, IEEE International Conference on Systems, Man and Cybernetics, Volume 4, 10-13 Oct. 2004 Page(s): 3239 - 3244 .

[3] E. Dasseni, V. Verykios, A. Elmagarmid and E. Bertino. “Hiding Association Rules by Using Confidence and Support”, in Proceedings of 4th Information Hiding Workshop, 369-383, Pittsburgh, PA, 2001.

[4] JIAWEI HAN and MICHALINE KAMBER, “Data Mining Conceptsd And Techniques”, Morgon Kaufman Publishers 2002 [5] Wang. S.L., Jafari, A. “Using unknowns for hiding sensitive predictive association rules”, Information Reuse and Integration, Conf, 2005. IRI -2005 IEEE International Conference on. 15-17 Aug. 2005 Page(s): 223 - 228.

[6] Marzena Kryszkiewicz. “Representative Association Rules”, In proceedings of PAKDD’98, Melbourne,Australia(Lecture notes in artificialIntelligence,LANI 1394, Springer-Verleg,1998,pp 198-209.

[7] Yucel Saygin, Vassilios S. Verykios, Chris Clifton. “Using unknowns to prevent discovery of association rules”, ACM SIGMOD Record Volume 30 Issue 4, pp. 45 - 54 , (2001)

[8] Yiqun Huang, Zhengding Lu, Heping Hu, “A method of security improvement for privacy preserving association rule mining over vertically partitioned data”, 9th International Database Engineering and Application Symposium, pp. 339 – 343, (2005)

[9] Saygin Y., Verykios V.S. and Elmagarmid A.K., “Privacy preserving association rule mining,” IEEE Proceedings of the 12th Int’l Workshop on Research Issues in Data Engineering, pp. 151 – 158, (2002)

[10] Aris Gkoulalas–Divanis;Vassilios S. Verykios “Asssociation Rule Hiding For Data Mining” Springer, DOI 10.1007/978-1-4419-6569-1, Springer Science + Business Media, LLC 2010