Journal of Computer Science 6 (12): 1494-1498, 2010
ISSN 1549-3636

Hybrid Algorithm for Privacy Preserving Association Rule Mining

Ila Chandrakar, Yelipe Usha Rani, Mortha Manasa and Kondabala Renuka
Department of Information Technology,
VNR Vignana Jyothi Institute of Engineering and Technology, Hyderabad, India

Abstract: Problem statement: The objective of the hybrid algorithm for privacy preserving data mining is to hide certain sensitive information so that they cannot be discovered through association rule mining techniques. **Approach:** The sensitive items whether in Left Hand Side (LHS) or Right Hand Side (RHS) of the rule cannot be inferred through association rule mining algorithms by combining the concept of Increase Support of Left Hand Side (ISL) and Decrease Support of Right Hand Side (DSR) algorithms i.e., by increasing and decreasing the support of the LHS and RHS item of the rule respectively. **Results:** The efficiency of the proposed approach is compared with alone Increase Support of Left Hand Side (ISL) approach for real databases on the basis of number of rules pruned. **Conclusion:** The hybrid approach of ISL and DSR algorithms prunes more number of sensitive rules with same number of database scans.

Key words: Hybrid algorithm, association rule mining, privacy preserving, mining algorithms, sensitive items, hiding association

INTRODUCTION

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 (Evfimievski *et al*., 2002). 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.

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 (Razali and Ali, 2009). 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.

The concept of privacy preserving data mining has been proposed in response to the concerns of preserving personal information from data mining algorithms (Saygin *et al*., 2002; Vaidya *et al*., 2008). 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. 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 (Liu *et al*., 2008; Yang *et al*., 2005). 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.

Given specific rules to be hidden, many data altering techniques for hiding association, classification and clustering rules have been proposed (Inan and Saygin, 2006; Poovammal and Ponnavaikko, 2009; Verykios *et al*., 2004 ; Kargupta *et al*., 2003; Inan *et al*., 2006). 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

Corresponding Author: Ila Chandrakar, Department of Information Technology,
VNR Vignan Jyothi Institute of Engineering and Technology Hyderabad, India

transactions so that the confidence of the association rules can be reduced. The efficiency of the proposed approach is further compared with ISL algorithm (Agrawal and Srikant, 1998; Wang and Jafari, 2005; Wang *et al*., 2004; 2007). It is shown that our approach prunes more number of rules.

The problem of mining association rules was introduced in (Yang *et al*., 2005). Let I = {i_1, i_2... i_n} be a set of literals, called items. Given a set of transactions D, where each transaction T is a set of items such that T⊆I, an association rule is an expression X⇒Y where X⊆I, Y⊆I and X∩Y = ϕ. 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 |X∪Y|÷|X|. The support of the rule is the percentage of transactions that contain both X and Y, which is calculated as|X∪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.

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 = >B (66%), B = >A (l00%), A = >C (66%), C = >A (l00%), B = >C (75%), C = >B (50%).

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.

Table 1: Database D

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

The following notation will be used in the paper. Each database transaction has three elements: T = <TID, list-of-elements, 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.

MATERIALS AND METHODS

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 o of X, i.e., the left hand side of the rule, but not support of X∪Y, or (2) decrease the support of the item set X∪Y (Poovammal and Ponnavaikko, 2009). 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∪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, RHS(U) is Right Hand Side of rule U, LHS(U) is the right hand side of the rule U, Confidence(U) is the confidence of the rule U.

Hybrid algorithm:

Input:
(1) A source database D,
(2) A minimum support min_support,
(3) A minimum confidence min_confidence,
(4) A set of hidden items X.

Output: A transformed database D, where rules containing X on Left Hand Side (LHS) or Right Hand Side (RHS) 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;

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 is shown in Table 2.

Table 2: Database D using the specified notation

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 3: 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 is 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 is 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. So Table 4 shows the modified database after hiding item A.

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.

Table 3: Updated table

TID	ABC	Size
T1	011	2
T2	011	2
T3	011	2
T4	010	2
T5	100	1
T6	101	2

Table 4: Updated table after hiding item A

TID	ABC	Size
T1	011	2
T2	011	2
T3	011	2
T4	010	1
T5	100	1
T6	001	1

Table 5: Updated table

TID	ABC	Size
T1	001	1
T2	001	1
T3	001	1
T4	010	1
T5	100	1
T6	001	1

Table 6: Updated table after hiding item B

TID	ABC	Size
T1	001	1
T2	001	1
T3	001	1
T4	010	1
T5	100	1
T6	001	1

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.

RESULTS

In results, we compare the performance of hybrid algorithm with ISL algorithm in terms of number of rules pruned and number of times database scanned for checking and updating the database. Proposed hybrid algorithm tries to hide all the rules in which item to be hidden is present. But ISL algorithm tries 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.

In our research, we applied ISL and Hybrid Algorithm in a real database called Teaching Assistant Evaluation (TAE) taken from the website of University of California and the database transa.txt version 2 which is used in implementation of Apriori algorithm by University of Regina. For both the database, we have taken minimum confidence = 60%. The results of both the algorithms for TAE database 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.

The Table 9 shows the results obtained from the ISL and Hybrid algorithm for the database transa.txt having details of transaction of a retail shop and used for Apriori algorithm.

Table 7: Comparison of algorithms for database D

Algorithm	No. of rules pruned	No. of times database scanned
ISL	1	4
Hybrid	4	4

Table 8: Comparison of algorithms for database TAE

Algorithm	No. of rules pruned	No. of times database scanned
ISL	1	6
Hybrid	6	6

Table 9: Comparison of algorithms for database transa.txt

Algorithm	No. of rules pruned	No. of times database scanned
ISL	9	15
Hybrid	15	15

DISCUSSION

Data mining: Data mining is the process of extracting patterns from data. Data mining is seen as an increasingly important tool by modern business to transform data into business intelligence giving an informational advantage.

Association rules: Association rules are statements of the form {X1,X2,….Xn }–>Y, meaning that if we find all of X1;X2; : : :;Xn in the market basket, then we have a good chance of finding Y.

Support of the rule: The support supp(X) of an item set X is defined as the proportion of transactions in the data set which contain the item set.

Confidence of the association rule: Confidence is the ratio of the number of transactions that include all items in the consequent as well as the antecedent (namely, the support) to the number of transactions that include all items in the antecedent:

$$\text{Conf}(X \rightarrow Y) = \text{supp}(XUY) / \text{supp}(X)$$

Privacy preserving data mining: Privacy Preserving Data Mining is a research area concerned with the privacy driven from personally identifiable information when considered for data mining. It provides security to protect data.

The hybrid algorithm for privacy preserving mining can be specially used for big retail shops, supermarket of different products like groceries, clothes etc. In this type of organization, they use association rule mining to mine the pattern of the sale like maximum which items customer is buying together. Using this information, they can decide their marketing strategies like to which

products together they should give offer or which items they should keep in nearby racks in the shop etc. But this information should not be revealed to the unauthorized person through association rule mining otherwise other competitors can use their information for their profit. So the proposed hybrid algorithm hides all of these sensitive rules from the unauthorized user by selectively modification of the database.

CONCLUSION

In this study, 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 can be developed which will prune all the sensitive rules with less number of database scans.

REFERENCES

Agrawal, R. and R. Srikant, 1998. Fast Algorithms for Mining Association Rules. In: Readings in Database Systems, Stonebraker, M. and J. Hellerstein (Eds.). Morgan Kaufmann, Massachusetts, ISBN: 1558605231, pp: 580-592.

Evfimievski, A., R. Srikant, R. Agrawal and J. Gehrke, 2002. Privacy preserving mining of association rules. Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, July 23-25, ACM Press, Edmonton, AB., Canada, pp: 1-12. DOI: 10.1145/775047.775080

Inan, A. and Y. Saygin, 2006. Privacy preserving spatio-temporal clustering on horizontally partitioned data. Lecture Notes Comput. Sci., 4081: 459-468. DOI: 10.1007/11823728_44

Inan, A., Y. Saygyn, E. Savas, A.A. Hintoglu and A. Levi, 2006. Privacy preserving clustering on horizontally partitioned data. Proceedings of the 22nd International Conference on Data Engineering Workshops, (DEA'06), IEEE Xplore Press, Atlanta, GA., USA., pp: 95-95. DOI: 10.1109/ICDEW.2006.115

Kargupta, H., S. Datta, Q. Wang and K. Sivakuma, 2003. On the privacy preserving properties of random data perturbation techniques. Proceedings of the 3rd IEEE International Conference on Data Mining, Nov. 19-22, IEEE Xplore Press, Washington, USA., pp: 99-106. DOI: 10.1109/ICDM.2003.1250908

Liu, L., M. Kantarcioglu, B. Thuraisingham, 2008. The applicability of the perturbation based privacy preserving data mining for real-world data. Data Knowl. Eng., 65: 5-21. DOI: 10.1016/j.datak.2007.06.011

Poovammal, E. and M. Ponnavaikko, 2009. Utility independent privacy preserving data mining on vertically partitioned data. J. Comput. Sci., 9: 666-673. DOI: 10.3844/jcssp.2009.666.673

Razali, A.M. and S. Ali, 2009. Generating treatment plan in medicine: A data mining approach. Am. J. Applied Sci., 6: 345-351. DOI: 10.3844/ajassp.2009.345.351

Saygin, Y., V.S. Verykios and A.K. Elmagarmid, 2002. Privacy preserving association rule mining. Proceedings of the 12th International Workshop on Research Issues in Data Engineering: Engineering E-Commerce/E-Business Systems, Feb. 24-25, IEEE Xplore Press, San Jose, CA., USA., pp: 151-158. DOI: 10.1109/RIDE.2002.995109

Vaidya, J., H. Yu and X. Jiang, 2008. Privacy-preserving SVM classification. Knowl. Inform. Syst., 14: 161-178. DOI: 10.1007/s10115-007-0073-7

Verykios, V.S., A.K. Elmagarmid, E. Bertino, Y. Saygin and E. Dasseni, 2004: Association rule hiding. IEEE Trans. Knowl. Data Eng., 16: 434-447. DOI: 10.1109/TKDE.2004.1269668

Wang, S.L. and A. Jafari, 2005. Hiding sensitive predictive association rules. Proceedings of IEEE International Conference on Systems, Man, Cybernetics, Oct. 10-12, IEEE Xplore Press, USA., pp: 164-169. DOI: 10.1109/ICSMC.2005.1571139

Wang, S.L., Y.H. Lee, S. Billis and A. Jafari, 2004. Hiding sensitive items in privacy preserving association rule mining. Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, Oct. 10-13, IEEE Xplore Press, USA., pp: 3239-3244. DOI: 10.1109/ICSMC.2004.1400839

Wang, S.L., B. Parikh and A. Jafari, 2007. Hiding informative association rule sets. Exp. Syst. Appli., 33: 316-323. DOI: 10.1016/j.eswa.2006.05.022

Yang, Z., S. Zhong and R.N. Wright, 2005. Privacy-preserving classification of customer data without loss of accuracy. Proceeding of the 5th SIAM International Conference on Data Mining, Apr. 21-23, National Science Foundation, Newport Beach, CA., pp: 1-11.