A Quantitative and Qualitative ANALYSIS of Blocking in Association Rule Hiding

Emmanuel D. Pontikakis
Dept. of Computer Engineering and Informatics, University of Patras
Research and Academic Computer Technology Institute, Athens, GR
pontikak@ceid.upatras.gr

Achilleas A. Tsitsonis
Dept. of Computer Engineering and Informatics, University of Patras
Research and Academic Computer Technology Institute, Athens, GR
tsitson@ceid.upatras.gr

Vassilios S. Verykios
Dept. of Computer and Comm. Engineering, University of Thessaly
Research and Academic Computer Technology Institute, Athens, GR
verykios@cti.gr

Yannis Theodoridis
Dept. of Informatics, University of Piraeus, Piraeus
Research and Academic Computer Technology Institute, Athens, GR
ytheod@cti.gr

Liwu Chang
Naval Research Laboratory, Washington DC, USA
lchang@itd.nrl.navy.mil

ABSTRACT[1]

Data mining provides the opportunity to extract useful information from large databases. Various techniques have been proposed in this context in order to extract this information in the most efficient way. However, efficiency is not our only concern in this study. The security and privacy issues over the extracted knowledge must be seriously considered as well. By taking this into consideration, we study the procedure of hiding sensitive association rules in binary data sets by blocking some data values and we present an algorithm for solving this problem. We also provide a fuzzification of the support and the confidence of an association rule in order to accommodate for the existence of blocked/unknown values. In addition, we quantitatively compare the proposed algorithm with other already published algorithms by running experiments on binary data sets, and we also qualitatively compare the efficiency of the proposed algorithm in hiding association rules. We utilize the notion of border rules, by putting weights in each rule, and we use effective data structures for the representation of the rules so as (a) to minimize the side effects created by the hiding process and (b) to speed up the selection of the victim transactions. Finally, we discuss the advantages and the limitations of blocking.

Categories & Subject Descriptors

Data Mining

General Terms

Algorithms, Security, Theory

WPES'04, October 28, 2004, Washington, DC, USA.

Keywords

Data Mining, Privacy, Association Rules Hiding, Blocking Technique.

1. ASSOCIATION RULE HIDING PROCESS

For a database *D,* a user mines the database to find association rules. We call the set of these rules *R*. Then, the user will select to hide a subset $R_H \subseteq R$ that she considers to be sensitive. A subset of rules is considered as sensitive if a certain rule in this subset should not be made public, either because this is enforced by a privacy policy or because if such a rule is disclosed we may provide our competitors with a business advantage. The sensitivity is not formally defined in this paper, but it is associated with the decrease of the support or the confidence of a rule *R*. A sensitive rule in R_H can be hidden by decreasing its confidence or by decreasing its support by a *Safety Margin* threshold (*SM*) below the Minimum Confidence Threshold (MCT) or the Minimum Support Threshold (MST), correspondingly. After the hiding process, the sanitized database *D* does not deduce the R_H and it is called D_M.

2. BLOCKING TECHNIQUE

Let us assume that we want to hide the sensitive rule *R* $(I_L \Rightarrow I_R)$ $conf(R) = \frac{\sup(I_L \cup I_R)}{\sup(I_L)}$ and $\sup(R)=\sup(I_L \cup I_R)$.

By blocking 1's, we can either reduce the *minconf(R)* below the MCT or the *minsup(R)* below the MST [1], [2]. When we want to reduce the support, we select some items of I_L or I_R and block them (replace 1's with ?'s) from the transactions that support the sensitive rule. If we select those items from the I_R then both the *minsup(R)* and the *minconf(R)* will be decreased. Otherwise, if we select items from I_L to block, the *minconf(R)* may not be decreased because both the numerator and the denominator are decreased. So, it is more effective to block items from the I_R, and by doing that, we reduce both the minimum support and the minimum confidence of the sensitive rule.

On the other hand, by blocking 0's we decrease *minconf(R)* by selecting transactions that partially supports R and replacing 0's with ?'s. Transactions that partially support R are those transactions in which exactly one item of I_L is 0, and at the same time at least one item of I_R must be 0. If we block the 0 item in I_L (replace 0 with a ?), then the minimum confidence of R will be reduced because the denominator of *conf(R)* will be increased while the numerator will remain the same.

3. BLOCKING ALGORITHM (BA)

This algorithm adds uncertainty in the database by adding question marks in a way that the database can be usable by a data miner that receives the database and at the same time an adversary cannot infer the sensitive rules that BA will hide. The algorithm aims to achieve the following two goals:

a) Reduce the minimum confidence of sensitive rules below (MCT-SM).
b) Do not reduce the minimum confidence of non-sensitive rules.

If the adversary finds the maximum confidence of all the rules in the modified database, she will find many new ghost rules that did not exist in the initial database so the adversary cannot assume with certainty which of the rules that have maximum confidence above MCT were the sensitive rules. On the other hand, a data miner who wants to find useful information from the database can find the minimum confidence of all the rules, excluding in that way the sensitive rules from his information.

4. EXPERIMENTAL RESULTS

The following rules were chosen randomly for hiding in a sample database with 50 items:

Rules	*Confidence*
1 ⇒ 33	80.6%
34 ⇒ 3	68.9%
1 33 ⇒ 48	100%
6 22 ⇒ 11	78.1%
7 41 ⇒ 3	98%

4.1 Rules changed in the database

In Figure 1 the side effects of the hiding process are presented for different *safety margin* values (10%, 20%, 40%, 60%). This figure indicates that the proposed BA algorithm performs better than CRA [1] when the *safety margin* is less than or equal to 40%. Especially, when the *safety margin* is relatively small (e.g., 10%), CRA does not choose the best transactions to hide first, in contrast with BA. When the *safety margin* becomes 60%, CRA performs better because BA blocks many 0's, and in that way, many ghost rules are created and the number of the side effects is increased. The trade-off between Privacy and Data Loss is clear, because if we raise the *safety margin* (and the number of items that we hide) then more side effects will be created.

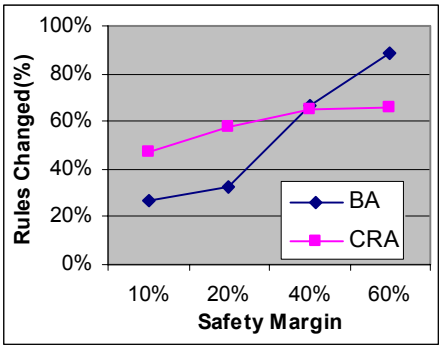

Figure 1. Rules lost or created after the hiding process

5. ANALYSIS OF BLOCKING TECHNIQUE

By using blocking as the hiding technique, the main advantage, compared to other techniques, is that the database is not distorted but only some values become unknown. In that way, a data miner can assume that all the remaining values of the database are true and she can construct models with the remaining data safely. In many real world applications this can be very useful, since it is not always feasible to distort the original database by adding some false information.

The main disadvantage of the blocking technique is the privacy breaches of the modified database. For example, an adversary that wants to infer which are the hidden rules, it could use the placement of the question marks and the remaining data in order to find the hidden values. Namely, all the rules that their generating itemsets contain question marks and their maximum confidence is above the MCT could be the sensitive rules that we want to hide from the adversary. To avoid this problem, an algorithm that implements the blocking technique should create *ghost* association rules so as an adversary will not be able to infer which of the rules that have maximum confidence above MCT are the sensitive and which are the ghost ones. These are called desirable side effects and should be maximized in order to protect the real identity of a specific rule.

REFERENCES

[1] Y. Saygin, V.S. Verykios, and C. Clifton, *Using Unknowns to Prevent Discovery of Association Rules*, SIGMOD Record **30** (2001), no. 4, 45–54.

[2] E. Dasseni, V.S. Verykios, A.K. Elmagarmid, and E. Bertino, *Hiding Association Rules by using Confidence and Support*, In Proceedings of the 4th Information Hiding Workshop (2001), 369–383.