NEW ALGORITHM FOR SENSITIVE RULE HIDING USING DATA DISTORTION TECHNIQUE

Kasthuri S[1] and Meyyappan T[2]

[1]Lecturer, Swami Dayananda College of Arts and Science, Manjakkudi
Kasthu.s@gmail.com
[2]Professor, Alagappa University, Karaikudi.
Meyslotus@yahoo.com

ABSTRACT

Data mining is the process of extracting hidden patterns of data. Association rule mining is an important data mining task that finds interesting association among a large set of data item. It may disclose pattern and various kinds of sensitive information. Such information may be protected against unauthorized access. Association rule hiding is one of the techniques of privacy preserving data mining to protect the association rules generated by association rule mining. This paper adopts 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. In the proposed technique, positions of sensitive items are altered while maintaining the support. The proposed technique uses the idea of representative rules to prune the rules first and then hides the sensitive rules.

KEYWORDS

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

1. INTRODUCTION

Data mining is a knowledge discovery process of analyzing data from different point of views and to work out into useful patterns. These patterns represent knowledge and are expressed in decision trees, clusters or association rules. The problem of data mining is that from non-sensitive data one is able to infer sensitive information including personal information facts or even patterns that are not supposed to be disclosed. For example the market basket database which is used to analyze customers purchasing behavior represented in the form of association rules. The disclosure of this business information or any knowledge extracted from the data may potentially reveal sensitive trade secrets, whose knowledge can provide a significant advantage to business competitors and thus can cause the data owner to lose business over his or her peers [10].

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. All association rule hiding algorithm aims to minimally modify the original database such that no sensitive association rule is derived from it.

Next section describes the association rule mining. Section 3 explains classes of association rule hiding algorithms. Section 4 presents the statement of the problem. Section 5 contains the related work on heuristic approaches based on data distortion technique. Section 6 presents the analysis of existing work. Section 6 presents the proposed algorithm for sensitive rule hiding. Section 8

Sundarapandian et al. (Eds): CoNeCo,WiMo, NLP, CRYPSIS, ICAIT, ICDIP, ITCSE, CS & IT 07,
pp. 373–379, 2012. © CS & IT-CSCP 2012 DOI : 10.5121/csit.2012.2436

shows the examples demonstrating the algorithm. Concluding remarks and future works are described in section 9.

2. ASSOCIATION RULE MINING

Let I = {i1,…., in} be a set of items. Let D be a set of transactions or database. Each transaction t Є D is an item set such that t is a proper subset of I. A 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∩Y= Ø. The support of rule X→Y can be computed by the following equation: Support(X→Y) = |X→Y| / |D|, where |X→Y| denotes the number of transactions in the database that contains the itemset XY, and |D| denotes the number of the transactions in the database D. The confidence of rule is calculated by following equation: 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 scan the database of transactions and calculate the support and confidence 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 algorithms prevents the sensitive rules from being disclosed. The problem can be stated as follows: "Given a transactional database D, minimum confidence, minimum support and a set R of rules mined from database D. A subset RH of R is denoted as set of sensitive association rules which are to be hidden. The objective is to transform D into a database D‟ in such a way that no association rule in RH will be mined and all non sensitive rules in R could still be mined from D‟.

3. CLASSES OF ASSOCIATION RULE HIDING ALGORITHMS

Association rule hiding algorithms can be divided into three distinct classes, namely *heuristic* approaches, *border-based* approaches and *exact* approaches. The first class of approaches involves efficient, fast algorithms that selectively sanitize a set of transactions from the database to hide the sensitive knowledge. Due to their efficiency and scalability, the heuristic approaches have been the focus of attention for the vast majority of researchers in the knowledge hiding field. However, there are several circumstances in which they suffer from undesirable side-effects that lead them to suboptimal solutions. The second set of approaches considers the task of sensitive rule hiding through modification of the original borders in the lattice of the frequent and the infrequent patterns in the dataset. In these schemes, the sensitive knowledge is hidden by enforcing the revised borders (which accommodate the hiding of the sensitive itemsets) in the sanitized database. The algorithms in this class differ both in the borders that they track and use for the hiding strategy, and in the methodology that they follow to enforce the revised borders in the sanitized dataset. Finally, the third class of approaches contains non-heuristic algorithms which conceive the hiding process as a constraint satisfaction problem that they solve by using integer or linear programming.

4. PROBLEM STATEMENT

The expression data mining indicates a wide range of tools and techniques to extract useful information which can be sensitive from a large collection of data. Data should be manipulated or distorted in such a way that information cannot be discovered through data mining techniques. While dealing with sensitive information it becomes very important to protect data against unauthorized access. The key problem faced is the need to balance the confidentiality of the disclosed with the legitimate needs of the data user. The proposed approach is based on modifying the database transaction.

5. PREVIOUS WORK

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

In Distortion based Technique on the basis of sensitive item (proposed by shyue-liang wang et al.) proposed in this work tries to hide certain specific items that are sensitive and proposes two algorithms to modify data in the Dataset. Concept of this paper says that if the sensitive item is on the LHS of the rule then increase its support and if the sensitive item is on the right of the rule then decrease its support.

6. ANALYSIS OF EXISTING TECHNIQUES

Approach in [1] tries to hide every single rule without checking if rules can be pruned after some transactions have been changed. Approach in [2] definitely hides all the rules which has sensitive items either in the left or in the right and for this it runs two different algorithms one if sensitive item is on the LHS and another is the sensitive item is on the RHS i.e. it fails to hide all the rules containing sensitive item and takes more number of passes to prune all the rules containing sensitive items.

7. 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. Then these rules are represented in representative rules (RR) format with sensitive item on the LHS/RHS of the rules. Select a rule from the set of RR's, which has sensitive item on the LHS/RHS of the rule. Select a transaction that completely support RR i.e. it contains all the items in the RR. Based on this a new approach for modifying database without changing the support of the sensitive item and still maintaining the secrecy of sensitive data has been proposed.

7.1 Representative Association Rule

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

$$C(X => Y) = \{X \cup Y => V \mid Z , V \subseteq Y \text{ and } Z \cap V = \phi \text{ and } V \neq \phi\}$$

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

In general, the process of generating representative rules may be decomposed in to two sub processes: frequent item-sets generations and generation of RR from frequent item-sets. Let be a frequent itemset and $\phi \neq X \subset Y$. The association rule $X \Rightarrow Z$ /X is representative rule if there is no

association rule ($X \Rightarrow Z' / X)'$ where $Z \subset Z'$, and there is no association rule $(X' \Rightarrow Z / X')$ such that $X \supset X'$. 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 and 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. Input to this algorithm is a Database, value of min_support, min_confidence, and a set of sensitive items. This algorithm computes the large item sets of all the sizes from the given dataset. Then it selects all the rules, which contain sensitive item from the association rules generated. The rules containing sensitive items are represented in the representative rules format and then the sensitive item is deleted from a transaction, which fully supports the selected RR and added to a transaction, which partially supports RR. The detailed steps is given below

Algorithm: Hiding Of Sensitive Association Rules Using Support and Confidence

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 RHS/LHS will be hidden
1. Find all large item sets from D;
2. For each sensitive item h∈ H **{**
3. If h is not a large item set then H=H- {*h*};
4. If H is empty then EXIT;
5. Select all the rules with min_supp containing h and store in U//*h* can either be on LHS or RHS
6. Repeat {
7. Select all the rules from U with same LHS
8. Join RHS of selected rules and store in R; //make representative rules
9. }Until (U is empty);
10.Sort R in descending order by the number of supported items;
11. Select a rule *r* from R
12.Compute confidence of rule *r*.
13.If conf>min_conf then {//change the position of sensitive item h.
14.Find T1={t in D|t completely supports *r* ;
15. If t contains x and h then
16.Delete h from t
17.Find T1={t in D|t does not support LHS(*r*) and Partially supports x;
18.Add h to t
19. Repeat
20. **{**
21. Choose the first rule from U;
22. Compute confidence of *r;*
23. **}** Until(U 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

Steps of the algorithm are given below:

1. Selects all the rules containing sensitive item(s) either in the left or in the right.
2. Convert these rules in representative rules (RR) format.
3. Selects a rule from the set of RR's, which has sensitive item on the left of the RR is selected.
4. Deletes the sensitive item(s) from the transaction that completely supports the RR i.e. it contained all the items in of RR selected and add the same sensitive item to a transaction which partially supports RR i.e. where items in RR are absent or only one of them is present.
5. Recomputed the confidence of the rules in U.

8. 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	ABC
T2	ABCD
T3	BCE
T4	ACDE
T5	DE
T6	AB

For the Dataset given in Table - 1 at a min_supp of 33% and a min_conf of 70 % and sensitive item H={C} we choose all the rules containing 'C' 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 C)

AR	SUPP	CONF
A => C	50	75
C => A	50	75
A, D => C	33.333	100
C, D => A	33.333	100
B => C	50	75
C => B	50	75

From this rules set select the rules that can be represented in the form of representative rules Like C=> A and C=>B can be represented as C=> AB Now delete C from a transaction where ABC all the three are present and add C to a transaction where A and B both are absent or only one of them is present. For this we change transaction T2 to ABD and transaction T5 to CDE. 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 – C)

TID	ITEMS
T1	ABC
T2	ABD
T3	BCE
T4	ACDE
T5	CDE
T6	AB

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
B=> A	50	75
A=> B	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.

9. CONCLUSIONS

In this paper, we presented the database privacy problems caused by data mining technology. Association rule mining is an important data mining task that finds interesting association among a large set of data item. It may disclose pattern and various kinds of sensitive information. Such information may be protected against unauthorized access.

The main aim of this work is to propose a new method to hide the sensitive association rules. Data will be distorted in such a way that sensitive information cannot be discovered through data mining techniques. The proposed work analyses the existing techniques and gives their limitations.

The proposed method uses the idea of representative rules. The confidence of the sensitive rules will be reduced but the support remains the same. It doesn't modify the database transactions. This method will hide all the rules containing the sensitive items. Example demonstrating the proposed algorithm is shown.

In the proposed method, the confidence of the rule is computed and it is represented as representative rules. After hiding the sensitive rules the confidence of RR is again computed even if it falls below the min_conf threshold. So it has to be avoided in the future.

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 "Association Rule Hiding For Data Mining" Springer, DOI 10.1007/978-1-4419-6569-1, Springer Science + Business Media, LLC 2010