INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & MANAGEMENT

HIDING SENSITIVE DATA ITEM USING ASSOCIATION RULE MINING

Satish Choudhary[1], Arvind Upadhyay[2]
Institute of Engineering & Science IPS Academy, Indore (M.P) India
Satishchoudhary1989@gmail.com

ABSTRACT

Data mining useful technology through which sensitive information and the relationships among the items in a database are identified .the users and originations are need to share their data and they should manage the data for preserving the privacy of sensitive data for their progress. The concept of preserving the privacy of information in data mining was introduced to manage the sharing of information. This paper presents a hybrid algorithm with distortion technique with both support-based and confidence-based approaches for privacy preserving. The proposed algorithm tries to hide sensitive rules from the perspective of the database owner and maintain useful association rules.

Keywords: Data Mining, Association Rule, Sensitive data set, Support ,Confidence

I. INTRODUCTION

Data mining is the process of extracting hidden patterns from data. Amount of data doubling every three years as more data is gathered, an important tool to transform this data into knowledge is data mining. Wide range of applications used data mining for the marketing and fraud detection and the purpose of scientific discovery. Data mining can be applied to data sets of any size, and to uncover hidden patterns it can be used, it cannot uncover patterns which are not already present in the data set. For an effective analysis and decision means the useful knowledge is extracted by data mining.

An automated extraction of novel is knowledge discovery in databases (kdd), in large databases understandable and potentially useful patterns implicitly is stored.

Data mining is an essential step in the process of knowledge discovery in databases, in which in order to extract patterns, intelligent methods are applied. Other steps in knowledge discovery process include pre-mining tasks such as data cleaning (inconsistency and noise removing of data) and data integration (to merge the data in single location from multiple source), as well as post mining tasks such as pattern evaluation (representing knowledge of interesting pattern that is evaluated) and knowledge presentation (the extracted rules are presented using visualization and knowledge representation techniques).

Association rule mining

Association rule mining finds interesting associations in large data base and/or correlation relationships among large sets of data items [1]. The data item that occurs frequently together in a given dataset is shown by association rules. A typical and widely-used example of association rule mining is market basket analysis [2]. For example, data are collected using bar-code scanners in supermarkets. Large number of transaction records consists in such market basket data base. By a customer on a single purchase transaction, each record lists all items bought. The groups of items which are consistently purchased together managers would be interested to know if certain. They could use this data for adjusting store layouts (with respect to each other placing items and item cross-selling , for promotions of items, for catalog design and buying pattern is used to identify customer segments. The information of this type in the form of "if-then" statements is provided by association rule. Unlike the if-then rules of logic, association rules are probabilistic in nature. These rules are computed from the data

In addition "if" part is called antecedent and the consequent is called the "then" part, and the degree of uncertainty about the rule is expressed by these two rules. During the rule analysis the sets of items (called item sets) are antecedent and consequent that are disjoint (do not have any items in common). The first number is called the support for the rule. The percentage of the total number of records in the database is expressed as a support or it is the item in antecedent and consequent part of rule is support. Confidence of the rule is known by other number. Confidence is the ratio of the number of transactions that include all items in the consequent ("then" part) as well as the antecedent ("if" part or namely, the support) to the number of transactions that include all items in the antecedent.

I.LITERATURE SURVEY

Data mining is often defined as the process of discovering meaningful, and interesting patterns and trends through implicit extraction of non trivial set of data, and extraction of unknown information from repositories of large amount of data, using pattern recognition as well as statistical and mathematical techniques. An sql query is usually stated OR written to retrieve specific data, while data miner are not sure what they are exactly require. The literature of privacy-preserving association rule-mining, researchers presented different privacy-preserving data mining problems based on the classifications of the authors. These classifications are good but we believe that from the point of view of the targeted people (individuals and parties who want to protect their sensitive data), it is difficult to understand.

In the literature of privacy-preserving association rule-mining, researchers presented different privacy-preserving data mining problems based on the classifications of the authors. These classifications are good but we believe that from the point of view of the targeted people (individuals and parties who want to protect their sensitive data), it is difficult to understand.

Association rule hiding

The association rule hiding problem was probed in 1999 firstly. After that proposed many method and approaches. Generally, there are two groups: data sanitization data modification approaches and sanitization of knowledge and reconstruction of data approaches. The basic idea of data sanitization is **data modification approaches**. They sanitizing the original data d directly to hide the sensitive items and the released database d' gated directly from d, most of the existing methods belong to this data modification track, it can be further classified by means of different modification into the two techniques of distortion of data and blocking of data. However, approaches of data modification cannot control the hiding effects intuitively as on the data level sanitization is performed. Moreover, data sanitization can produce a lot of i/o operations, especially a large number of transactions included in the original database. The other solution is the **data reconstruction approaches** towards the association rule hiding problem. The basic idea is knowledge sanitization and data reconstruction. They put the original data aside unlike data modification, and start from sanitizing the so-called "knowledge base" k. From the sanitized knowledge base k data d' (apostrophe) which is new released data base is then reconstructed. By the recently emerging inverse frequent set mining problem the new idea is inspired.

Using unknowns to prevent discovery of association rules

This technique depends on the assumption that in order to hide a rule a → b either the support of the item set asb should be decreased below the minimum support threshold (mst) or the confidence of the rule should be decreased below the minimum confidence threshold (mct). Based on the above, we might have the following cases for item set a which is contained in a sensitive association rule:

- A remains sensitive when minsup (a) ≥ mst.
- A is not sensitive when maxsup (a) < mst.
- A is sensitive with a degree of uncertainty when minsup (a) ≤ mst ≤ maxsup (a).

According to [8] the only way to decrease the support of a rule a → b is to replace 1s by ?s for the items in asb in the database. In this process, the minimum support value will be changed while the maximum support value will stay the same. Also, the confidence of a rule a → b can be decreased by replacing both 1s and 0s by?s.

In the same year (2001), dasseni et al. [6] proposed another approach that is based on perturbing support and/or confidence to hide association rules.

II. PROBLEM FORMULATION

To hide certain sensitive information is main problem in data mining so that they cannot be discovered through data mining techniques. The problem is to hide the sensitive and useful data from others by using the transaction database, min_sup_thr.(mst), min_conf_thr.(mct) , a set of strong rules(above the threshould value) is given with the set of sensitive items, how can we modify the database such that using the same mst and mct, in the modified database satisfies all the constraints with in the set of strong rules: 1) no sensitive rule, 2) no lost rule, and 3) no false rule? In association rule mining, we assume that only sensitive item set in transaction database are given and problem is to hide sensitive items so that it cannot be inferred through association rules mining algorithms. More specifically, given a transaction database, a minimum support, a minimum confidence and a set of items to be hidden, the objective is to modify the database such that no association rules containing on the right hand side or left hand side will be discovered.

III. PROPOSED ALGORITHM

The task of mining association rules over market basket data [1] is considered a core knowledge discovery activity. An useful mechanism is provide by association rule mining for item discovering correlations that belonging to customer transactions in a market basket database. Let d be the database of transactions and j = {j1, ..., jn} be the set of items. A transaction t includes one or more items in j . An association rule has the form $x \rightarrow y$, where x and y are non-empty sets of items (i.e. X and y are subsets of j) such that $x \cap y = null$. A set of items is called an itemset, while x is called the antecedent. The support of an item (or itemset) x is the percentage of transactions from d in which that item or itemset occurs in the database the confidence or strength c for an association rule $x \rightarrow y$ is the ratio of the number of transactions that contain x or y to the number of transactions that contain x.

Proposed Algorithm

- **Step 1: transaction data base, rule data base, mct (minimum confidence threshold) are the inputs.**
- **Step 2: enter the sensitive element**
- **Step 3: find all those rules in the rule data base which contains sensitive element on the rhs & whose confidence is greater than the mct.**
- **Step 4: for each rule which contains a sensitive item on rhs repeat step 4**
- **Step 5: while the data set is not empty**
- **Find all those transactions where sensitive item = 1 and lhs = 1**
- **Then put sensitive item = 0 in all those transactions. In this way, the confidence will become less than the mct (minimum confidence threshold)**
- **Step 6: exit**

IV. RESULT ANALYSIS

A data set

Let assume that a data set of assemble part of computer is given

TABLE : SAMPLE DATA SET

T-id	processor ram mouse	bitmap
t1	processor, ram	110
t2	processor, ram, mouse	111
t3	processor, ram, mouse	111
t4	processor	100
t5	processor, mouse	101

Association rules	**confidence**
Processor→ ram	60%
Processor →mouse	60%
Ram →processor	100%
Mouse→ processor	100%

Let the minimum support of the above transaction data set is 33% and minimum confidence of rule is 55%. Suppose that the rule processor→ ram need to hide with sensitive item ram. The rule cannot be hide by previous algorithm as the increasing support of left hand side the transaction table is become as it is as show the support of left hand side is already enough, and we cannot further increase it.

TABLE : MODIFIED DATA SET BY ISL

T-id	processor ram mouse	bitmap
t1	processor, ram	110
t2	processor, ram, mouse	111
t3	processor, ram, mouse	111
t4	processor	100
t5	processor, mouse	101

By Proposed Approach:

Suppose that the item processor need to be hide, for this, first take rules in which processor is in rhs. The rule is ram →processor and mouse→ processor, the confidence is greater in both the rule. Search in transaction data base by taking the rule mouse→ processor first. And select the transaction which supports both mouse and processor i.e., the t2, t3, t5 are four transactions with processor = ram = 1. Now on the place of item processor put 0 in all the four transactions in transaction table. After modifying the transaction, here given the

TABLE: MODIFIED DATA SET BY ISL

T-id	processor ram mouse	bitmap
t1	processor, ram	010
t2	processor, ram, mouse	011
t3	processor, ram, mouse	011
t4	processor	000
t5	processor, mouse	001

Now 0% is the confidence of rule mouse–>processor after calculation, it is which is less than minimum confidence so now this rule is hidden.
Now taking rule the rule is ram →processor, and calculating the confidence of the rule which is already below the minimum threshold, so again we are successful to hiding the rule.

. A data set
Suppose there is a database of transactions as below:

Tid	items
t1	abd
t2	b
t3	acd
t4	ab
t5	abd

Fig 4.1: a data set

One has also given a mst of 60% and a mct of 70%. One can see four association rules can be found as below

A → b (60%, 75%)
B → a (60%, 75%)
A → d (60%, 75%)
D → a (60%, 100%)

Now there is a need to hide d and b.

Previous methods:
One can see that by simple **isl** algorithm if someone want to hide d and b, then the transaction t2 can be check after modification from b to bd (i.e. From 0100 to 0101).but still isl cannot hide the rule d → a. It can be see with following example

Tid	items	bit map
t1	abd	1101
t2	b	0100
t3	acd	1011
t4	ab	1100
t5	abd	1101

(rule d → a hide by isl approach)

T-id	items	bit -map
t1	abd	1101
t2	b	0101
t3	acd	1011
t4	ab	1100
t5	abd	1101

So by above explanation it is clear that rule d → a can not be hidden by isl approach because by modifying t2 from b to bd (i.e. From 0100 to 0101) rule d → a will have support and confidence 60% and 75% respectively.

By dsr approach:

T-id	items	bit -map
t1	abd	1101
t2	b	0100
t3	acd	1011
t4	ab	1100
t5	abd	1101

(rule d → a hide by dsl approach)

T-id	items	bit- map
t1	abd	***0101***
t2	b	0100
t3	acd	1011
t4	ab	1100
t5	abd	1101

Now the rule d → a is hidden by dsr technique as its support 40% its confidence is now 60% , but as a result the rule a → d is also hidden as a side effect.

Result analysis of proposed algorithm 2:

A data set

Suppose there is a database of transactions as below:

Tid	items
t1	abc
t2	abc
t3	abc
t4	ab
t5	a
t6	ac

Fig 4.2: a data set

Suppose mct is 50%.

Tid	abc
t1	111
t2	111
t3	111
t4	110
t5	100
t6	101

The all possible rules with confidences are:

a–>b(66.66%) ,
a–>c (66.66%),
b–>a(100%),
b–>c (75%),
c–>a(100%),
c–>b (75%).

By hybrid approach and proposed algorithm 2:

Suppose that the item a need to be hide , for this, first take rules in which a is in rhs. The rule are c–>a and b–>a , the confidence is greater in both the rule. Search in transaction data base by taking the rule b–>a first . And select the transaction which supports both b and a i.e., b = a = 1.the t1, t2, t3, t4 are four transactions with a = b = 1. Now on the place of item a put 0 in all the four transactions in transaction table. After modifying the transaction , here given the table 3 as the resultant modified table.

Tid	abc
t1	011
t2	011
t3	011
t4	010
t5	100
t6	101

Now 0% is the confidence of rule b–>a after calculation, it is which is less than minimum confidence so now this rule is hidden. Now search transactions with rule c–>a in which the value of a = c = 1, t6 only transaction which has the value a = c = 1, by putting 0 instead of 1 update transaction in place of
A. Now the confidence of rule c–>a is 0% after calculation, which is less than the minimum confidence so now this rule is hidden. Now those rules are taken in which a is in lhs.

Tid	abc
t1	011
t2	011
t3	011
t4	010
t5	100
t6	001

The rules are a–>b and a–>c but confidence of both the rules is less than minimum confidence so these rules are week rule and its not required to hide. So after hiding item a table 4 shows the modified database . So the data base unnecessarily scans by hybrid algorithm . Because to find the same sensitive item a in lhs it scans the data base .and the item a is already hidden in the data base so that it doesn't make any difference . Proposed algorithm 2 removes this problem of hybrid algorithm.

RESULT COMPARISON:

TABLE : COMPARISON OF ALGORITHMS

ALGORITHM/RULE	HYBRID	PROPOSED ALGORITHM 2
D → A	**NOT HIDE**	**HIDE**
RAM →PROCESSOR	**NOT HIDE**	**HIDE**
MOUSE→ PROCESSOR	**NOT HIDE**	**HIDE**

V. CONCLUSION & FUTURE WORK

Proposed algorithm introduced an effective privacy-preserving technique. It also studied the existing sanitizing algorithms and stated their drawbacks. We showed that previous methods remove more knowledge than necessary. The proposed algorithm also capable to hide the strong rule which is not hide by the previous algorithm as seen in result analysis. Privacy-preserving data mining can be applied in different domains. The focus in this thesis is on the association rule mining domain. The goal of association rule mining is to find (in databases) all patterns based on some hard thresholds, such as the min_ support and the min_confidence . Some patterns that are sensitive nature might need to hide owners of the databases. The degree of sensitivity of item and sensitivity of item decided to help the data owner is decided by the data miner. Nowadays, determining the most effective way to protect sensitive patterns while not hiding non-sensitive ones as a side effect is a crucial research issue.

REFERENCES

I. r. Agrawal, t. Imielinski, and a. Swami. *Mining association rules between sets of items in large databases*. In proceedings of the acm sigmod conference on management of data, pages 207–216, new york, ny, usa, may 1993. Acm press.

II. a. K. Pujari. *Data mining techniques* (book), 2001. University press (india) limited.

III. r. Chen, k. Sivakumar, and h. Kargupta. *Distributed web mining using bayesian networks from multiple data streams*. In n. Cercone, t. Young lin, and x. Wu, editors, proceedings of the 2001 ieee international conference on data mining (icdm'01), pages 75–82, san jose, california, usa, november 2001. Ieee computer society.

IV. s. Goldwasser. *Multi-party computations: past and present*. In proceedings of the 16th annual acm symposium on the principles of distributed computing, pages 1–6, santa barbara, california, usa, 1997. Acm press.

V. m. Atallah, e. Bertino, a. Elmagarmid, m. Ibrahim, and v. Verykios. *Disclosure limitation of sensitive rules*. In proceedings of 1999 ieee knowledge and data engineering exchange workshop (kdex'99 pages 45–52, chicago, illinois usa, november 1999. Ieee computer society.

VI. e. Dasseni, v. S. Verykios, a. K. Elmagarmid, and e. Bertino. *Hiding association rules by using confidence and support*. In i. S. Moskowitz, editor, proceedings of the 4th information hiding workshop, volume 2137, pages 369–383, pittsburg, pa, usa, april 2001. Springer veralg lecture notes in computer science.

VII. s. R. M. Oliveira and o. R. Zaiane. *Algorithms for balancing privacy and knowledge discovery in association rule mining*. In proceedings of the 7th international database engineering and applications symposium (ideas'03), pages 54–65, hong kong, china, july 2003. Ieee computer society.

VIII. y. Saygin, v. S. Verykios, and c. Clifton. *Using unknowns to prevent discovery of association rules*. In acm sigmod record, volume 30(4), pages 45–54, new york, ny, usa, december 2001. Acm press.

IX. m. Kantarcioglu and c. Clifton. *Privacy-preserving distributed mining of association rules on horizontally partitioned data*. In ieee transactions on knowledge and data engineering journal, volume 16(9), pages 1026–1037, piscataway, nj, usa, september 2004. Ieee educational activities department.

X. c. Clifton and d. Marks. *Security and privacy implications of data mining*. In workshop on data mining and knowledge discovery, pages 15–19, montreal, canada, february 1996. University of british columbia, department of computer science.

XI. y. Saygin, v. S. Verykios, and a. K. Elmagarmid. *Privacy preserving association rule mining*. In z. Yanchun, a. Umar, e. Lim, and m. Shan, editors, proceedings of the 12th international workshop on research issues in data engineering: engineering e-commerce/e- business systems (ride'02), pages 151–158, san jose, california, usa, february 2002. Ieee computer society.

XII. w. Du and m. J. Atallah. *Secure multi-party computation problems and their applications: a review and open problems*. In v. Raskin, s. J. Greenwald, b. Timmerman, and d. M. Kienzle, editors, proceedings of the new security paradigms workshop, pages 13–22, cloudcroft, new mexico, usa, september 2001. Acm press.

XIII. j. Vaidya and c. Clifton. *Privacy preserving association rule mining in vertically partitioned data*. In proceedings the 8th acm sigkdd international conference on knowledge discovery and data mining, pages 639–644, edmonton, alberta, canada, july 2002. Acm press.

XIV. y. Lindell and b. Pinkas. *Privacy preserving data mining*. In crypto-00, volume 1880, pages 36–54, santa barbara, california, usa, 2000. Springer verlag lecture notes in computer science.