Computer and Information Science
www.ccsenet.org/journal.html
Vol. 1, No. 2
May 2008

Identification of Sensitive Items in Privacy Preserving - Association Rule Mining

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

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

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. For example, through data mining, sensible information such as private information or patterns may be inferred from non-sensible information or unclassified data. As large repositories of data contain confidential rules that must be protected before published, association rule hiding becomes one of important privacy preserving data mining problems. There have been two types of privacy concerning data mining. Output privacy tries to hide the mining results by minimally altering the data. Input privacy tries to manipulate the data so that the mining result is not affected or minimally affected. For some applications certain sensitive predictive rules are hidden that contain given sensitive items. To identify the sensitive items an algorithm SENSIDENT is proposed. The results of the work have been given.

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

1. Introduction

In recent years, data mining or knowledge discovery in databases has developed into an important technology of identifying patterns and trends from large quantities of data. 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. While all of the applications of data mining can benefit commercial, social and human activities, there is also a negative side to this technology: the threat to data privacy. 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. The concept of privacy preserving data mining has recently been proposed in response to the concerns of preserving privacy information from data mining algorithms (Agrawal, Srikant, 2000). 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 (Dasseni, Verykios, Elmagarmid, Bertino, 2001, 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 (Evfimievski, 2002).

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 (Clifton, Marks, 1996). Suppose we (as purchasing directors of BigMart, a large supermarket chain) are negotiating a deal with the Dedtrees paper company. They offer to us a reduced price if we agree to give them access to our database of customer purchases. We accept this deal and Dedtrees starts mining our data. By using association rule mining tool, they find that people who purchase skim milk also purchase Green paper. Dedtrees now runs a coupon marketing campaign “50 cents off skim milk when you buy Dedtrees products,” cutting heavily into the sales of Green paper, who increase prices to us based on the lower sales. When we next go to negotiate with Dedtrees, we find that with reduced competition, they are unwilling to offer us as low a price, and we start to lose business to our competitors. This example indicates the need to prevent disclosure not only of confidential personal information from summarized or aggregated data, but also to prevent data mining techniques from discovering sensitive knowledge which is not even known to the database owners. So the highest sales product can be easily identified using this rule generation. This makes the supplier of the product to demand or hike the product rate. Such products are considered to be sensitive. In the previous works (Dasseni, Verykios, Elmagarmid, Bertino, 2001, Oliveira, Zaiane, 2003 a) of association rule hiding the sensitive items are manually selected by the user. In the proposed work, the selection of sensitive items would require data mining process to be executed first. Based on the discovered rules sensitive items are selected. While generating the association rules the sensitive items has to be hidden. To identify the sensitive items an algorithm SENSIDENT is proposed.

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 selecting sensitive items. Section 5 shows the example of the proposed algorithm. Section 6 shows the experimental results of the performance of the algorithm. Concluding remarks and future work are described in Section 7.

2. Related Work

In (Dasseni, Verykios, Elmagarmid, Bertino, 2001) the authors selected the rules manually, in order to hide the sensitive knowledge by reducing the support and confidence of the rules. In (Oliveira, Zaiane, 2003 a), in order to protect the sensitive knowledge in association rule mining, the sensitive rules are selected. In (Oliveira, Zaiane, 2003 b) authors introduced a heuristic approach to hide restrictive association rules that requires the victim item. The victim item has to be identified manually and it has to be removed from its transactions. In (Wenliang Du, Zhijun Zhan, 2003), the authors selected the sensitive attributes and protect the attributes by randomization process.

Instead of selecting the sensitive rules and sensitive data manually, the proposed algorithm helps to identify the sensitive data.

3. Problem Statement

The problem of mining association rules was introduced in (Agrawal, Imielinski, Swami, 1993). Let $I = \{i1, i2, \ldots, im\}$ 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

However, the objective of privacy preserving data mining is to hide certain sensitive information so that they cannot be discovered through data mining techniques (Oliveira, Zaiane, 2003 a, Oliveira, Zaiane, 2003 b). In this work, sensitive information is identified using the proposed algorithm.

4. Proposed Algorithm

In order to select the sensitive item(s) the frequent item sets and the association rules are generated based on the given support and confidence values (Agrawal, Imielinski, Swami, 1993). By selecting the consequent of the rule, the sensitive item can be identified. One or more items can be selected as sensitive items. The algorithm SENSIDENT is described below:

Input:

(1) input database D

(2) minimum support

(3) minimum confidence

Output: sensitive item(s) to be hidden.

Algorithm:

```
1. Find the frequent item sets and generate the association rules from D using minimum support and minimum
   confidence.
2. Identify the rules with single antecedent and consequent.
     (i.e.) x -> y
3. Sort the rules in descending order based on the confidence values.
4. Select the rules with the highest confidence.
5. Count the frequency of the consequent (y) in the rules.
6. Sort the count in descending order.
7. Choose the highest two different counts c1, c2
8. If c1-c2 < threshold value
             If more than one consequent with the same count
                       Select the corresponding consequents
             else
                       Select the corresponding consequents
   else
             If one (or) more consequent with the same count c1
                       Select the consequent(s)
9. Selected consequent(s) will be the sensitive item(s).
```

The threshold value is the user specified value, the minimum difference between the two counts. The selected sensitive items are further used to hide sensitive rules.

The algorithm SENSIDENT first tries to find the frequent item sets and the association rules using the Apriori algorithm (Agrawal, Imielinski, Swami, 1993). Rules are sorted and the frequency of the consequent items are considered and compared with the threshold value. The selected consequent items are finally considered as sensitive items.

5. Example

This section shows the example to demonstrate the proposed algorithm to select the sensitive items .

TID	Items
T1	ABC
T2	ABC
T3	ABC
T4	AB
T5	A
T6	AC

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 rules with highest confidence are

C->A	1.0
B->A	1.0

From the above rules the item A is selected as the sensitive item with the threshold value 2.

6. Experimental Results

6.1 Methodology

In order to better understand the characteristics of the proposed algorithm numerically, a series of experiments is performed to measure various characteristics. The experiments are conducted on a PC, with Pentium IV Processor with 512 MB of RAM running on Windows XP Operating System. To measure the effectiveness, the dataset used are mushroom dataset from UCI machine learning repository (www.ics.uci.edu/mlearn), the two synthetic datasets are csc and c20d10.The mushroom dataset contains 128 different items, with 8124 transactions. The c20d10 dataset has 2000 transactions and 386 items. The csc contains 298 transactions and 88 items.

The Apriori algorithm is used to generate the frequent item sets and the association rules. The minimum support and minimum confidence, which are used to generate the frequent item sets and the rules, are given by the user. From that the rules with single antecedent and consequent are identified. Sorting is performed to find the highest confidence. The two consequents with highest counts are compared with the user threshold value. The threshold value is given to select the sensitive items with minimum difference in their counts.

6.2 Performance Evaluation

For each dataset, the frequent item sets and the association rules are generated with minimum support 0.40 and minimum confidence 0.50. Figure 1 shows the time effects of the various sizes of the data sets, in frequent item set generation. The mushroom dataset produces the frequent item sets in 14,843 ms, c20d10 in 3969 ms and csc in 79ms. The time can be varied with various minimum support and minimum confidence values. Figure 2 shows the selection of the rules with single antecedent and consequent. The mushroom data set is used for this effect. Under the minimum supports 0.4, 0.5 and 0.6 with the minimum confidence 0.75 the rules generated are 31, 19 and 7. The count of rule rapidly decreases with the increase in minimum support value. Figure 3 shows the count of sensitive items with various threshold value. The mushroom data set is used for this effect. With the threshold value 1, 5 and 10, under the minimum support value 0.5 and the minimum confidence value 0.75 the sensitive items identified are 1, 1 and 3. The counts of sensitive items are increased for higher values of threshold value.

7. Conclusion and Future work

In this work it is discussed about the objective of privacy preserving data mining, to hide certain sensitive information so that they cannot be discovered through data mining techniques. In the previous works for association rule hiding the sensitive information is selected manually.

The proposed algorithm is used to identify the sensitive information using the set of association rules. Results illustrating the algorithm are given. In future the sensitive items can be identified using the classification, clustering and regression techniques. Also the proposed work can be integrated with association rule hiding.

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

Agrawal R, Srikant R (2000) Privacy preserving data mining. *In ACMSIGMOD Conference on Management Of Data*, Dallas, Texas, pp 439–4501.

Clifton C, Marks D (1996) Security and privacy implications of data mining. *In: SIGMOD Workshop on Research Issues on Data Mining and knowledge Discovery.*

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.

Oliveira S, Zaiane O (2003 a) Algorithms for balancing privacy and knowledge discovery in association rule

mining. *In: Proceedings of7th 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.*

Wenliang Du , Zhijun Zhan(2003)Using Randomized Response Techniques for Privacy Preserving Data mining. *SIGKDD'03*

www.ics.uci.edu/mlearn

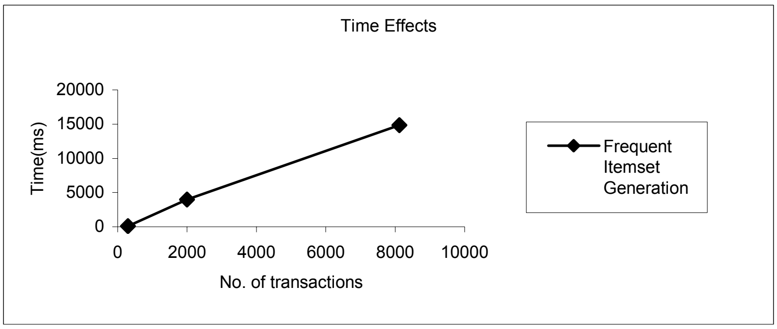

Figure 1. Time effects

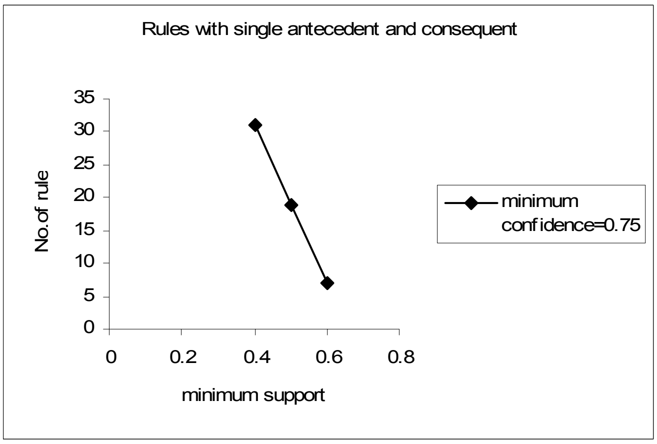

Figure 2. Rules with single antecedent and consequent

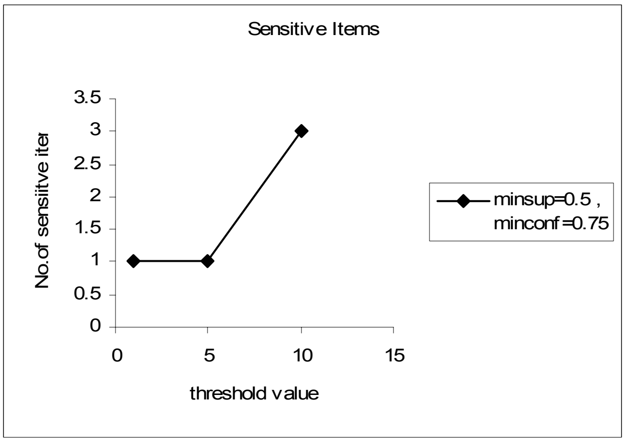

Figure 3. Sensitive items