Volume 5, Issue 4, 2015 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Sensitive Representative Association Rules

Maryam Nourafkan*, Hamid Rastegari, Mohammad Naderi Dehkordi

Department of Computer Engineering, Islamic Azad University,
Najafabad Branch, Isfahan, Iran

Abstract— Association rule mining is an important data-mining technique that finds interesting association among a large set of data items. Since it may disclose patterns and various kinds of sensitive knowledge that are difficult to find otherwise, it may pose a threat to the privacy of discovered confidential information. The first step for hiding certain information and/or confidential knowledge in the data set, and having a new database for non-confidential access is, finding sensitive association rules. The number of sensitive association rules may be huge. This leads to increase run-time, and changes which are applied to the database for immunization process. Clustering, finding representative association rules, and some other techniques would help to reduce the numbers of sensitive rules. This leads to reduce run-time, and changes which are applied to the database. This study investigates how to mine sensitive representative association rules which is a least set of rules that covers all association rules by means of the cover operator. In this connection, sensitive representative rules are mined based an algorithm named GSRR.

Keywords—Data Mining, Sensitive Itemsets, Representative Association Rules

I. INTRODUCTION

The data explosion in recent years urges the need of data mining in the form of tools for the better management of immense data and the investigation of the new relation and information. In cases, organizations require to reveal their data. Since confidential data might exist here in this giant size of data, organizations might not be willing to expose the confidential sections. The first step for hiding certain information and/or confidential knowledge in the data set, and having a new database for non-confidential access is, finding sensitive association rules. The number of sensitive association rules may be huge. This leads to increase run-time, and changes which are applied to the database for immunization process. Clustering, finding representative association rules, and some other techniques would help to reduce the numbers of sensitive rules. This leads to reduce run-time, and changes which are applied to the database. This study investigates how to mine sensitive representative association rules which is a least set of rules that covers all association rules by means of the cover operator. At first, the data mining owner identifies the sensitive item sets. Then, sensitive representative rules which includes the members of this set on the right hand side is mined through the recommended algorithm. The rest of the paper is organized as follows. Section 2 presents the statement of the problem and the notation used in the paper. Section 3 presents a review of related performed work. Proposed algorithms are presented in section 4. Section 5 shows example of the proposed algorithms. And section 6 presents future works, discussions and conclusion.

II. PROBLEM STATEMENT

Let $I = \{i_1, i_2 \dots, i_m\}$ be a set of m distinct literals, called items. Given a set of transactions D, where each transaction T is a set of items such that $T \subseteq I$. An association rule is an implication of the form $X \to Y$ where $X \subset I, Y \subset I, X \cap Y = \phi$. X and Y are called antecedent/body and consequent /head of the rule respectively. Strength of a rule whether it is strong or not is measured by two parameters called support and confidence of the rule. These two parameters help in deciding the interestingness of a rule ([1], [6], [14]). For a given rule $X \to Y$ support is the percentage of transaction that contains both X and Y $(X \cup Y)$ or is the proportion of transactions jointly covered by the LHS and RHS and is calculated as:

$$Supp\,(X \to Y) = Supp\,(X \cup Y) = \frac{|X \cup Y|}{|N|} \qquad (1)$$

Where, N is the number of transactions. Confidence is the percentage for a transaction that contains X also contains Y or is the proportion of transactions covered by the LHS that are also covered by the RHS and is calculated as:

$$Conf\,(X \to Y\,) = \frac{|X \cup Y|}{|X|} = \frac{Supp\,(X \cup Y)}{Supp\,(X)} \qquad (2)$$

For a database of transactions with certain sets of items, there can be too much association rules potentially. A rule is significant if its support and confidence is higher than the user specified minimum support threshold (MST) and minimum confidence threshold (MCT). In this way, algorithms do not retrieve all the association rules that may be derivable from a database, but only a very small subset that satisfies the minimum support and minimum confidence requirements set by the users. An association rule-mining algorithm works as follows. It finds all the sets of items that appear frequently enough to be considered relevant and then it derives from them the association rules that are strong enough to be considered interesting. We aim at preventing some of these rules that we refer to as “sensitive rules”, from being disclosed. The problem can be stated as follows: Given a database D, a set F of frequent itemsets that are mined from D and a set X of sensitive items, how can we mine sensitive representative association rules from D? ([1], [6], [14])

III. BACKGROUND AND RELATED WORK

There are mainly 3 approaches for association rule hiding (i) Exact approach (ii) Border based approach (iii) Heuristic approach. In following, overview of these approaches is given in brief ([6], [13]).

Exact approach: This approach contains none heuristic algorithms which formulates the hiding process as a constraints satisfaction problem or an optimization problem which is solved by integer programming. These algorithms can provide optimal hiding solution with ideally no side effects. Border based approach: This approach hides sensitive association rule by modifying the borders in the lattice of the frequent and the infrequent itemsets of the original database. Heuristic approach: This approach involves efficient, fast and scalable algorithms that selectively sanitize a set of transactions from the original database to hide the sensitive association rules. Various heuristic algorithms are based on mainly two techniques: Data distortion technique and blocking technique. Blocking is the replacement of an existing value with a “?”. It inserts unknown values in the data to fuzzify the rules. In some applications where publishing wrong data is not acceptable, then unknown values may be inserted to blur the rules. Data distortion is done by the alteration of an attribute value by a new value. It changes 1‟s to 0‟s or vice versa in selected transactions to increase or decrease support or confidence of sensitive rule. Heuristic algorithms cannot give an optimal solution because of undesirable side effects to none sensitive rules, e.g. lost rules and new rule. Algorithms proposed using heuristic approach can be divided into rule hiding and itemset hiding algorithms.

In most cases finding sensitive association rules is the first step of immunization. Most approaches try to hide every single sensitive association rule without checking if some rules could be pruned out after some changes have been made in the database while hiding some rules previously. If the number of sensitive association rules is too large then the number of passes taken by these approaches is equal to the number of sensitive rules, which can be a great overhead for hiding algorithms. This process increases the number of lost rules and run-time as well. Clustering, finding representative association rules, and some other techniques would help to reduce the numbers of sensitive rules. This leads to reduce run-time, and changes which are applied to the database. Olivera & Zaiane ([8], [9]) for the first time proposed the method of Multiple Rule Hiding. To hide there is a need to scan the dataset twice, regardless of the number of Sensitive Items. The first scan is to make index files to speed up the process of finding sensitive transactions and allow for efficient retrieval of data. The second scan is to apply algorithm Dataset selectively. Shah, Takkar and Ganatra [13] proposed two association rule hiding algorithms, ADSRRC (Advanced Decrease Support of R.H.S. items of Rule Cluster) and RRLR (Remove and Reinsert L.H.S. of Rule), based on heuristic approach. Both algorithms are based on algorithm DSRRC (Decrease Support of R.H.S. items of Rule Cluster) proposed in [7]. Algorithm DSRRC depends on ordering of transactions for removing items from database. Also it requires sorting of database each time item is removed from database. Algorithm ADSRRC is proposed to overcome these limitations. Algorithm DSRRC cannot hide rule having multiple R.H.S. items. To overcome this limitation algorithm RRLR is proposed. DSRRC and ADSRRC use clustering in order to reduce the numbers of sensitive association rules. They clusters the sensitive rules based on R.H.S. It means sensitivity is the sum of the sensitivities of all association rules in cluster. Cluster sensitivity defines the rule cluster which is most affecting to the privacy. In DSRRC transaction sensitivity is different for each cluster but ADSRRC calculates transaction sensitivity irrespective of clusters. It means for all clusters transaction sensitivity is same. Jain et al [3] proposed approach uses the data distortion technique where the position of the sensitive items is altered but its support is never changed. The size of the database remains the same. The proposed heuristics use the idea of representative rules to prune the rules first and then hides the sensitive rules. It uses the idea of representative rules to prune the rules first and then hides the sensitive rules. Due to this property by hiding a sensitive representative rule, sub rules that are covered by this rule will be concealed as well. This leads to reduce run-time, and changes which are applied to the database.

Discovering association rules between items in a large database is an important database mining problem. The number of association rules may be huge. Kryszkiewicz [4] introduced a notion of a cover operator which transforms an association rule into the set of association rules by syntactic transformation of the initial rule. Representative association rules are defined as a least set of rules that covers all association rules satisfying certain user specified support and confidence. A user may be provided with a set of representative association rules instead of the whole set of association rules. In this paper, an algorithm for computing representative association rules is offered. In this paper, to check whether a candidate rule is representative the algorithm required comparing the rule with longer representative rules, which was quite time consuming operation. Kryszkiewicz [5] investigated some properties of representative association rules and propose a new efficient algorithm for representative association rules mining. The new algorithm generates representative rules independently from other representative rules. Unlike the algorithm proposed in [4], the new algorithm exploits solely the information about the supports of frequent itemsets.

IV. PROPOSED ALGORITHM

In this paper, an algorithm for mining sensitive representative association rules is presented. In this connection, representative association rules which is a least set of rules that covers all association rules by means of the cover operator are mined based an algorithm named Generate Sensitive Representative Rules (GSRR).

A. Cover Operator

A notion of a cover operator was introduced in ([4], [5], [12]) for deriving a set of association rules from a given association rule without accessing a database. The cover C of the rule $r : X \rightarrow Y, Y \neq \emptyset$ is defined as follows:

$$C(r : X \rightarrow Y) = \{X \cup U \rightarrow V \mid U, V \subseteq Y, U \cap V = \emptyset, and\ V \neq \emptyset\} \quad (3)$$

Each rule in $C(r : X \rightarrow Y)$ consists of a subset of items occurring in the rule $X \rightarrow Y$. The antecedent of any rule r covered by $X \rightarrow Y$ contains X and perhaps some items from Y, whereas r's consequent is a non-empty subset of the remaining items in Y. It was proved in [4] that each rule r in the cover $C(r')$, where r' is an association rule having support s and confidence c, belongs in $AR(s,c)$. Hence, if r belongs in $AR(s,c)$ then every rule r'in $C(r)$also belongs in $AR(s,c)$.

1) *Example*: For a given set of transactional data given in Table 1. Considering *r: (BC → ADE)*, support $s = 3$, and confidence *c=75%* Table 2 contains all rules belonging in the cover *C(r)*:

TABLE I DATABASE

TID	A	B	C	D	E	F	G	H
T1	1	1	1	1	1	0	0	0
T2	1	1	1	1	1	1	0	0
T3	1	1	1	1	1	0	1	1
T4	1	1	0	0	1	0	0	0
T5	0	1	1	1	1	0	1	1

TABLE IIII THE COVER OF THE ASSOCIATION RULE (BC →□ADE)

#	Cover	Support	Confidence
1	BC→ADE	3	75%
2	BC→A	3	75%
3	BC→D	4	100%
4	BC→E	4	100%
5	BC→AD	3	75%
6	BC→AE	3	75%
7	BC→DE	4	100%
8	ABC→D	3	100%
9	ABC→DE	3	100%
10	ABC→E	3	100%
11	BCD→A	3	75%
12	BCD→AE	3	75%
13	BCD→E	4	100%
14	BCE→A	3	75%
15	BCE→AD	3	75%
16	BCE→D	4	100%
17	ABCD→E	3	100%
18	ABCE→D	3	100%
19	BCDE→A	3	100%

B. Representative Association Rules

In this section we describe a notion of representative association rules which was introduced in ([4], [5], [12]). Informally speaking, a set of all representative association rules is a least set of rules that covers all association rules by means of the cover operator. A set of representative association rules with minimum support s and minimum confidence c will be denoted by $RR(s,c)$ and defined as follows:

$$RR(s,c) = \{r \in AR(s,c) \mid \nexists r' \in AR(s,c), r \neq r'\ and\ r \in C(r')\} \quad (4)$$

If s and c are understood than $RR(s,c)$ will be denoted by RR. Each rule in RR is called a representative association rule. By the definition of RR no representative association rule may belong in the cover of another association rule.

1) *Example:* Given minimum support *s = 3* and minimum confidence *c = 75%*, the following representative rules *RR(s,c)* would be found for the database *D* from Table 1:

{A→BCDE, C→ABDE, D→ABCE, B→CDE, E→BCD, B→AE, E→AB}.

C. GSRR Algorithm

GSRR algorithm which is a modification of the FastGenAllRepresentatives given below: [5]

```
Algorithm GSRR
INPUT: (1) all frequent itemsets F,
       (2) a min_confidence c,
        $(3)$ $a$ $set$ $of$ $items$ $X$;
$OUTPUT$: $sensitive$ $representative$ $rules$ $SRR$;
$forall$ $x \in X$ $do$ $begin$
    $If$ $(x$ $isn't$ $in$ $F_1)$ $then$ $X = X - \{x\}$;
$endfor$;
$If$ $(X$ $is$ $empty)$ $then$ $EXIT$;
$Select$ $all$ $itemsets$ $from$ $F$ $which$ $have$ $x$ $and$ $store$ $in$ $FS$;
$forall$ $Z \in FS$ $do$ $begin$
    $K = |Z|$; $maxsupp = max(\{supp(Z') \mid Z \subset Z' \in F_{K+1}\} \cup \{0\})$;
    $If$ $Z.supp \neq maxsupp$ $then$ $begin$
        $A_1 = \{\{Z[1]\},\{Z[2]\},\ldots,\{Z[k]\}\}$;
        /* loop1 */
        $for$ $(i = 1; (A_i \neq \emptyset) and (i < K); i++)$ $do$ $begin$
            $forall$ $X \in A_i$ $do$ $begin$
                $find$ $Y \in F_i$ $such$ $that$ $Y = X$;
                $XCount = Y.Count$;
                $if$ $(Z.supp / XCount \geq c)$ $then$ $begin$
                    $if$ $(maxsupp/XCount < c)$ $then$
                        $print$ $(X, " \rightarrow ", Z|X,$ "with support:", $Z.supp,$ "and confidence:", $Z.supp / XCount)$;
                    $A_i = A_i \setminus \{X\}$;
                $endif$;
            $endfor$;
            $A_{i+1} = AprioriGen(A_i)$;
        $endfor$;
    $endif$;
$endfor$;
```

A brief description of important steps of the algorithm is given below:

The GSRR algorithm computes sensitive representative association rules from each itemset in FS. Let Z be a considered itemset in FS. Only k-rules, $K = |Z|$are generated from Z. First, $maxsupp$ is determined as a maximum from the supports of these itemsets in F_{k+1} which are supersets of Z. If there is no superset of Z in F_{k+1} then $maxsupp = 0$. Loop1 starts. In general, the i-th iteration of Loop1 looks as follows: Each candidate $X \rightarrow Z \backslash Y$, where $X \subset Z$ belongs in i-itemsets A_i, is considered. Z is frequent, so X, which is a subset of Z, is also frequent. In order to check if $X \rightarrow Z \backslash Y$ is an association rule its confidence: $supp(Z) / supp(x)$ has to be determined. $supp(Z) = Z.supp$, while $supp(x)$ is computed as Y. Only association rules that satisfy $maxsupp \leq s$ or $maxsupp/XCount < c$ are representative. The antecedent X of each association rule $X \rightarrow Z \backslash Y$ is removed from A_i. Having found all representative k-rules with i-antecedents from Z, (i+1)-itemset antecedents A_{i+1} are built from A_i by the AprioriGen function given in ([4], [5]).

V. EXAMPLE

The proposed algorithms can be illustrated with the following example for a given set of transactional data given in Table 1. Considering $s = 3$, the following itmesets in Table 3 are mined:

TABLE IVVVI FREQUENT ITEMSETS

Frequent itemsets (F)	Support (S)
A	4
B	5
C	4
D	4
E	5
AB	4
AC	3
AD	3
AE	4
BC	4
BD	4
BE	5
CD	4
CE	4
DE	4
ABC	3

ABD	3
ABE	4
ACD	3
ACE	3
ADE	3
BCD	4
BCE	4
BDE	4
CDE	4
ABCD	3
ABCE	3
ABDE	3
ACDE	3
BCDE	4
ABCDE	3

Applying the GSRR algorithm on the database represented in Table 1 for minimum confidence $c = 0.75$ and a set of sensitive items $X = \{A\}$, the following sensitive representative rules are found:
$SRR = \{A \rightarrow BCDE, C \rightarrow ABDE, D \rightarrow ABCE, B \rightarrow AE, E \rightarrow AB\}$

VI. CONCLUSION AND FUTURE WORK

In this paper, an algorithm named GSRR for pruning extracted rules from a database by using the concept of representative rules was presented. Most approaches try to hide every single sensitive association rule without checking if some rules could be pruned out after some changes have been made in the database while hiding some rules previously. If the number of sensitive association rules is too large then the number of passes taken by these approaches is equal to the number of sensitive rules, which can be a great overhead for hiding algorithms. This process increases the number of lost rules and run-time as well. Therefore, the proposed algorithm can combine with other hiding algorithms in order to increase the efficiency of immunization process. Using the idea of sensitive representative association rules would prune the rules first and then hide the sensitive rules. Due to this property by hiding a sensitive representative association rule, sub rules that are covered by this rule will be concealed as well. This leads to reduce run-time, and changes which are applied to the database. We can also have a combination of proposed algorithm and algorithms presented in ([15], [16]) for our future work, to hide sensitive representative association rules.

Different algorithms to obtain a set of representative association rules are presented .In order to increase the efficiency of the proposed algorithm, we can improve representative association rules mining algorithms. We can use and improve algorithm which presented in ([10], [11], [12]).

REFERENCES

[1] Berry MJ, Linoff GS, Data mining techniques: for marketing, sales, and customer relationship management. New Jersey: *The Wiley*; 2004.

[2] Dasseni E, Verykios VS, Elmagarmid AK, Bertino E. Hiding association rules by using confidence and support. *Proceedings of the 4th International Workshop on Information Hiding*; 2001. p. 369-383.

[3] Jain D, Sinhal A, Gupta N, Narwariya P, Saraswat D, Pandey A. Hiding sensitive association rules without altering the support of sensitive item(s). *International Journal of Artificial and Applications (IJAIA)*. 2012; 3(2): 75-84.

[4] Kryszkiewicz M. Representative association rules. *Proceedings of the 1998 Springer-Verlag Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'98)*; 1998. p. 198-209.

[5] Kryszkiewicz M. Fast discovery of representative association rules. *Proceedings of the 1998 Springer-Verlag Rough Sets and Current Trends in Computing (RSCTC'98)*; 1998. p. 214-221.

[6] Luo Y, Zhao Y, Le J. A survey on the privacy preserving algorithm of association rule mining. *Proceedings of the 2009 IEEE 2nd International Symposium on Electronic Commerce and Security (ISECS'09)*; 2009. p. 241-245.

[7] Modi CN, Rao UP, Patel DR. Maintaining privacy and data quality in privacy preserving association rule mining. *Proceedings of the 2010 IEEE International Conference on Computing Communication and Networking Technologies (ICCCNT'10)*; 2010. pp. 1-6.

[8] Oliveira SRM, Zaiane OR. Privacy preserving frequent itemset mining. *Proceedings of the 2002 IEEE International Conference on Privacy, Security and Data Mining (PSDM'02)*; 2002; Maebashi City, Japan. p. 43-54.

[9] Oliveira SRM, Zaiane OR. Protecting sensitive knowledge by data sanitization. *Proceedings of the IEEE International Conference on Data Mining (ICDM'03)*; 2003. p. 211-218.

[10] Pasquier N, Bastide Y, Taouil R, Lakhal L. Efficient mining of association rules using closed item set lattices. *Information Systems*. 1999; 24(1): 25-46.

[11] Pasquier N, Bastide Y, Taouil R, Lakhal L. Discovering frequent closed itemsets for association rules. Proceedings of the 1999 Springer- Heidelberg International Conference on Database Theory (ICDT '99); 1999. p. 398-416.

[12] J. Saquer, J.S. Deogun, Using Closed itemsets for discovering representative association rules, *Proceedings of the 2000 Springer-Verlag International Symposium on Methodologies for Intelligent Systems (ISMIS'10)*, p. 495-504, 2010.

[13] K. Shah, A. Thakkar and A. Ganatra, Association Rule Hiding by Heuristic Approach to Reduce Side Effects & Hide Multiple R.H.S. Items. *International Journal of Computer Applications*. 2012; 45(1): 1-7.

[14] Verykios VS, Emagarmid AK, Bertino E, Saygin Y, Dasseni E. Association rule hiding. *IEEE Transactions on Knowledge and Data Engineering*. 2004; 16(4): 434-447.

[15] Wang SL, Jafari A. Hiding sensitive predictive association rules. *IEEE International Conference on Systems, Man and Cybernetics*, p. 164-169, 2005.

[16] Wang SL, Parikh B, Jafari A. Hiding informative association rule sets. *Journal of Expert Systems with Applications*. 2007; 33(2): 316-323.