Volume 3, Issue 9, September 2013 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

A Generalized Association Rule Based Method for Privacy Preserving in Data Mining

Vijay Patidar* *M.Tech IT ITM Bhilwara, India*

Vikash Shrivastava *Senior Architect I-Gate Solution Pvt. Limited Noida New Delhi (India)*

Vivek Shrivastava *Assistant Professor, HOD IT ITM Bhilwara, India*

Abstract: Mining association rules from huge amounts of data is an important issue in data mining, with the discovered information often being commercially valuable. Moreover, companies that conduct similar business are often willing to collaborate with each other by mining significant knowledge patterns from the collaborative datasets to gain the mutual benefit. However, in a cooperative project, some of these companies may want certain strategic or private data called sensitive patterns not to be published in the database. Therefore, before the database is released for sharing, some sensitive patterns have to be hidden in the database because of privacy or security concerns. To solve this problem, sensitive-knowledge-hiding (association rules hiding) problem has been discussed in the research community working on security and knowledge discovery. The aim of these algorithms is to extract as much as non sensitive knowledge from the collaborative databases as possible while protecting sensitive information. Sensitive-knowledge-hiding problem was proven to be a nondeterministic polynomial-time hard problem. After that, a lot of research has been completed to solve the problem. In this article, we will introduce a new modified hybrid algorithm for privacy preserving. There are many approaches to hide certain association rules which take the support and confidence as a base for algorithms ([1, 2, 6,7]). Our approach is a modification of work done by [7]. Our algorithm takes lesser number of passes to hide a specific association rule.

Keywords: Association Rule Mining, Sensitive Rule Hiding, Support, Confidence.

1. Introduction

Data mining extracts novel and useful knowledge from large repositories of data and has become an effective analysis and decision means in corporation. The sharing of data for data mining 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 protected before published. Motivated by the multiple conflicting requirements of data sharing, privacy preserving and knowledge discovery, privacy preserving data mining 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). Concretely, the problem of KHD can be described as follows:

Given a data set D to be released, a set of rules R mined from D, and a set of sensitive RS to be hided, how can we get a new data set D', such that the rules in RS cannot be mined from D', while the rules in R-RS can still be mined as many as possible.

Typically, when *D* is a transaction database and *R* is specific to the set of association rules mined from *D* with minimum support threshold *MST* and minimum confidence threshold *MCT*, the problem of KHD becomes association rule hiding problem.

Clifton in [5] provided a well designed scenario which clearly shows the importance of the association rule hiding problem. In the scenario, by providing the original unaltered database to an external party, some strategic association rules that are crucial to the data owner are disclosed with serious adverse effects. The sensitive association rule hiding problem is very common in a collaborative association rule mining project, in which one company may decide to disclose only part of knowledge contained in its data and hide strategic knowledge represented by sensitive rules. These sensitive rules must be protected before its data is shared. Besides, by hiding some association rules, data owners can prevent the rule-based vicious inferences used for unwarrantable purposes, e.g. uncovering private data, as discussed in [8]. The authors in [3] first proposed the concept of "data sanitization" to settle the association rule hiding problem. Its main idea is to select some transactions to modify (delete or add items) from original database through some heuristics. They also proved that the optimal sanitization is an NP-hard problem. After that, many approaches have been proposed in data sanitization framework. Association rule hiding based on data sanitization framework operates simply. However, data sanitization techniques cannot control the hiding effects of confidential rules obviously. The hiding effects can only be validated after sanitization. In other words, they suffer from the weakness of providing a way of fine tuning the

generation of the released database. Moreover, data sanitization can produce a lot of I/O operations, which greatly increase the time cost, especially when the original database includes a large number of transactions. Different from the data sanitization framework, the authors in [4] proposed a novel framework that can be regarded as "knowledge sanitization" approach, which is inspired by the inverse frequent set mining problem. The new proposed framework first performs sanitization on an itemset lattice called a knowledge base from which association rules can be derived. The itemset lattice is defined as all partial ordered subset items generated from given transactions. Then a reconstruction procedure reconstructs a new released dataset from the sanitized itemset lattice. In one word, this approach conceals the sensitive rules by sanitizing itemset lattice rather than sanitizing original dataset. Compared with original dataset, itemset lattice is a medium production that is closer to association rules. In this way, one can easily control the availability of rules that can be mined from original dataset and control the hiding effects directly. However, as a rudimental work, the approach proposed in [4] is still very incomplete and limited in the following two aspects:

1) It does not give concrete guidance on how to sanitize the itemset lattice according to the sensitive association rules.

2) The feasibility of the data reconstruction process is restricted to whether the knowledge sanitization process can produce an itemset lattice with consistent support value configuration relationship. But, in fact, the proposed knowledge sanitization process cannot guarantee that one can always find a consistent one within a polynomial time.

2. Related Work

The problem of association rule hiding was first probed in [3]. After that, many approaches were proposed. Roughly, they can fall into two groups: data sanitization data modification approaches (data modification for short) and knowledge sanitization data reconstruction (data reconstruction) approaches. Most of the researchers have worked on the basis of reducing the support and confidence of sensitive association rules ([1, 2, 6, 7]). ISL and DSR are the common approaches used to hide the sensitive rules.

In 2008, Belwal et al[7] Presented an algorithm. To hide any specified association rule $X \rightarrow Y$ our algorithm works on the basis of confidence ($X \rightarrow Y$) and support ($X \rightarrow Y$).To hide the rule $X \rightarrow Y$ (containing sensitive element X on LHS),our algorithm increases the special variable 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. Our approach is a modification of work done by [7]. Our algorithm takes lesser number of passes to hide a specific association rule.

3. Problem Statement

The problem of sensitive rule hiding is described as follows:

> Given a transaction database, MST, MCT, a set of strong rules, and a set of sensitive items, how can we modify the database such that using the same MST and MCT, the set of strong rules in the modified database satisfies all the constraints: 1) no sensitive rule, 2) no lost rule, and 3) no false rule?

Let D be the database of transactions and J = {J1, ..., Jn} be the set of items. A transaction T includes one or more items in J . An association rule has the form $X \rightarrow Y$, where X and Y are non-empty sets of items (i.e. X and Y are subsets of J) such that $X \cap Y = Null$. A set of items is called an itemset, while X is called the antecedent. The support of an item (or itemset) x is the percentage of transactions from D in which that item or itemset occurs in the database The confidence or strength c for an association rule $X \rightarrow Y$ is the ratio of the number of transactions that contain X or Y to the number of transactions that contain X.

The problem of mining association rule is to find all rules that have support and confidence greater then user specified minimum support threshold (MST) and minimum confidence threshold (MCT).

4. Proposed Algorithm

To hide any specified association rule $X \rightarrow Y$ this algorithm works on the basis of confidence ($X \rightarrow Y$) and support ($X \rightarrow Y$). To hide any sensitive rule $X \rightarrow Y$, this algorithm first finds the value of support (sup) and confidence (conf) in the available rule and then it computes the support and confidence of the sensitive rule using following:

Confidence ($X \rightarrow Y$) = (conf * factor);

Support ($X \rightarrow Y$) = (sup * factor);

Input:

1. A database of transactions
2. A database of rules
3. A set of sensitive items X
4. A minimum support threshold (MST)
5. A minimum confidence threshold (MCT)

Output:

A transformed database of rules with modified support and confidence where rules containing X will be hidden.

Procedure:

//find value of support and confidence

Select confidence into conf from database.
Select support into supp from database.

For each X
{
// Now check all the rules containing sensitive element x.
For each rule R which contain X on LHS or RHS.
{
While (conf(R) >= MCT)
{
Set confidence(X → Y) = (conf * factor);
(If the number of transactions are more than 1000 then choose 1/3 as factor. If the no of transactions are between 100 and 1000 then choose 1/5 as factor. If less then 100 then choose 1/10 as factor)
Set support (X → Y) = (sup * factor);
}
}
}
End of procedure

5. Example :

A Data Set

Suppose there is a database of transactions as below:

TID	Items
T1	ABD
T2	B
T3	ACD
T4	AB
T5	ABD

Fig 1: A Data Set

One has also given a MST of 60% and a MCT of 70%. One can see four association rules can be found as below

A → B (60%, 75%)
B → A (60%, 75%)
A → D (60%, 75%)
D → A (60%, 100%)

Now there is a need to hide D and B.

Previous Methods:

One can see that by simple **ISL** algorithm if someone want to hide D and B, then he can check it by modifying the transaction T2 from B to BD (i.e. from 0100 to 0101).but still ISL cannot hide the rule D → A. Let us see by following example

TID	Items	Bit Map
T1	ABD	1101
T2	B	0100
T3	ACD	1011
T4	AB	1100
T5	ABD	1101

(Hiding D → A by ISL approach)

TID	Items	Bit Map
T1	ABD	1101
T2	B	***0101***
T3	ACD	1011
T4	AB	1100
T5	ABD	1101

So by above explanation it is clear that rule D → A can not be hidden by ISL approach because by modifying T2 from B to BD (i.e. from 0100 to 0101) rule D → A will have support and confidence 60% and 75% respectively.

By DSR approach:

TID	Items	Bit Map
T1	ABD	1101
T2	B	0100
T3	ACD	1011
T4	AB	1100
T5	ABD	1101

(Hiding D → A by DSR approach)

TID	Items	Bit Map
T1	ABD	***0101***
T2	B	0100
T3	ACD	1011
T4	AB	1100
T5	ABD	1101

By **DSR** approach rule D → A is hidden as its support and confidence is now 40% and 66% respectively, but as a side effect the rule A → D is also hidden.

By Hiding Counter Approach:

TID	Items
T1	ABD
T2	B
T3	ACD
T4	AB
T5	AB

	support ,	confidence,	special variable
A → B	60%	75%	0
B → A	60%	75%	0
A → D	60%	75%	0
D → A	60%	100%	0

First to hide B

TID	Items
T1	ABD
T2	B
T3	ACD
T4	AB
T5	ABD

	support ,	confidence,	special variable
A →B	60%	75%	0
B →A	50%	60%	1**(Rule is hidden)**
A →D	60%	75%	0
D →A	60%	100%	0

Now to hide D

TID	Items
T1	ABD
T2	B
T3	ACD
T4	AB
T5	ABD

	support	**, confidence,**	**special variable**
A →B	60%	75%	0
B →A	50%	60%	1**(Rule is hidden)**
A →D	60%	75%	0
D →A	43%	60%	2

By Proposed Algorithm :

To hide A. After 1 pass the status of database is as follows:

	support , confidence	
A →B	20%	25%
B →A	50%	60%
A →D	20%	25%
D →A	43%	60%

6. Result Comparison & Conclusion:

So it is clear that this approach is hiding all the given sensitive rules successfully without any side effect in small as well as large databases. Previous algorithm (Hiding counter) is also hiding all sensitive rules without any side effect in small datasets. But when the transactions are too many then in that case this algorithm is having a major problem. It takes too many passes to bring down the support below the threshold. For example: Suppose a database has 100 transactions. If support of a rule says X -> Y is 80%. If MST is 30% then this algorithm will require 170 passes to reduce the support to 29%. But our proposed algorithm 1 will take 1 pass to reduce the support below 30%. So it is clear that the proposed algorithm is more efficient. Also previous algorithm hides all those rules in which sensitive items occur in left side of that rule but our algorithm hides all those rules in which sensitive items occur either in left side or right side of that rule.

References:

1. Shyue-Liang Wang, Yu-Huei Lee, Steven Billis, Ayat Jafari "Hiding Sensitive Items in Privacy Preserving Association Rule Mining" 2004 International Conference on Systems, Man and Cybernetics.
2. Vassilios S. Verykios, Ahmed K. Elmagarmid, Elisa Bertino, Yucel Saygin and Elena Dasseni"Association Rule Hiding", IEEE Transactions on Knowledge and Data Engineering, Vol. 16No. 4, April 2004
3. Atallah, M., Bertino, E., Elmagarmid, A., Ibrahim, M., and Verykios, V.S. Disclosure limitation of sensitive rules. In: Scheuermann P, ed. *Proc. of the IEEE Knowledge and Data Exchange Workshop (KDEX'99)*. IEEE Computer Society, 1999. 45-52.
4. Chen, X., Orlowska, M., and Li, X. A new framework for privacy preserving data sharing. In: *Proc. of the 4*P *th* P *IEEE ICDM Workshop*: *Privacy and Security Aspects of Data Mining*. IEEE Computer Society, 2004. 47-56.
5. Clifton, C. and Marks, D. Security and privacy implications of data mining. In: *Proc. of the ACM SIGMOD Workshop Data Mining and Knowledge Discovery*. 1996. 15-19.
6. Vi-Hung Wu, Chia-Ming Chiang, and Arbee L.P. Chen, Senior Member, IEEE Computer Society Hiding Sensitive Association Rules Limited Side Effects IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 19, NO.1, JANUARY 2007
7. Belwal, Varsheney, Khan, Sharma, Bhattacharya. *Hiding sensitive association rules efficiently by introducing new variable hiding counter*. Pages 130-134, 978-2008, IEEE xplore.
8. Fienberg, S. and Slavkovic, A. Preserving the confidentiality of categorical statistical data bases when releasing information for association rules. *Data Mining and Knowledge Discovery*, 11(2):155–180, 2

