A Novel Data Mining Approach for Information Hiding

Shikha Sharma
PG Student,
Shri Vaishnav Institute of Technology and Science, Indore (MP), India

Prof. Pooja Jain
Assistant Professor, Department of Computer Science and Engineering
Shri Vaishnav Institute of Technology and Science, Indore (MP), India

Abstract

Data mining services require accurate input data for their results to be meaningful, but privacy concerns may influence users to provide spurious information. To preserve client privacy in the data mining process, a variety of techniques based on random perturbation of data records have been proposed recently. One known fact which is very important in data mining is discovering the association rules from database of transactions where each transaction consists of set of items. Two important terms support and confidence are associated with each of the association rule. Actually any rule is called as sensitive if its disclosure risk is above a certain privacy threshold. Sometimes we do not want to disclose sensitive rules to the public because of confidentiality purposes. There are many approaches to hide certain association rules which take the support and confidence as a base for algorithms ([1, 2, 6] and many more). Our approach is a modification of ISL (increase support of LHS) and DSR (decrease support of RHS) and has some modifications so that it hides any desired association rule as previous work sometimes can not.

Our work has the basis of reduction of support and confidence of sensitive rules but in our work we are not editing or disturbing the given database of transactions directly. Our algorithm use some modified definition of support and confidence so that it would hide any desired sensitive association rule without any side effect. Actually we are using the same method (as previously used method) of getting association rules but we are modifying the definitions of support and confidence.

1 Data Mining

Data mining is the process of extracting hidden patterns from data. As more data is gathered, with the amount of data doubling every three years, data mining is becoming an increasingly important tool to transform this data into knowledge. It is commonly used in a wide range of applications, such as marketing, fraud detection and scientific discovery. Data mining can be applied to data sets of any size, and while it can be used to uncover hidden patterns, it cannot uncover patterns which are not already present in the data set.

Data mining extracts novel and useful knowledge from data and has become an effective analysis and decision means in organization.

Data sharing can bring a lot of advantages for research and business collaboration. However, large repositories of data contain private data and sensitive rules that must be preserved before published. Motivated by the multiple conflicting requirements of data sharing, privacy preserving and knowledge discovery, privacy preserving data mining (PPDM) has become a research hotspot in data mining and database security fields.

Two problems are addressed in PPDM: one is the protection of private data; another is the protection of sensitive rules (knowledge) contained in the data.
The former settles how to get normal mining results when private data cannot be accessed accurately; the latter settles how to protect sensitive rules contained in the data from being discovered, while non-sensitive rules can still be mined normally. The latter problem is called knowledge hiding in database (KHD) which is opposite to knowledge discovery in database (KDD). And association rule hiding problem we focus is one of problems in KHD

2 Association Rules

In data mining, association rule learning is a popular and well researched method for discovering interesting relations between variables in large databases. Piatetsky-Shapiro describes analyzing and presenting strong rules discovered in databases using different measures of interestingness. Based on the concept of strong rules, Agrawal et al introduced association rules for discovering regularities between products in large scale transaction data recorded by point-of-sale (POS) systems in supermarkets.

For example, the rule

\[ \{\text{onions, potatoes}\} \Rightarrow \{\text{beef}\} \]

Found in the sales data of a supermarket would indicate that if a customer buys onions and potatoes together, he or she is likely to also buy beef. Such information can be used as the basis for decisions about marketing activities such as, e.g., promotional pricing or product placements. In addition to the above example from market basket analysis association rules are employed today in many application areas including Web usage mining, intrusion detection and bioinformatics.

3 Backgrounds and Related Work

The security impact of DM is analyzed in [12] and some possible approaches to the problem of inference and discovery of sensitive knowledge in a data mining context are suggested. The proposed strategies include fuzzyfying and augmenting the source database and also limiting the access to the source database by releasing only samples of the original data. Clifton [13] adopts the last approach as he studies the correlation between the amount of released data and the significance of the patterns which are discovered. He also shows how to determine the sample size in such a way that data mining tools cannot obtain reliable results.

Clifton and Marks in [12] also recognize the necessity of analyzing the various data mining algorithms in order to increase the efficiency of any adopted strategy that deals with disclosure limitation of sensitive data and knowledge. The solution proposed by Clifton in [13] is independent from any specific data mining technique; other researchers [14], [15] propose solutions that prevent disclosure of confidential information for specific data mining algorithms such as association rule mining and classification rule mining.

Classification mining algorithms may use sensitive data to rank objects; each group of objects has a description given by a combination of non sensitive attributes. The sets of descriptions, obtained for a certain value of the sensitive attribute, are referred to as description space. For Decision-Region based algorithms, the description space generated by each value of the sensitive attribute can be determined a priori. The authors in [8] first identify two major criteria which can be used to assess the output of a classification inference system and then they use these criteria, in the context of Decision-Region based algorithms, to inspect and to modify, if necessary, the description of a sensitive object so that they can be sure that it is not sensitive.

There is a large amount of work related to association rule hiding. Maximum researchers have worked on the basis of reducing the support and confidence of sensitive association rules ([1, 2, 6]). ISL and DSR are the common approaches used to hide the sensitive rules. Actually any given specific rules to be hidden, many approaches for hiding association, classification and clustering rules have been proposed. Some of the researchers have used data perturbation techniques ([5]) to modify the confidential data values in such a way that the approximate data mining results could be obtained from the modified version of the database. Some researchers also recognize the necessity of analyzing the various data mining algorithms in order to
increase the efficiency of any adopted strategy that deals with disclosure limitation of sensitive data and knowledge. Also disclosure limitation of sensitive knowledge by data mining algorithms, based on the retrieval of association rules, has been recently investigated. Our work also has the basis of reduction of support and confidence of sensitive rules but in our method we are using some modified terms and some new variable to do the job. Also our work specifies that we can hide any given association rule, as some of the previous work can not.

4. Problem Statement

Let

\[ I = \{i_1, i_2, \ldots, i_n\} \]

be a set of \( n \) binary attributes called items.

Let

\[ D = \{t_1, t_2, \ldots, t_m\} \]

d be a set of transactions called the database.

Each transaction in \( D \) has a unique transaction ID and contains a subset of the items in \( I \). A rule is defined as an implication of the form

\[ X \Rightarrow Y \]

where

\[ X, Y \subseteq I \text{ and } X \cap Y = \emptyset. \]

The sets of items (for short item sets) \( X \) and \( Y \) are called antecedent (left-hand-side or LHS) and consequent (right-hand-side or RHS) of the rule.

The support \( \text{supp}(X) \) of an item set \( X \) is defined as the proportion of transactions in the data set which contain the item set

Confidence can be interpreted as an estimate of the probability \( P(Y \mid X) \), the probability of finding the RHS of the rule in transactions under the condition that these transactions also contain the LHS

An association rule is an implication of the form \( X \Rightarrow Y \), where \( X \subseteq I \), \( Y \subseteq I \) and \( X \cap Y = \emptyset \). We say the rule \( X \Rightarrow Y \) holds in the database \( D \) with confidence \( c \) if \[ \frac{|X \cup Y|}{|X|} \geq c \]. We also say the rule \( X \Rightarrow Y \) has support \( s \) if \[ \frac{|X \cup Y|}{|D|} \geq s \]. Note while the support is a measure of the frequency of a rule, the confidence is a measure of the strength of the relation between sets of items. The well-known association rule mining problem aims to find all significant association rules. A rule is significant if its support and confidence is no less than the user specified minimum support threshold (MST) and minimum confidence threshold (MCT). To find the significant rules, an association rule mining algorithm first finds all the frequent itemsets and then derives the association rules from them. On the contrary, the association rule hiding problem aims to prevent some of these rules, which we refer to as “sensitive rules”, from being mined. In our approach we are computing confidence and support as follows

\[ \text{Confidence}(X \Rightarrow Y) = \frac{|X \cup Y|}{|X| + \text{counter of rule}} \]

\[ \text{Support}(X \Rightarrow Y) = \frac{|X \cup Y|}{|N| + \text{counter of rule}} \]

5 Proposed Approach

To hide any specified association rule \( X \Rightarrow Y \) our algorithm works on the basis of confidence \( (X \Rightarrow Y) \) and support \( (X \Rightarrow Y) \) as discussed in previous section. To hide the rule \( X \Rightarrow Y \) (containing sensitive element \( X \) on LHS), our algorithm repeatedly increases the hiding counter of the rule \( X \Rightarrow Y \) until confidence \( (X \Rightarrow Y) \) goes below a minimum specified threshold confidence (MCT). As the confidence \( (X \Rightarrow Y) \) goes below MCT (minimum specified confidence threshold), rule \( X \Rightarrow Y \) is hidden i.e. it will not be discovered through data mining algorithm.

Input:

www.cirworld.com
1: A source database D.
2: MST (Minimum Support Threshold).
3: MCT (Minimum Confidence Threshold).
4: A set of sensitive items X.
5: A set of initialization variables for all rules (which are initially set to zero).
6: Confidence(X→Y) = (X U Y)/(X+ counter of rule)
   Support(X→Y) = (X U Y)/(N+ counter of rule)

Process:
To hide the an association rule containing sensitive element X on LHS, our algorithm repeatedly increases the hiding counter of that rule until confidence goes below a minimum specified threshold confidence (MCT). As the confidence of that rule goes below MCT (minimum specified confidence threshold), rule is hidden i.e. it will not be discovered through data mining algorithm.

Output:
Output the rules which do not contain sensitive elements on the left hand side.

New Procedure :-
Initially set the sv’s of all the rules equal to 0.
   Ilcheck for all sensitive elements.
   for each x in X where x belongs to X
      {  
      // Now check all the rules containing sensitive element x.
      for each rule R which contain x on LHS
         {  
         // IlCheck whether Modified confidence of the rule Ilgoes below MCT or not.
         sv(R) = X-(X U Y *100 / MCT)
         while (Modifiedconfidence (R >= MCT))
            // increase the sv of rule R by 1
            
            {  
            sv(R)=sv(R)+1
            }  
        }  
    }  
End of procedure

Example
Suppose we have given a database of transactions as below

<table>
<thead>
<tr>
<th>TID</th>
<th>Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1</td>
<td>ABD</td>
</tr>
<tr>
<td>T2</td>
<td>B</td>
</tr>
<tr>
<td>T3</td>
<td>ACD</td>
</tr>
<tr>
<td>T4</td>
<td>AB</td>
</tr>
<tr>
<td>T5</td>
<td>ABD</td>
</tr>
</tbody>
</table>

We have also given a MST of 60% and a MCT of 70%.

We can see four association rules can be found as below:

- A \rightarrow B (60%, 75%)
- B \rightarrow A (60%, 75%)
- A \rightarrow D (60%, 75%)
- D \rightarrow A (60%, 100%)

Now we have to hide D and B.

By previous methods: We can see that simple by ISL algorithm if we want to hide D and B, we check it by modifying the transaction T2 from B to BD (i.e. from 0100 to 0101) we can not hide the rule D \Rightarrow A.

(Taking Hiding D \Rightarrow A by ISL approach)

<table>
<thead>
<tr>
<th>T1</th>
<th>ABD</th>
<th>1101</th>
</tr>
</thead>
<tbody>
<tr>
<td>T2</td>
<td>B</td>
<td>0100</td>
</tr>
<tr>
<td>T3</td>
<td>ACD</td>
<td>1011</td>
</tr>
<tr>
<td>T4</td>
<td>AB</td>
<td>1100</td>
</tr>
<tr>
<td>T5</td>
<td>ABD</td>
<td>1101</td>
</tr>
</tbody>
</table>

So by above explanation we can see that rule D \Rightarrow A can not be hidden by ISL approach because by modifying T2 from B to BD (i.e., from 0100 to 0101) rule D \Rightarrow A will have support and confidence 60% and 75% respectively. Now we will check it by DSR approach....
(Hiding $D \Rightarrow A$ by DSR approach)

T1   ABD   0101
T2   B     0100
T3   ACD   1011
T4   AB    1100
T5   ABD   1101

We see by DSL approach rule $D \Rightarrow A$ is hidden as its support and confidence is now 40% and 66% respectively, but as a side effect the rule $A \Rightarrow D$ is also hidden. Similarly we can check same is the condition for $B \Rightarrow A$.

**Our Approach:**

T1   ABD
T2   B
T3   ACD
T4   AB
T5   ABD

<table>
<thead>
<tr>
<th>M support , M confidence, SV</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A \Rightarrow B$</td>
</tr>
<tr>
<td>$B \Rightarrow A$</td>
</tr>
<tr>
<td>$A \Rightarrow D$</td>
</tr>
<tr>
<td>$D \Rightarrow A$</td>
</tr>
</tbody>
</table>

*First we hide B*

T1   ABD
T2   B
T3   ACD
T4   AB
T5   ABD

<table>
<thead>
<tr>
<th>M support , M confidence, SV</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A \Rightarrow B$</td>
</tr>
<tr>
<td>$B \Rightarrow A$</td>
</tr>
<tr>
<td>$A \Rightarrow D$</td>
</tr>
<tr>
<td>$D \Rightarrow A$</td>
</tr>
</tbody>
</table>

*Now we hide D*

T1   ABD
T2   B
T3   ACD
So we clearly see that our approach is hiding all the given sensitive rules successfully without any side effect.

6 Analysis and Conclusion

As from our example we see that our approach is better in the way that it hides any rule which can not be hidden by some of the previous works. We see in the example that proposed method is hiding the given association rules (with sensitive items on the left hand side of the rule) without any side effect. Our algorithm is also simpler in the sense that we have to do only one step of modification as we are only incrementing the hiding counter each time (to decrease the confidence of sensitive rule) rather then checking all transactions again and again and ordering them in increasing or decreasing order as we had to do in some of the previous works (which work on the basis of reducing the support and confidence of the sensitive association

Expected Contributions

The expected contributions of my dissertation are:
1) It will provide an effective association rule hiding method.
2) It will provide database security administrators with a credible association rule mining tool protecting both the private data and confidential rules contained in the data.

References


[6] Vi-Hung Wu, Chia-Ming Chiang, and Arbee L.P. Chen, Senior Member, IEEE Computer Society
Hiding Sensitive Association Rules with Limited Side Effects IEEE TRANSACTIONS ON
KNOWLEDGE AND DATA ENGINEERING, VOL. 19, NO.1, JANUARY 2007

Databases", In Proceedings of ACM SIGMOD International Conference on Management of Data

Rule Mining", Proceedings of 71 th International Database Engineering and Applications SYmposium
(IDEAS03), Hong Kong, July 2003.

IEEE Transactions on Knowledge and Data Engineering, 2007,19(1):29-42.

[10] Fienberg, S. and Slavkovic, A. Preserving the confidentiality of categorical statistical data bases when

Fovino Loredana Parasiliti Provenza, Yucel Saygin, Yannis Theodoridis SIGMOD Record, Vol. 33, No.1,

Workshop Data Mining and Knowledge Discovery, 1996.


[14] T. Johnsten and V.V. Raghavan, “Impact of Decision-Region Based Classification Mining Algorithms