Computer and Information Science

www.ccsenet.org/journal.html

Vol. 1, No. 3
August 2008

A New Approach to Sensitive Rule Hiding

Dr.K. Duraiswamy
K.S. R. College of Technology
Tiruchengode- 637 209, Tamil Nadu, India
E-mail: kduraiswamy@yahoo.co.in

Dr.D. Manjula
Department of Computer Science and Engineering
Anna University
Chennai, Tamil Nadu, India
E-mail: manju@annauniv.edu

N. Maheswari (Corresponding Author)
P.G. Department of Computer Science
Kongu Arts and Science College
Erode-638 107, Tamil Nadu, India
E-mail: mahii_14@yahoo.com

Abstract

Privacy preserving data mining is a novel research direction in data mining and statistical databases, which has recently been proposed in response to the concerns of preserving personal or sensible information derived from data mining algorithms. There have been two types of privacy proposed concerning data mining. The first type of privacy, called output privacy, is that the data is altered so that the mining result will preserve certain privacy. The second type of privacy, called input privacy, is that the data is manipulated so that the mining result is not affected or minimally affected. For output privacy in hiding association rules, current approaches require hidden rules or patterns to be given in advance. However, to specify hidden rules, entire data mining process needs to be executed. For some applications, only certain sensitive rules that contain sensitive items are required to hide. In this work, an algorithm SRH (Sensitive Rule Hiding) is proposed, to hide the sensitive rules that contain sensitive items, so that sensitive rules containing specified sensitive items on the right hand side of the rule cannot be inferred through association rule mining. Example illustrating the proposed approach is given. The characteristics of the algorithm are discussed.

Keywords: Data Mining, Privacy Preserving, Association Rules, Sensitive Rules, Clustering, Minimum Support, Minimum confidence

1. Introduction

The concept of Privacy-Preserving has recently been proposed in response to the concerns of preserving personal or sensible information derived from data mining algorithms. Successful applications of data mining have been demonstrated in marketing, business, medical analysis, product control, engineering design, bioinformatics and scientific exploration, among others. The current status in data mining research reveals that one of the current technical challenges is the development of techniques that incorporate security and privacy issues. The main reason is that the increasingly popular use of data mining tools has triggered great opportunities in several application areas, which also requires special attention regarding privacy protection. There have been two types of privacy concerning data mining. The first type of privacy, called output privacy, is that the data is minimally altered so that the mining result will preserve certain privacy (Evfimievski, 2002, Oliveira, Zaiane, 2003 a, Oliveira, Zaiane, 2003 b). The second type of privacy, input privacy, is that the data is manipulated so that the mining result is not affected or minimally affected (Dasseni, Verykios, Elmagarmid, Bertino, 2001).

For example, through data mining, one is able to infer sensitive information, including personal information, or even patterns from non-sensitive information or unclassified data. As a motivating example of privacy issue in data mining discussed in (Yi-Hung Wu, Chia-Ming Chiang, and Arbee L.P. Chen, 2007). Consider a supermarket and two breads suppliers A and B. If the transaction database of the supermarket is released, A (or B) can mine the association rules related to his/her breads and apply the rules to the sales promotion and the goods supply. As a result, a supplier is willing to exchange a lower price of goods for the database with the supermarket. From this aspect, it is good for the supermarket to release the database. However, the conclusion can be opposite if a supplier uses the mining methods in a different way. For instance, if A finds the association rules related to B's breads, saying that most customers who buy cheese also buy B's breads, he/she can run a coupon that gives a 10 percent discount when buying A's breads together with cheese. Gradually, the amount of sales on B's breads is down and B cannot give a low price to the supermarket as before. Finally, A monopolizes the bread market and is unwilling to give a low price to the supermarket as before. From this aspect, releasing the database is bad for the supermarket. Therefore, for the supermarket, an effective way to release the database with sensitive rules hidden is required. This leads to the research of sensitive rule hiding.

In this work, the sensitive rules are given and the algorithm SRH is proposed to modify data in database so that sensitive rules containing specified sensitive items on the right hand side of rule cannot be inferred through association rule mining. The proposed algorithm is based on modifying or perturbing the database transactions so that the confidence of the association rules can be reduced.

The rest of the paper is organized as follows. Section 2 gives the view of the previous works. Section 3 presents the statement of the problem. Section 4 presents the proposed algorithm for sensitive rule hiding. Section 5 shows the example of the proposed algorithm. Section 6 analyzes the characteristics of the algorithm. Concluding remarks and future work are described in Section 7.

2. Related Work

In output privacy, given specific rules or patterns to be hidden, many data altering techniques for hiding association, classification and clustering rules have been proposed. For association rules hiding, two basic approaches have been proposed. The first approach (Saygin Y, Verykios V, Clifton C, 2001, Verykios V, Elmagarmid A, Bertino E, Saygin Y, Dasseni E ,2004) hides one rule at a time. It first selects transactions that contain the items in a give rule. It then tries to modify transaction by transaction until the confidence or support of the rule fall below minimum confidence or minimum support. Either removing items from the transaction or inserting new items to the transactions does the modification of transaction. The second approach (Oliveira, Zaiane, 2002 a, Oliveira, Zaiane, 2002 b, Oliveira, Zaiane, 2003 a, Oliveira, Zaiane, 2003 b) deals with groups of restricted patterns or association rules at a time. It first selects the transactions that contain the intersecting patterns of a group of restricted patterns. Depending on the disclosure threshold given by users, it sanitizes a percentage of the selected transactions in order to hide the restricted patterns. However, both the above approaches require hidden rules or patterns been given in advance.

The work presented here differs from the related work in some aspects are as follows: First, correlations among the sensitive rules are considered. Second, avoids the modification in transactions unnecessarily, if the confidence of the sensitive rule gets reduced. Third, alter the transactions in the cluster and finally changes can be updated in the database, which reduces the time period of database updating.

3. Problem Statement

The problem of mining association rules was introduced in (Agrawal, Imielinski, Swami, 1993). Let $I = \{i1, i2, . . . , im\}$ be a set of literals, called items. Given a set of transactions D, where each transaction T in D *is* a set of items such that $T \leq I$, an association rule is an expression $X \Rightarrow Y$ where $X \leq I$, $Y \leq I$, and $X \cap Y = \Phi$. The X and Y are called respectively the body (left hand side) and head (right hand side) of the rule. An example of such a rule is that 90% of customers buy hamburgers also buy Coke. The 90% here is called the confidence of the rule, which means that 90% of transaction that contains X (hamburgers) also contains Y (Coke). The confidence is calculated as $|X\ U\ Y| / |X|$, where $|X|$ is the number of transactions containing X and $|X\ U\ Y|$ is the number of transactions containing both X and Y. The notation U here is not the set union operator. The support of the rule is the percentage of transactions that contain both X and Y, which is calculated as $|XU\ Y| / N$, where N is the number of transactions in D. In other words, the confidence of a rule measures the degree of the correlation between item sets, while the support of a rule measures the significance of the item sets. A typical association rule-mining algorithm first finds all the sets of items that appear frequently enough to be considered significant and then it derives from them the association rules that are strong enough to be considered interesting. The problem of mining association rules is to find all rules that are greater than the user-specified minimum support and minimum confidence.

The objective of data mining is to extract hidden or potentially unknown but interesting rules or patterns from databases. However, the objective of privacy preserving data mining is to hide certain sensitive information so that they cannot be discovered through data mining techniques (Agrawal, Imielinski, Swami, 1993, Evfimievski A, Gehrke J, Srikant R, 2003). In this work, an algorithm SRH (Sensitive Rule Hiding) is proposed, to hide the sensitive rules that contain sensitive items, so that sensitive rules containing specified sensitive items on the right hand side of rule cannot be inferred through association rule mining. More specifically, given a transaction database *D*, a minimum support, a minimum confidence and a set of sensitive items Y, the objective is to minimally modify the database *D* such that no sensitive rules containing sensitive items *Y* on the right hand side of the rule will be discovered.

4. Proposed Algorithm

In order to hide an sensitive rule, $X => Y$, it can be either decrease its supports, (|X|/N or |X U Y |/N), to be smaller than pre-specified minimum support or its confidence (|X U Y |/|*X*|) *to* be smaller than pre-specified minimum confidence. In the transactions containing both *X* and *Y*, to decrease the support of *Y* only, the right hand side of the rule, it would reduce the confidence of the rule. In order to hide sensitive rules, when considering hiding sensitive rules with 2 items, x => z, where z is a sensitive item and x is a single large one item. In theory, association rules may have more specific rules that contain more items, e.g., xY => z, where Y is a large item set. However, for such rule to exist, its confidence must be greater than the confidence of x => z, i.e., conf (xY => z) > conf (x => z) or |xY z| > conf (x => z)∗|xY|. For higher confidence rules, such as conf (x => z) = 1, there will be no more specific rules. In addition, once the more general rule is hidden, the more specific rule might be hidden as well.

The algorithm tries to decrease the support of the right hand side of the rule.

```
Algorithm SHR
Input:
1.        source database D
2.        min support
3.        min conf
4.        sensitive items Y
Output:
a transformed database D', where rules containing Y on RHS will be hidden.
Step 1:
Generate the association rule.
Selecting the Sensitive rules with single antecedent and consequent with the sensitive item in the consequent.
Step 2:
Clustering the rules with the right hand side has the common item and indexing the rules
Step 3:
Check all the rules in the cluster.
for cluster i= 1 to n
{
 do
{

1         select the rule with high confidence
2         find K= mincount
3         |Tr| = no of sensitive rules that are correlated
Such that k<|Tr|
4         Tk = find first k transactions that contains the items in the correlated rules
5         sort Tk in ascending order by the number of items.
6         For i= 1 to k
     {
7           Choose the first transaction t from Tk.
8         Modify t so that y is not supported
```

```
    Find the confidence of every rule
    If confidence of all the rules < min conf
  {
    Update the transactions in the clusters
    Exit
    Else
    Update the transactions in the clusters
   }
 }
    unchecked the rules< minconf
    }while(rules in cluster[i]==checked)
     }
```

Step 4: update database D as transformed database D'.

The algorithm tries to generate the association rule using (Agrawal, Imielinski, Swami, 1993). Then it selects the sensitive rules with the sensitive items in the right hand side. Cluster the rules with the common item in the right hand side of the rule and index the rule using (Oliveira, Zaiane, 2003 a).

The mincount=minimum (MSC, MCC) (Yi-Hung Wu, Chia-Ming Chiang, and Arbee L.P. Chen, 2007)

MSC x->y =count (xUy)-[|D| * min support] +1

MCCC x->y = count (xUy)-[count(x) * min conf]+1

MPCC x->y =[(count(xUy)-count(x)*min conf)/(1-min conf)] + 1,

MCC=minimum (MCCC, MPCC).

Therefore it would take maximum of k no of executions to hide the rule. The rules in every cluster will be hidden.

5. Example

This section shows the example to demonstrate the proposed algorithm to hide the sensitive rules.

The items in database can be represented as a bit vector 1 and 0.(Saygin Y, Verykios V, Clifton C, 2001)

TID	Items	Items
T1	ABC	111
T2	ABC	111
T3	ABC	111
T4	AB	110
T5	A	100
T6	AC	101

Frequent item sets are generated with minimum support 0.50. The following are the frequent item sets and its support value A(1.0), B(0.6),C(0.6), AB(0.6), AC(0.6), BC(0.5), ABC(0.5). Association rules with minimum confidence 0.75 are generated. The rules and the corresponding confidence values are as follows:

C->B0.75, B->C0.75, C->A1.0, B->A1.0, BC->A1.0, AC->B0.75, AB->C0.75, C->AB0.75, and B -> AC0.75.

The item A is considered as sensitive item. The sensitive rule with single antecedent and consequent is C->A, B->A. The rules are clustered and to hide such rules the no. of transactions required is 2. In the first and second transaction the item A will be removed and placed as 0.

Then the database will be updated as

TID	Items	Items
T1	BC	011
T2	BC	011
T3	ABC	111
T4	AB	110

T5	A	100
T6	AC	101

After updating the database the rules C->A, B->A will have the confidence as 0.5, which is less than minimum confidence and are hidden. While hiding the two rules the rules BC->A, C->AB, B -> AC are also hidden and the two rules AB->C and AC->B are lost as side effects.

6. Analysis

This section analyses some of the characteristics of the proposed algorithm. The first characteristic is the time effect. The time taken to scan the database to search the sensitive rules in the database is reduced because of clustering the sensitive rules. The second characteristic is the database effect. The minimum numbers of transactions are modified because of correlation among the sensitive rules. The third characteristic is the efficiency of the algorithm. The database is updated after all the rules are hidden that saves the updating time. The fourth characteristic is the transaction effect. The alteration in the transactions are stopped when the confidence of the sensitive rules are reduced than the minimum confidence.

7. Conclusion

In this work, the database privacy problems caused by data mining technology are discussed and the algorithm for hiding sensitive rules is presented. The proposed algorithm here can automatically hide sensitive rule sets. The previous works does not consider the characteristics that are discussed here. Example illustrating the proposed algorithm is given and the characteristics of the algorithm are analyzed. Further the efficiency of the algorithm will be analyzed and improved by reducing the side effects.

References

Agrawal R, Imielinski T, Swami A (1993). Mining association rules between sets of items in large databases.*In: Proceedings of ACM SIGMOD International Conference on Management of Data*, Washington DC.

Dasseni E, Verykios V, Elmagarmid A, Bertino E (2001). Hiding association rules by using confidence and support. *In: Proceedings of 4th Information Hiding Workshop*, Pittsburgh, PA, pp. 369– 383.

Evfimievski A (2002). Randomization in privacy preserving data mining. *SIGKDD Explorations* 4(2), Issue 2:43–48.

Evfimievski A, Gehrke J, Srikant R (2003). Limiting Privacy Breaches in Privacy Preserving Data Mining. *PODS 2003*, San Diego, CA.

Evfimievski A, Srikant R, Agrawal R, Gehrke J (2002). Privacy preserving mining of association rules. *In: Proceedings of the 8th ACMSIGKDD Int'l Conference on Knowledge Discovery and Data Mining*, Edmonton, Canada.

Oliveira S, Zaiane O (2002 a). Privacy preserving frequent item set mining. *In: Proceedings of IEEE International Conference on Data Mining*, November 2002, pp. 43–54.

Oliveira S, Zaiane O (2002 b) A framework for enforcing privacy in mining frequent patterns. *Technical report, TR02-13*, Computer Science Department, University of Alberta, Canada.

Oliveira S, Zaiane O (2003 a). Algorithms for balancing privacy and knowledge discovery in association rule mining. *In: Proceedings of 7th International Database Engineering and Applications Symposium (IDEAS03)*, Hong Kong.

Oliveira S, Zaiane O (2003 b). Protecting sensitive knowledge by data sanitization. *In: Proceedings of IEEE International Conference on Data Mining.*

Saygin Y, Verykios V, Clifton C (2001). Using unknowns to prevent discovery of association rules. *SIGMOND Record* 30(4): 45– 54.

Verykios V, Elmagarmid A, Bertino E, Saygin Y, Dasseni E (2004). Association Rules Hiding. *IEEE Trans Knowedge and Data Eng* 16(4): 434– 447.

Yi-Hung Wu, Chia-Ming Chiang, and Arbee L.P. Chen, (2007). Hiding Sensitive Association Rules with Limited Side Effects, IEEE *Trans Knowledge and Data Eng*, Vol. 19, No. 1.