Hybrid Algorithm for Privacy Preserving Association Rule Mining

Ila Chandrakar
Assistant Professor, Department of Information Technology
VNR Vignana Jyothi Institute of Engineering and Technology
Hyderabad, India

Abstract—Privacy-preserving data mining, is a novel research direction in data mining and statistical databases, where data mining algorithms are analyzed for the side effects they incur in data privacy. For example, through data mining, one is able to infer sensitive information, including personal information or even patterns, from non sensitive information or unclassified data. There have been two types of privacy concerning data mining. The first type of privacy is that the data is altered so that the mining result will preserve certain privacy. The second type of privacy is that the data is manipulated so that the mining result is not affected or minimally affected.

Given specific rules to be hidden, many data altering techniques for hiding association, classification and clustering rules have been proposed. However, to specific hidden rules, entire data mining process needs to be executed. For some applications, we are only interested in hiding certain sensitive items. In this work, we assume that only sensitive items are given and propose one algorithm to modify data in database so that sensitive items whether in LHS (Left Hand Side) or RHS (Right Hand Side) of the rule cannot be inferred through association rule mining algorithms. The efficiency of the proposed approach is further compared with ISL (Increase Support of Left Hand Side) approach. It is shown that our approach prunes more number of rules.

Keywords—association rule mining, privacy preserving.

I. INTRODUCTION (HEADING 1)

Privacy preserving data mining is a novel research direction in data mining and statistical databases where data mining algorithms are analyzed for the side-effects they incur in data privacy. Here is the introduction to data mining and association rule mining and later on “privacy preserving the association rule mining” is explored in more details, which is the base of this research.

A. Data Mining

Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their databases. Association rule induction is a powerful method for so-called market basket analysis, which aims at finding regularities in the shopping behavior of customers of supermarkets, mail-order companies, on-line shops and the like. For example, a famous Indian supermarket named Big Bazaar uses association rules for deciding their marketing strategies like offers should be given in which products, which products should be placed together in shelves [9].

B. Background And Related Work

The concept of privacy preserving data mining has been recently been proposed in response to the concerns of preserving personal information from data mining algorithms [1]. There have been two broad approaches. The first approach is to alter the data before delivery to the data miner so that real values are obscured. One technique of this approach is to selectively modify individual values from a database to prevent the discovery of a set of rules [2, 3, 4, 7]. They apply a group of heuristic solutions for reducing the number of occurrences (support) of some frequent (large) item sets below a minimum user specified threshold. The advantage of this technique is that it maximizes the amount of available data, although it does not ensure the integrity of the data. The second type of privacy is that the data is manipulated so that the mining result is not affected or minimally affected [5, 6, 11].

Given specific rules to be hidden, many data altering techniques for hiding association, classification and clustering rules have been proposed. However, to specify hidden rules, entire data mining process needs to be executed. For some applications, we are only interested in hiding certain sensitive items that appeared in association rules. In this work, we assume that only sensitive items are given and propose one hybrid algorithm based on already proposed ISL algorithm to modify data in database so that sensitive items cannot be inferred through association rules mining algorithms. The proposed algorithm is based on modifying the database transactions so that the confidence of the association rules can be reduced. The efficiency of the proposed approach is further compared with ISL algorithm [3, 4]. It is shown that our approach prunes more number of rules.

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 the proposed algorithm for hiding association rules that contain the specified sensitive items. Section 4 shows an example of the proposed algorithm. Section 5 analyses the result of the efficiency of proposed
algorithm and further compared with ISL approach. Concluding remarks and future works are described in section 6.

II. PROBLEM STATEMENT

A. Mining of Association Rules

The problem of mining association rules was introduced in [10]. Let \( I = \{ i_1, i_2, ..., i_n \} \) be a set of literals, called items. Given a set of transactions \( D \), each transaction \( t \) is a set of items such that \( T \subseteq I \), an association rule is an expression \( X \Rightarrow Y \) where \( X \subseteq I \), \( Y \subseteq I \) and \( X \cap Y = \phi \). The X and Y are called respectively the body (left hand side) and head (right hand side) of the rule. An example of such a rule is that 90% of customers buy milk also buys bread. The 90% here is called the confidence of the rule, which means that 90% of transaction that contains X also contains Y. The confidence is calculated as \( \frac{|X \cap Y|}{|X|} \). The support of the rule is the percentage of transactions that contain both X and Y, which is calculated as \( \frac{|X \cap Y|}{|N|} \). In other words, the confidence of a rule measures the degree of the correlation between item sets, while the support of a rule measures the significance of the correlation between item sets. The problem of mining association rules is to find all rules that are greater than the user-specified minimum support and minimum confidence.

| TID | Items |
|-----|-------|
| T1  | A B C |
| T2  | A B C |
| T3  | A B C |
| T4  | A B  |
| T5  | A    |
| T6  | A C  |

As an example, for a given database in table 1, a minimum support of 33% and a minimum confidence of 70%, four association rules can be found as follows: \( A \Rightarrow B \) (66%), \( B \Rightarrow A \) (100%), \( A \Rightarrow C \) (66%), \( C \Rightarrow A \) (100%), \( B \Rightarrow C \) (75%), \( C \Rightarrow B \) (50%).

B. Problem Description

The objective of data mining is to extract hidden or potentially unknown interesting rules or patterns from databases. However, the objective of privacy preserving data mining is to hide certain sensitive information so that they cannot be discovered through data mining techniques. In this work, we assume that only sensitive items are given and propose one algorithm to modify data in database so that sensitive items cannot be inferred through association rules mining algorithms. More specifically, given a transaction database \( D \), a minimum support, a minimum confidence and a set of items \( H \) to be hidden, the objective is to modify the database \( D \) such that no association rules containing \( H \) on the right hand side or left hand side will be discovered.

The following notation will be used in the paper. Each database transaction has three elements: \( T=<\text{TID}, \text{list-of-elements}, \text{size}> \). The TID is the unique identifier of the transaction \( T \) and list-of-elements is a list of all items in the database. However, each element has value 1 if the corresponding item is supported by the transaction and 0 otherwise. Size means the number of elements in the list-of-elements having value 1. For example, if \( I = \{ A, B, C \} \), a transaction that has the items \( \{ A, C \} \) will be represented as \( t = <T1, 101, 2> \). In addition, a transaction \( t \) supports an item set \( I \) when the elements of \( t \).list-of-elements corresponding to items of \( I \) are all set to 1. A transaction \( t \) partially supports an item set \( I \) when the elements are not all set to 1. For example, if \( I = \{ A, B, C \} = \{111\}, p=<T1, \{111\}, 3>, \) and \( q=<T2, \{001\}, 1> \), then we would say that \( p \) supports \( I \) and \( q \) partially supports \( I \).

III. PROPOSED ALGORITHM

In order to hide an association rule, we can either decrease its support or its confidence to be smaller than pre-specified minimum support and minimum confidence. To decrease the confidence of a rule, we can either (1) increase the support of \( X \), i.e., the left hand side of the rule, but not support of \( X \cap Y \), or (2) decrease the support of the item set \( X \cap Y \) [4]. For the second case, if we only decrease the support of \( Y \), the right hand side of the rule, it would reduce the confidence faster than simply reducing the support of \( X \cap Y \). To decrease support of an item, we will modify one item at a time by changing from 1 to 0 or from 0 to 1 in a selected transaction.

Based on these two strategies, we propose one data-mining algorithm for hiding sensitive items in association rules called hybrid algorithm. This algorithm first tries to hide the rules in which item to be hidden i.e. \( X \) is in right hand side and then tries to hide the rules in which \( X \) is in left hand side. For this algorithm \( t \) is a transaction, \( T \) is a set of transactions, \( U \) is used for rule, \( \text{RHS}(U) \) is Right Hand Side of rule \( U \), \( \text{LHS}(U) \) is the right hand side of the rule \( U \), \( \text{Confidence}(U) \) is the confidence of the rule \( U \).

Hybrid Algorithm:

**Input:**
(1) A source database \( D \),
(2) A minimum support \( \text{min\_support} \),
(3) A minimum confidence \( \text{min\_confidence} \),
(4) A set of hidden items \( X \),

**Output:** A transformed database \( D \), where rules containing \( X \) on \( \text{LHS} \) (Left Hand Side) or \( \text{RHS} \) (Right Hand Side) will be hidden.
Steps of Algorithm:
1. Find all possible rules from given items X;
2. Compute confidence of all the rules.
3. For each hidden item h
4. For each rule containing h, Compute confidence of rule U
5. For each rule U in which h is in RHS
5.1. If confidence (U) < min conf , then Go to next large 2-itemset; Else go to step 6
6. Decrease Support of RHS i.e. item h.
6.1. Find T = t in D | t fully support U ;
6.2. While (T is not empty)
6.2.1. Choose the first transaction t from T;
6.2.2. Modify t by putting 0 instead of 1 for RHS item;
6.2.3. Remove and save the first transaction t from T; End While
6.3. Compute confidence of U;
6.4. If T is empty, then h cannot be hidden; End For
7. For each rule U in which is in LHS
8. Increase Support of LHS;
8.1. Find T = t in D | t does not support U ;
8.2. While (T is not empty)
8.2.1. Choose the first transaction t from TR;
8.2.2. Modify t by putting 1 instead of 0 for LHS item;
8.2.3. Remove and save the first transaction t from T; End While
8.3. Compute confidence of U;
8.4. If T is empty, then h cannot be hidden; End For End Else
End For
Output updated D, as the transformed D;

IV. EXAMPLE

This section shows an example for demonstrating the proposed algorithm in hiding sensitive items in association rules mining. Consider table 1 as a database, a minimum converted database according to the specified notation given in section 2.2 is shown in table 2.

| TID | ABC | Size |
|-----|-----|------|
| T1  | 111 | 3    |
| T2  | 111 | 3    |
| T3  | 111 | 3    |
| T4  | 110 | 2    |
| T5  | 100 | 1    |
| T6  | 101 | 2    |

The all possible rules with confidences are: A→B (66.66%) (less), A→C (66.66%)(less), B→A (100%)(greater), B→C (75%)(less), C→A (100%)(greater), C→B (75%)(greater).

Suppose we first want to hide item A, for this, first take rules in which A is in RHS. These rules are B→A and C→A and both have greater confidence.
First take rule B→A and search for transaction which supports both B and A i.e. B=A=1. There are four transactions T1, T2, T3, T4 with A=B=1. Now update the table as follows: Put 0 for item A in all the four transactions. After this modification, we get table 3 as the modified table. Now calculate confidence of B→A, it is 0% which is less than minimum confidence so now this rule will be hidden. Now take rule C→A, search for transactions in which A=C=1, only transaction T6 has A=C=1, update transaction by putting 0 instead of 1 in place of A. Now calculate confidence of C→A, it is 0% which is less than the minimum confidence so now this rule will be hidden. Now take the rules in which A is in LHS. There are two rules A→B and A→C but both rules have confidence less than minimum confidence so there is no need to hide these rules.

To hide item B, first take rules in which B is in RHS. These rules are A→B and C→B. But only rule C→B has confidence greater than minimum confidence. So search for transaction having B=C=1. Using same procedure as above, table 5 will be the updated table. Now calculate the confidence of rule C→B, it is 0%, which is less than minimum confidence so now this rule will be hidden. Now take rules in which B is in LHS. These are B→A and B→C. But B→A is already hidden so take rule B→C. For hiding this rule, search for transaction which doesn’t support both B and C i.e. B=C=0. Transaction T5 has B=C=0. Update the table as put 1 in place of 0 for B. The table 6 is the updated table. Now calculate the confidence of rule B→C, it is 0%, which is less than the minimum confidence so this rule will be hidden.

To hide item C, first take rules B→C and A→C. Both rules are already hidden. Now take rules C→A and C→B. Both rules are already hidden. So in all, our hybrid algorithm has hidden four rules.

V. EXPERIMENT AND RESULTS

In this section, we will compare the performance of hybrid algorithm with ISL algorithm in terms of number of rules pruned and number of times database scanned. Proposed hybrid algorithm will try to hide all the rules in which item to be hidden is present. But ISL algorithm will try to hide only those rules in which item to be hidden are in LHS. Table 7 shows the comparison between the algorithms for database D.

| Algorithm | Number of Rules Pruned | Number of Times Database Scanned |
|-----------|------------------------|---------------------------------|
| ISL       | 1                      | 4                               |
| Hybrid    | 4                      | 4                               |

In our research work, we applied ISL and Hybrid Algorithm in a real database called Teaching Assistant Evaluation(TAE) taken from the website of University of
California [12] and the database transa.txt which is used in implementation of Apriori algorithm by University of Regina [13]. For both the database, we have taken minimum confidence = 60%. The results of both the algorithms are shown in table 8. The reason why hybrid approach prunes more number of rules is that it tries to prune all the rules whether item to hide is in LHS or RHS first then it will try for another item.

### TABLE VIII. COMPARISON OF ALGORITHMS FOR TAE DATABASE

| Algorithm | Number of Rules Pruned | Number of Times Database Scanned |
|-----------|------------------------|---------------------------------|
| ISL       | 1                      | 6                               |
| Hybrid    | 6                      | 6                               |

### VII. CONCLUSION AND FUTURE WORK

In this work, we have proposed one algorithm for hiding sensitive data in association rules mining which is a hybrid approach of previous algorithms and based on modifying the database transactions so that the confidence of the association rules can be reduced. The efficiency of the proposed approach is further compared with ISL approach and shown that this approach prunes more number of hidden rules with same number of times database scanned. In future, better algorithm should be developed which will prune all the sensitive rules with less number of database scans.

### REFERENCES

[1] Benny Pinkas, Yehuda Lindell.: Privacy Preserving Data Mining. In Proceedings of the 20th Annual International Cryptology Conference on Advances in Cryptology, pages 36-54. (2000)

[2] Ahmed K. Elmagarmid, Vassilios S. Verykios and Yucel Saygin.: Privacy Preserving Association Rule Mining. In Proceedings of the 12th International Workshop on Research Issues in Data Engineering: Engineering E-Commerce/E-Business Systems, Washington, USA. (2002)

[3] Ayat Jafari, Shyue-Liang Wang.: Hiding Sensitive Predictive Association Rules”.IEEE International Conference on Systems, Man and Cybernetics, 164-169. (2005)

[4] Shyue-Liang Wang, Yu-Huei Lee, Steven Billia, Ayat Jafari, “Hiding Sensitive Items in Privacy Preserving Association Rule Mining”

[5] Jaideep Vaidya , Chris Clifton.: Privacy preserving association rule mining in vertically partitioned data. In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. July 23-26, Edmonton, Alberta, Canada.( 2002)

[6] Hillol Kargupta and Souptik Datta and Qi Wang and Krishnamoorthy Sivakumar.: On the Privacy Preserving Properties of Random Data Perturbation Techniques. In Proceedings of the Third IEEE International Conference on Data Mining, Washington, Washington, USA, 99. (2003)

[7] Vassilios S. Verykios , Ahmed K. Elmagarmid , Elisa Bertino , Yucel Saygin ,Elena Dasseni.: Association Rule Hiding. IEEE Transactions on Knowledge and Data Engineering, 434-447. (2004)

[8] Rakesh Agrawal,Ramakrishnan Srikant.: Fast Algorithms for Mining Association Rules. Morgan Kaufmann Series In Data Management Systems, Readings in database systems, 580-592. (1998)

[9] Kishore Biyani and Dipayan Baishya.: It Happened in India. New Delhi, India, Rupa and Co. (2007)

[10] Rakesh Agrawal , Tomasz Imieliński , Arun Swami.: Mining association rules between sets of items in large databases, Proceedings of the 1993 ACM SIGMOD international conference on Management of data, Washington, D.C., United States, 207-216. (1993)

[11] Alexandre Evfimievski , Ramakrishnan Srikant, Rakesh Agrawal, Johannes Gehrke.: Privacy preserving mining of association rules, In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, Edmonton, Alberta, Canada. (2002)

[12] http://www.ics.uci.edu/~mlearn/MLSummary.html - University of California at Irvine Machine Learning Repository.

[13] http://www2.cs.uregina.ca/~dbd/cs831/notes/itemsets/transa.txt - Website of University of Regina.