International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 November 2023*

Preserving the Support of Sensitive Item(s) while Hiding Sensitive Association Rules

Ashoktaru Pal[1], Dr. Ajay R. Raundale [2]
[1]Ph.D. Research Scholar, Dept. of Computer Science and Engineering, Dr. A. P. J. Abdul Kalam University, Indore, India
Email: pal.ashoktaru@gmail.com.
[2]Assistant Professor, Dept. of Computer Science and Engineering, Dr.A.P.J.Abdul Kalam University, Indore, India.
Email : arraundale@gmail.com

Abstract – An essential data-mining method for identifying intriguing relationships among a sizable collection of data objects is association rule mining. It may be a threat to the privacy of uncovered confidential information since it may reveal patterns and other types of sensitive knowledge that are hard to obtain in other ways. Such data needs to be shielded against unwanted access. Numerous tactics had been put forth to conceal the knowledge. Some employ data disruption, clustering, data distortion, and distributed databases across multiple sites. The need to strike a balance between the user's legitimate needs and the secrecy of exposed data is a challenge with hiding sensitive rules that has not yet received enough attention. The suggested method makes advantage of the data distortion methodology, which modifies the sensitive elements' position without changing their support. The database is still the same size. It first prunes the rules using the concept of representative rules, and then it conceals the rules that are sensitive. This strategy has the advantage of hiding the maximum number of rules; in contrast, the existing ways are unable to conceal all the needed rules, which should be concealed in the fewest number of passes. The suggested method is also contrasted with current approaches in the paper.

Keywords: *Privacy preserving data mining; Association rule; Association rule hiding.*

1 Introduction

The practice of extracting meaningful information and patterns from massive databases is known as data mining. Data mining has been extremely important in recent years because it allows managers to access information that is concealed and utilize it to make decisions. Safeguarding data from unwanted access becomes crucial when working with sensitive information.[1], [2A] significant challenge is striking a balance between the rightful demands of data consumers and the confidentiality of the exposed data. This necessitates changing the association rules and data value(s) and relationships. It might be challenging to strike a real balance between disclosing and withholding [2], [4]. The implementation of hiding rules that reveal the sensitive portion of the data can help achieve this. Since most data users understand association among the data, one such strategy is to hide the association rule. If data is in the hands of a malevolent person, this association rule vulnerability poses a serious risk to the data.

If data is in the hands of a malevolent person, this association rule vulnerability poses a serious risk to the data.

There are two popular approaches to stop data from being exploited. Before sending it to the data miner, the first strategy modifies the date [3]. Using distributed database architecture, the second strategy merely releases a portion of the entire data. In the literature, algorithms for hiding rules have been proposed. In order to lower the confidence in the rules, the suggested algorithms work by altering the database transactions. Hiding association rule by using support and confidence is discussed in [3], Basically, the approach method conceals a particular rule, whereas the association rule mining [4] technique conceals rules about sensitive item(s) to the left or right of the rule. These methods, meanwhile, fall short of fully concealing all the desired rules, which should be concealed in the fewest possible passes.

In this research, we offer a set of techniques and strategies for minimally perturbing values based privacy preservation and knowledge concealment from data. The suggested method makes use of the data distortion methodology, in which the size of the database stays constant but the position of the sensitive item or items is changed, but their support is never modified.

The suggested heuristics hide the sensitive rules after first pruning the rules using the concept of representative rules. With respect to the specified limitations, this method generates a notably smaller number of rules while preserving the minimal set of pertinent association rules and the capacity to construct the full set of association rules [1], [2]. The benefit of the suggested strategy is that, unlike the

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 November 2023*

previous approaches, which either raised or lowered the size of the database, it maintains the same size while supporting the sensitive item or items in the same way. The sensitive item(s) support is maintained, but its location has been altered—that is, it is removed from one transaction and added to another where it does not exist. This strategy also has the benefit of hiding the greatest amount of rules with the fewest database modifications.

Additionally, an algorithm is suggested for this, and examples are given. The suggested method is contrasted with earlier methods that were in use [5].

This is the arrangement for the remainder of the paper. An explanation of association rule mining is given in Section 2. Section 3 gives an explanation of the association rule. Section 4 contains the problem definition. In section 5, the proposed plan is shown. Results of the scheme's simulation evaluation are presented in Section 6, and the work is wrapped up in Section 7.

2 Association Rule Mining

Let I = { i_1, i_2, … .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 -> Y where $X \subset I$ $Y \subset I$ and $X \cap Y = \phi$. X And Y are called antecedent/body and consequent/head of the rule respectively [6].

Support and trust in the rule are the two metrics used to determine if a rule is strong or not. A rule's level of interest can be determined using these two factors [5],[7].

For a given rule $X \Rightarrow Y$

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: $S = | X \cup Y | / |N|$

Where, N is the number of transactions.

Confidence is measured as the percentage of transactions covered by the LHS that are also covered by the RHS, or as the proportion of transactions that contain both of these elements.

$$C = |X \cup Y| / |X|$$

For the database given in Table1, with a minimum support of 33% and minimum confidence 70% following nine association rules could be found:

C => A (66.667%, 100%), A, B => C (50%, 75%),
B => C.A (50%, 75%), C.B=> A (50%, 100%)
C => A.B (50%, 75%), C.A=>B (50%, 75%)
B => C (50%, 75%), C=>B (50%, 75%)
B => A(66.667%, 100%)

Table 1. Set of transactional data

TID	ITEMS
T1	ABC
T2	ABC
T3	ABC
T4	AB
T5	A
T6	AC

3 Representative Association Rule

In general, a given database contains a relatively large number of associations. A significant portion of these regulations are found to be superfluous and unnecessary. All of these should be made available to users since they are unique, intriguing, and novel. This problem was addressed by [6], who proposed the idea of representative rules (RR), a condensed (loss less) representation of association rules. Without requiring access to a database, RR is the minimal collection of rules that permits the deduction of every association rule. In a notion of cover operator was introduced for driving a set of association rules from a given association rule. The cover C of the rule X=>Y, $Y \neq \phi$, is defined as follows:

$$C (X => Y) = (X \cup Y => V | Z.\ V \subseteq Y \text{ and } Z \cap V = \phi \text{ and } V \neq \phi\}$$

Each rule in C (X => Y) consists of a subset of items occurring in the rule X => Y . The number of different rules in the cover of the association X => Y is equal to $3^m - 2^m$, = | Y |.

In general, the process of generating representative rules may be decomposed in to two subprocesses: frequent item-sets generations and generation of RR from frequent item-sets. Let be a frequent itemset and $\phi \neq X \subset Z$. The association rule $X \Rightarrow Z/X$ is representative rule if there is no association rule ($X \Rightarrow Z\ '/X\ )'$ where $Z \subset Z'$, and there is no association rule ($X' \Rightarrow Z/X'$) such that $X \supset X'$. Formally, a set of representative rules (RR) for a given association rules (AR) can be defined as follows:

$$RR = \{ r \in AR \mid \neg \exists\ r' \in AR\ r' \neq r \text{ and } r \in C(r')\}$$

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 November 2023*

Each rule in RR is called representative association rule and no representative rule may belong in the cover of another association rule [8], [9].

4 Problem Definition

The term "data mining" refers to a broad range of instruments and methods for deriving valuable insights—which may include intriguing rules or sensitive information—from massive data sets. The purpose of this work is to suggest a novel approach to prevent sensitive data extraction. It is important to modify or distort data so that data mining tools are unable to find sensitive information. It becomes critical to safeguard data against unwanted access while working with sensitive information. Finding a solution to protect the privacy of revealed data while yet meeting the rightful demands of data consumers is the main challenge. Both approaches have advantages and disadvantages when it comes to the item's support. The proposed algorithm is based on altering the database transaction in order to decrease the trust in the rules.

A new algorithm is proposed in the next section to conceal sensitive rules (sensitive rules are those that contain sensitive item(s)).

5.1 Proposed approach

The focus of this proposed work centers on the final technique outlined in the paper, which involves concealing sensitive rules by adjusting the support and confidence thresholds of association rules or frequent item sets. Data mining primarily revolves around generating association rules, making it crucial to manipulate these parameters effectively. Much of the effort undertaken by data miners is dedicated to the exploration and creation of association rules. As association rules are widely recognized as significant entities, they possess the potential to compromise confidential information within businesses, defense sectors, or organizations. This necessity underscores the importance of concealing such information, specifically in the form of association rules. As highlighted in this paper, understanding associations within data is paramount for most data users, thereby mandating the modification of data values and relationships, particularly those governed by association rules. Researchers such as Saygin [1] and Wang [2] have proposed algorithms aimed at reducing the support and confidence levels of these rules. The following section delves into the approaches put forth by Saygin [1] and Wang [2], elucidating them through practical examples.

5.2 Proposed Algorithm

In this work, a novel approach to hiding association rules involves the concept of 'not altering the support' of sensitive items. This strategy forms the basis of a proposed algorithm, which assumes a set of sensitive items as input. The algorithm then distorts the original database in a manner that prevents sensitive rules from being discovered through Association Rule Mining algorithms. The proposed algorithm takes as input a database, minimum support (min_supp), minimum confidence (min_conf), and a sensitive item or set of items (each represented by $\overline{h}$) that need to be hidden. The objective is to distort the database in a way that prevents any association rules containing h either on the left-hand side or the right-hand side from being discovered.

ALGORITHM:

Input (1) A source database D

(2) A min_support.

(3) A min_confidence.

(4) A set of sensitive items H.

Output

A transformed database D' where rules containing H on RHS/LHS will be hidden

1. Find all large itemsets from D;
2. For each sensitive item $h \in H$ {
3. If h is not a large itemset then H=H- {*h*};
4. If H is empty then EXIT;
5. Select all the rules containing h and store in U

//*h* can either be on LHS or RHS

6. Select all the rules from U with *h* alone on LHS
7. Join RHS of selected rules and store in R;

//make representative rules

8. Sort R in descending order by the number of supported items;
9. Select a rule *r* from R
10. Compute confidence of rule *r*.
11. If conf>min_conf then {

//change the position of sensitive item h.

12. Find T_1={t in D|t completely supports *r* ;
13. If t contains x and h then
14. Delete h from t
15. Else
16. Go to step 19
17. Find T_1 ={t in D|t does not support LHS(*r*) and partially supports x;
18. Add h to t
19. Repeat
20. {
21. Choose the first rule from R;

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 November 2023*

22. Compute confidence of *r* ;
23. } Until(R is empty);
24. } //end of if conf>min_conf
25. Update D with new transaction *t*;
26. Remove h from H;
27. Store U [i] in R; //if LHS (U) is not same
28. i++, j++;
29. Go to step 7;
30. }//end of for each h∈H

After applying the proposed algorithm, the updated database D' is obtained. The algorithm begins by identifying all association rules that involve sensitive items, whether on the left-hand side (LHS) or right-hand side (RHS). These rules are then organized into a representative rules (RR) format. From the set of RR's, a rule is selected where the sensitive item appears on the left-hand side of the rule.

Now, remove the sensitive item(s) from the transaction that fully supports the representative rule (RR), meaning it contains all the items in the selected RR. Subsequently, add the same sensitive item to a transaction that partially supports the RR, where either some items in RR are absent or only one of them is present.

For example in Table1 at a min*_supp of 33% and a min*_conf of 70 % and sensitive item

H = {C }, choose all the rules containing 'C' either in RHS or LHS

C=>A(66.667%, 100%), A,B=> C (50%, 75%),

B=>C.A(50%, 75%), C,B=> A(50%, 100%),

C=>A.B (50%, 75%), C,A=> B(50%, 75%),

B =>C(50%, 75%), C=> B(50%, 75%)

And represent them in representative rule format.

Like

C → and C → B can be represented as C → AB

Now delete from a transaction where and are present and add to a transaction where both of them (A and B) are either absent or only one of them is present. If transaction T_1 is modified to and transaction T_5 is modified to AC then the rules that will be hidden are:

C → B , C →A , C →AB, B→ C, AB→ C, B→AC and AC → B

i.e. seven rules out of eight rules containing sensitive item(s) are hidden.

6 Results and Analysis

We conducted our experiments on an Intel P5 workstation equipped with a 500 MHz processor and 500 MB RAM, running the Microsoft Windows XP operating system using the -minor tool. The following section demonstrates the functionality of the aforementioned algorithm for concealing sensitive items, presenting some of the results obtained.

1. In Table 1, we conceal seven of the eight rules that contain sensitive item(s) for a specific database, provided that the transaction is modified to, with a minimum support of 33% and a minimum confidence of 70%. In the paper's section above, this has been covered in detail.

2. For database in Table 1, if H = {B} i.e. if sensitive item is B and the rule which is to be hidden is B ⇒ C (B⇒ A and B⇒ C represented as B ⇒ AC) then change transaction T1 to and transaction T5 to AB.

For a database given in Table 2 with a minimum support of 33% and minimum confidence of 70% following association rules are mined:

A⇒B , A⇒C , B⇒C, B⇒A , C⇒A , C⇒B , AD⇒C , CD⇒A

Select the rules containing sensitive items either in the LHS or in the RHS

A⇒C , B⇒C , B⇒A , C⇒A , C⇒B , AD⇒C , CD⇒A

Representation of rules in representative rule format is:

C⇒A and C⇒B can be represented as C⇒AB

Delete C from a transaction in which and are present and add it in a transaction where

both and are absent or only one of them is present

This results in modification of the database by changing the transaction T_2 to ABD and

transaction T_5 to CDE

Out of 6 rules containing sensitive items all of them are hidden.

Similarly for database in Table 2. if H= {B} *i.e.* and the rule which is to be hidden is B⇒ C (B⇒ A and B ⇒ C represented as B ⇒AC) if sensitive item is B then change T2 to ACD and T_5 ⇒ BDE transaction to and tra transaction.

Out of 4 rules containing sensitive items all of them are hidden.

Table 2 : Set of transactional data

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 November 2023*

TID	ITEMS
T1	ABC
T2	ABCD
T3	BCE
T4	ACDE
T5	DE
T6	AB

6. Analysis:

This section looks at and compares some of the properties of the proposed algorithm with the existing algorithms.

As mentioned in section 1 of the study, the first characteristic of the proposed algorithm is that the support of the sensitive item is not altered. The only thing that changes is the location of the sensitive object. Examples of this trait are provided in the preceding section, and Tables 3 and 4 provide a summary of it.

The effectiveness of the suggested algorithm in comparison to earlier methods is the second feature we examine.

For the suggested approach, four database scans are needed for Table 1, and seven rules must be trimmed.

Table 3. Database before and after hiding C and B

TID	D	D_1 (C sensitive)	D_2 (B sensitive)
T1	ABC	AB	AC
T2	ABC	ABC	ABC
T3	ABC	ABC	ABC
T4	AB	AB	AB
T5	A	AC	AB
T6	AC	AC	AC

Table 4. Database before and after hiding C and B

TID	D	D_1 (C sensitive)	D_2 (B sensitive)
T1	ABC	ABC	ABC
T2	ABCD	ABD	ACD
T3	BCE	BCE	BCE
T4	ACDE	ACDE	ACDE
T5	DE	CDE	BDE

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 November 2023*

T6	AB	AB	AB

The algorithm in [3] performs four database scans and prunes zero rules. Wang has an approach that involves trimming two rules and performing three database scans. More rules are pruned in the same number of database scans using the recommended method, as can be seen from both tables [3]. This feature is displayed in Table 5, while Table 6 provides a summary of the same features for the Table 2 database.

One of the main causes of the failure of current methods is that they try to conceal every rule in a set without taking into account the possibility of trimming rules after making changes to specific transactions.

Only those rules that have sensitive items on the left or right are hidden. It utilizes two distinct algorithms, depending on whether the sensitive item is subsequent or antecedent. Furthermore, this tactic prevents the hiding of additional rules.

Table 5. Database scans and rules pruned in hiding item C using proposed algorithm

	DB Scans	Rules Pruned	
		Table 1	Table 2
Proposed Algorithm	4	7	6
ISLF	3	2	3
[3] Dasseni et al.	4	0	1

Table 6. Database scans and rules pruned in hiding item B using proposed algorithm

	DB Scans	Rules Pruned	
		Table 1	Table 2
Proposed Algorithm	4	6	4
ISLF	3	2	2
[3] Dasseni et al.	4	1	1

However, proposed approach hides almost all the rules, which contain sensitive item(s) (either on the left or on the right [11], [12],[13], [14] and [15].

7 Conclusion

In this article, we discussed the risks to database security and privacy brought about by the quick development of data mining. The phrase "data mining" refers to a broad spectrum of instruments and methods for obtaining valuable information—which may include sensitive information like intriguing rules—from a sizable body of data. One crucial data-mining operation that uncovers intriguing relationships among a sizable collection of data objects is association rule mining. It might be a threat to the privacy of uncovered confidential information since it might reveal patterns and other sensitive knowledge types that are hard to obtain in other ways. Preventing unwanted access to such data is the goal. Our work aims to suggest a novel approach to prevent sensitive data from being extracted. In order to prevent sensitive information from being found via data mining techniques, data should be twisted or modified. The concerns that the rapid expansion of data mining and related technologies poses to database security and privacy are covered in this study.

An approach that was previously described is covered in the proposed work along with some of its methods. It explains

the limitations of the current methods and analyzes them. Drawing on a critical analysis, this study presents a novel algorithm that employs the concept of representative rules to conceal sensitive rules, thereby encapsulating sensitive information in the form of association rules.

Unlike existing algorithms that either increase or decrease the support of the sensitive item to modify the database transactions, the proposed approach takes a different approach to modify the database transactions so that the confidence of the sensitive rules can be reduced without changing the support of the sensitive item. The suggested approach's effectiveness is contrasted with the methods that are currently in use.

In terms of the amount of database scans and hidden rules, the suggested algorithm outperforms the current methods.

8. Future Work

In this work confidence the rules, which could be represented as representative rules, is also recomputed even if the confidence of the RR falls below the min_conf threshold. There is a need to find out a method, which can avoid the computation of the confidence of the rules from which the RR is made i.e. if the confidence of the RR falls below the min_conf threshold then the rules from which this RR evolved should not be computed.

References

[1] V. S. Verykios, Ahmed K. Elmagermld, Elina Bertino, Yucel Saygin, Elena Dasseni, "Association Rule Hiding," IEEE Transactions on knowledge and data engineering, vol. 6, no.4, (2004).

[2] S-L. Wang, Yu-Huei Lee, S.Billis and A. Jafari, "Hiding sensitive items in privacy preserving association rule mining," IEEE International Conference on Systems, Man and Cybernetics, vol. 4, pp. 3239 – 3244, (2004).

[3] E. Dasseni, V. Verykios, A. Elmagarmid and E. Bertino, "Hiding Association Rules by Using Confidence and Support," In Proceedings of 4th Information Hiding Workshop, Pittsburgh, PA, pp. 369-383, (2001).

[4] V. S. Verykios, A. K. Elmagarmid, B. Elisa, D. Elena, and Y. Saygin, "Association Rule Hiding," IEEE Transactions on Knowledge and Data Engineering, pp. (2000).

[5] S-L. Wang and A. Jafari, "Using unknowns for hiding sensitive predictive association rules," In IEEE International Conference on Information Reuse and Integration, pp. 223 – 228, (2005).

[6] Marzena Kryszkiewicz. "Representative Association Rules", In proceedings of PAKDD'98. Melbourne, Australia (Lecture notes in artificial Intelligence,LANI 1394, Springer-Verleg, pp 198-209, (1998).

[7] Yucel Saygin, Vassilios S. Verykios, Chris Clifton. "Using unknowns to prevent discovery of association rules", ACM SIGMOD Record Volume 30 Issue 4, pp. 45 - 54, (2001).

[8] Yiqun Huang, Zhengding Lu, Heping Hu, "A method of security improvement for privacy preserving association rule mining over vertically partitioned data", 9th International Database Engineering and Application Symposium, pp. 339 – 343, (2005).

[9] Saygin Y., Verykios V.S. and Elmagarmid A.K., "Privacy preserving association rule mining," IEEE Proceedings of the 12th Int'l Workshop on Research Issues in Data Engineering, pp. 151 – 158, (2002).

[10] Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke, "Privacy preserving mining of association rules", In Proc. Of the 8th ACM SIGKDD Int'l Conference on Knowledge Discovery and Data Mining, Edmonton, Canada, (2002).

[11] R. Agrawal and R. Srikant, "Privacy preserving data mining", In ACM SIGMOD Conference on Management of Data, pages 439-450, Dallas, Texas, (2000).

[12] Agrawal R., Imielinski T. and Swami A., "Mining Association Rules Between Sets of ITEMS in Large Databases", In Proceedings Of ACM SIGMOD International Conference on Management of Data, Washington D.C. May 1993,pp.207-216.

[13] Sun D., Teng S., Zhang W. and Zhu H.: An Algorithm to Improve the Effectiveness of Apriori, In the 6th IEEE Int. Conf. on Cognitive Informatics (ICCI'07), 385-390, (2007).

[14] Ping D. and Yongping G.: A New Improvement of Apriori Algorithm for Mining Association Rules. In the International Conference on Computer Application and System Modeling (ICCASM 2010), 529-532, (2010).

[15] Lu-Feng W:Association Rule Mining Algorithm Study and Improvement, In the 2nd International Conference on Software Technology & Engineering (ICSTE),362-364, (2010).

[16] Benny Pinkas ,HP Labs, Cryptographic Technique for Privacy Preserving Datamining, SIGKDD Explorations, Vol.4, Issue 2, 2002.

[17] Chirs Clifton, Murat Kantarcioglu Jaideep Vaidya, Purdue University, DCS, Xiaodong Lin ,Michael Y.Zhu, Purdue University, DOS, Tool for Privacy Preserving Distributed Data Mining ,Vol .4, Issue –2.

[18] Aggarwal C., Pei J., Zhang B. A Framework for

Privacy Preservation against Adversarial Data Mining. ACM KDD Conference, 2006.

[19] M. Evfimievski, "Randomization in Privacy Preserving Data Mining", Proceedings of the 9th ACM International Conference on Knowledge Discovery and Data Mining, (2008), pp. 43-48.

[20] A. P. Felty and S. Matwin, "Privacy-Oriented Data Mining by Proof Checking", Proceedings of the 6th European Conference on Principles of Data Mining and Knowledge Discovery, (2007) December 23-25, pp. 138-149.

[21] Y. Guo (2007), "Reconstruction-Based Association Rule Hiding", Proceedings of the Workshop on Innovative Database Research, (2007) March 12-14, pp. 511-543.

[22] J. Natwichai, X. Sun and Xue, "A Heuristic Data Reduction Approach for Associative Classification Rule Hiding", Proceedings of the 10th Pacific Rim international conference on artificial intelligence, 2008) July, pp. 140-151.