BRBA: A Blocking-Based Association Rule Hiding Method

Peng Cheng,[1,2] **Ivan Lee,**[3] **Li Li,**[1] **Kuo-Kun Tseng,**[2] **Jeng-Shyang Pan**[2]
[1] School of Computer and Information Science, Southwest University, P.R. China
[2] Shenzhen Graduate School, Harbin Institute of Technology, P.R. China
[3] School of Information Technology and Mathematical Sciences, University of South Australia, Australia
chengp.mail@gmail.com

Abstract

Privacy preserving in association rule mining is an important research topic in the database security field. This paper has proposed a blocking-based method to solve the association rule hiding problem for data sharing. It aims at reducing undesirable side effects and increasing desirable side effects, while ensuring to conceal all sensitive rules. The candidate transactions are selected for sanitization based on their relations with border rules. Comparative experiments on real datasets demonstrate that the proposed method can achieve its goals.

Introduction

When people benefit from using data mining techniques to discover unknown knowledge, they also face the risk of disclosing sensitive knowledge. When the data is shared with different organizations or released to the public, other people may expose confidential knowledge which should be kept private to the data owner. Thus, the shared data often need to be sanitized in order to protect the sensitive knowledge contained in it. In this paper, we focus on privacy preserving in association rule mining.

Problem Description

Basic Notation

Let $I = \{I_1, I_2, \ldots, I_m\}$ be a set of items available. An itemset X is a subset of I. A transaction t is an ordered pair, denoted as $t = < ID, X >$, where ID is a unique identifier number and X represents an itemset. A transactional database D is a relation consisting of a set of transactions.

The support of X is the percentage of the transactions in database which contain itemset X, denoted as $supp(X)$. An itemset X is called frequent if $supp(X)$ is at least equal to a minimum relative support threshold (denoted as MST) specified by the user.

The notion of confidence is relevant to association rules. An association rule has the form of $X \rightarrow Y$. It means that the antecedent X infers to the consequent Y, where both X and Y are itemsets. $X \cap Y = \emptyset$. The confidence of a rule is computed as $supp(X \cup Y)/supp(X)$, and denoted as $conf(X \rightarrow Y)$. It indicates a rule's reliability. Like MST, users may specify a minimum confidence threshold called MCT. A rule is considered strong if its support is no less than MST and its confidence is no less than MCT.

The task of association rule mining is to find all strong rules. Fig. 1 introduce an example to clarify the above concepts. The set of available items $I = \{a, b, c, d, e, f\}$. The example database contains 8 transactions. In each transaction, the value 1 indicates existence of the corresponding item, and the value 0 indicates non-existence. Assume $MST = 0.6$ and $MCT = 0.7$, then association rules can be derived.

ID	a	b	c	d	e	f
1	0	1	1	1	1	1
2	1	0	1	0	1	1
3	0	0	1	1	1	0
4	1	0	1	0	0	1
5	0	1	1	1	0	1
6	0	0	1	0	1	1
7	0	0	1	1	1	1
8	0	1	1	0	1	1

⇨

Rule	Conf.	Supp.
e → f	0.833	0.625
f → e	0.714	0.625
e → c	1	0.75
c → e	0.75	0.75
f → c	1	0.875
c → f	0.875	0.875

Figure 1: Database and association rules

Association Rule Hiding

The association rule hiding problem can be formulated as follows. Let R be the set of strong rules that can be mined from D with given MST and MCT. Let R_S denote a set of sensitive rules that need to be hidden, and $R_S \subset R$. The sanitization process is to transform D into a sanitized database D' so that sensitive rules can not be deduced from D' with the same thresholds, while non-sensitive rules in $R \backslash R_S$ still can be mined out to the maximum extent.

In sanitization, some non-sensitive rules may be falsely hidden because their supports or confidences drop below MST or MCT. In addition, some rules, termed as ghost rules, which are not strong in the original database but become strong in the sanitized counterpart because both their supports and confidences get above MST and MCT respectively. These two kinds of rules are called side effects.

Blocking Technique

Some association rule hiding strategies have been proposed (Gkoulalas-Divanis and Verykios 2010). However, most of them are distortion-based techniques, which conceal rules by turning 1's into 0's or 0's into 1's on some selected items. This means that cooperating parties cannot know which data is original and which has been modified into false values when they receive the shared data. This is unacceptable in some cases, such as medical or health applications.

In contrast, the blocking-based techniques sanitize the data through replacing some items with the unknown symbol "?". In this way, support or confidence of a rule is no longer a single value, but mapped into an interval [min_supp, max_supp] or [min_conf, max_conf]. A sensitive rule can be hidden by reducing its minimum support or minimum confidence below MST or MCT. A non-sensitive rule that is strong originally may be hidden after blocking because its minimum support or minimum confidence drops below MST or MCT. A rule that is not strong originally could become strong after blocking if both its maximum support and maximum confidence get above MST and MCT.

In this paper, we propose a blocking-based association rule hiding approach, named as BRBA (Border Rule based Blocking Algorithm).

BRBA Algorithm

BRBA utilizes the concept of border rules to select suitable transactions for sanitization. The algorithm aims to achieve the following goals:

1. Reduce the minimum support or confidence of each sensitive rule below MST or MCT by a Safety Margin (SM).
2. Minimize the undesirable side effects, i.e., the number of missing non-sensitive rules.
3. Maximize the desirable side effects, i.e., the number of ghost rules.

Border rules have the following features. Their supports or confidences are close to MST or MCT and they are very easy to become missing rules or ghost rules in sanitization. Border rules can be divided into positive border rules and negative border rules. A positive border rule is easy to be concealed mistakenly and a negative border rule is easy to become a ghost rule. Ghost rules are desirable because it beholds the same features with hidden sensitive rules. They both contain the symbol "?". Their maximum supports and confidences are greater than MST and MCT, but their minimum supports or confidences are less than MST or MCT. If malicious attackers try to expose sensitive rules with these traits in the released database, they will find many ones and cannot determine which ones are sensitive. So, increasing ghost rules favors preventing privacy breach.

Candidate transactions are evaluated according to positive and negative border rules they contain. Candidates containing fewer positive border rules and more negative border rules are given high priorities for sanitization.

In order to prevent malicious attackers from recovering the database by replacing all "?" with 1's or with 0's, BRBA performs sanitization by blocking 1's and blocking 0's simultaneously. The algorithm adopts the parameter A_{01} to control the proportion of blocked 0's.

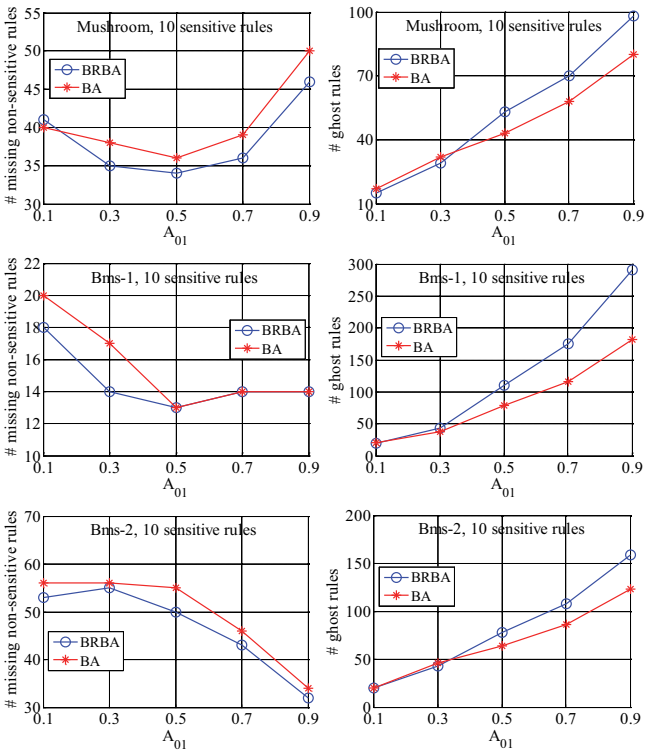

Figure 2: Side effects with increasing values of A_{01}

Experimental Results and Analysis

We compared BRBA empirically with the algorithm BA in (Verykios et al. 2007) on three real datasets: Mushroom, Bms-1 and Bms-2. The Safety Margin was applied in both algorithms. Fig. 2 shows the side effects of two algorithms with different A_{01} values (0.1, 0.3, 0.5, 0.7, 0.9).

As indicated in Fig. 2, the number of ghost rules grows accordingly with increasing values of A_{01}. It demonstrates that blocking more 0's can really increase the number of ghost rules. This is desirable since more ghost rules can improve the safety of sensitive rules. Generally, BRBA may generate more ghost rules, and less or approximately same missing non-sensitive rules than BA on each test case. It shows that utilizing the concept of border rules can guide the BRBA algorithm to produce more desirable side effects and less undesirable side effects.

References

Gkoulalas-Divanis, A., and Verykios, V. S. 2010. *Association Rule Hiding for Data Mining*, volume 41 of *Advances in Database Systems*. Kluwer.

Verykios, V. S.; Pontikakis, E. D.; Theodoridis, Y.; and Chang, L. 2007. Efficient algorithms for distortion and blocking techniques in association rule hiding. *Distributed & Parallel Databases* 22(1):85–104.