Association Rule Hiding Using ISL and DSR

Authors: Priyansh Mishra, Shashwat Singh***

*Author addresses: *Vellore Institute of Technology, Vellore, Tamil Nadu, 632014, **Vellore Institute of Technology, Vellore, Tamil Nadu, 632014*
Email: shashwatsingh.2016@vitstudent.ac.in, priyansh.pm@gmail.com

Abstract: *Privacy preserving data mining is a novel research direction in data mining and statistical database where data mining algorithm are analyzed for the side-effects they incur in data privacy. The security of the large database that contains certain crucial information, it will become a serious issue when sharing data to the network against unauthorized access. Association analysis is a powerful tool for discovering relationships which are hidden in large database. Association rules hiding algorithms get strong and efficient performance for protecting confidential and crucial data. Data modification and rule hiding is one of the most important approaches for secure data. One of the techniques in privacy preservation selectively modifies individual values from a database to prevent discovery of a set of rules. There are two known algorithms for it, ISL(Increase Support of Left) and DSR(Decrease Support of Right) both makes use of user specified values for MST(Minimum Support Threshold) & MCT(Minimum Confidence Threshold) as Input. The efficiency of the proposed algorithm is compared with ISL and DSR algorithms using real database, on the basis of number of rules hide, CPU time and number if transaction and got better results.*

Keywords: *Association Rules; Apriori Algorithm; ISL and DSR; Association Rule Hiding; Hybrid ISL and DSR; Privacy; Sensitive Items;*

1. Introduction

Data mining is known as knowledge discovery process of analyzing data from different point of views and to work out into useful information which can be applied in various application, including advertisement, bioinformatics, database marketing, fraud detection, e-commerce, health care, security web, financial forecasting etc[1]. Privacy preserving data mining(PPDM) provides solutions to the problem of maintaining the privacy of data as well as knowledge. It allows extraction of knowledge and also prevents the sensitive data or information from disclosure. PPDM algorithms refer to the techniques used for the selective modification of the data. the selective modification will help us to achieve higher utility for modified data[2]. The problem for finding most favourable purification of a database against association rule analysis research can be divided into hiding sensitive rules and sensitive items[3]. For association rules hiding, two basic approaches have been proposed. the first approach hides one rule at a time. First select the transactions that contains the item in a give rule. It then tries to modify transaction by transaction until the confidence or support of the rule fall below

minimum confidence or minimum support[4]. The modification is done by either removing items from the transaction or inserting new items to the transaction. The second approach deals with groups of restricted patterns or association rules at a time[5]. It first selects the transaction that contain the intersecting patterns of a group of restricted patterns. In our work we are concern of hiding certain association rules which contain some sensitive information which are on the Right hand side or left hand side of the rule, so that rules containing confidential item can't be reveal. Our approach is based on modifying the database in a way that confidence of association rule can be reduce with the help increase or decrease the support value of RHS or LHS correspondingly. As the confidence of the rule is reduce below a specified threshold, it is hidden or we can say it will not be disclosed.

2. Literature Survey

The problem of mining association rules was introduced in. The problem of mining association rules is to find all rules that are greater than the user-specified minimum support threshold and minimum confidence threshold. Association rule using support and confidence can be defined as follows. Let I = {i1, i2… in} be a set of literals, called items. Database D= {T1, T2,T3,…,Tn) is a set of transactions, where each transaction T is a set of items such that T$\subset$ I, an association rule is an expression,$X \rightarrow Y$where $X \subset I$, $Y \subset I$ and $X \cap Y = \emptyset$. The X and Y are called correspondingly 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 c is calculated as $|X \cup Y| \div |X| \geq c$. The support s of the rule is the percentages of transactions that contain both X and Y, which is calculated as $|X \cup Y| \div |D| \geq s$. 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. We consider user specified thresholds for support and confidence, MST (minimum support threshold) and MCT (minimum confidence threshold). There are many approaches have been proposed to preserve privacy for crucial knowledge or sensitive association rules in database. They can be classified in to following classes: Heuristic based, these approaches can be further divided in to two groups based on data modification techniques: data distortion techniques and data blocking techniques. Data distortion techniques try to hide association rules by decreasing or increasing support. To increase or decrease support, they replace 0's by 1's or vice versa in selected transactions. So, they can be used to address the complication issue. But they produce undesirable side effects in the new database, which lead them to suboptimal solution. The method of reduce the side effects in sanitized database, which are produced by other approaches. An efficient clustering-based approach to reduce the time complexity of the hiding process. Data blocking

techniques replace the 0 and 1 by unknowns "?" in selected transaction instead of inserting or deleting items. So, it is difficult for an opponent to know the value behind "?"

3. Problem Statement

The goal of data mining is to extract hidden or useful unknown interesting rules or patterns from database. However, the objective of privacy preserving data mining is to hide confidential data 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 deduced 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. Many researchers have worked on the basis of reducing the support and confidence of sensitive association rule. ISL and DSR are common approaches used to hide the sensitive rules. Some of the researchers have used data perturbation techniques to modify the confidential data value in such a way that the approximant data mining results could be obtained from the modified version of the database.

4. Proposed 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 A.

Output:
A transformed database D, where rules containing A on Left
Hand Side (LHS) or Right Hand Side (RHS) will be hidden.

Steps of the algorithm:
1. Generate all Frequent Item Set Fk from A Using Apriori Algorithm;
2. For all (Z ∈Fk , k >= 2) do begin
3. Max_Sup = max({sup(Z') l Z⊂Z ∈Fk+1} ∪{0});
4. if Z. sup ≠Max_Sup then begin
5. A1 = {{Z[1] }, {Z[2] } .....Z[k]]}; //create 1-
Antecedents
6. for (i = I; cx (Ai ≠ Ø) and (i< k); i++) do begin
7. for all X ∈ Ai do begin
    7.1 find Y∈ Fi such that Y = X;

7.2 X.Count = Y. count;
7.3 if (Z. count/XCount > c) then begin
7.4 if (Max_Sup/X.Count < c) then
7.5 print(X ⇒" Z\X with support: ", Z. count, and confidence: " Z.count / X.Count");
7.6 Ai = Ai \ {X}
7.7 End if
7.8 End for
7.9 Ai+1 = Apriori_Gen(Aj)
8. End for
9. End if
10. End for
11. Compute confidence of all the Representative rules.
12. for each hidden item h
13. For each rule containing h, compute confidence of rule R
14. For each rule R in which h is in RHS
14.1.1 If confidence (R) < min conf, then Go to next RR;
14.1.2 Else go to step 6
15. Decrease Support of RHS i.e. item h.
15.1 Find T = t in D | t fully support R;
15.2 While (T is not empty)
15.2.1 Choose the first transaction t from T;
15.2.2 Delete the item set which is in RHS Item of RR;
15.2.3 End While
15.3 Compute confidence of R;
15.4 If T is empty, then h cannot be hidden;
End For
16. For each rule R in which is in LHS
17. Increase Support of LHS;
17.1 Find T = t in D | t does not support R;
17.2 While (T is not empty)
17.2.1 Choose the first transaction t from T;
17.2.2 ADD the item set which is in LHS Item of RR;
17.2.3 End While
17.3 Compute confidence of R;
17.4 If T is empty, then h cannot be hidden;
End For
End Else
End For
18. Output updated D, as the transformed D;

5. Dataset Implementation

TID	LIST OF ITEMS
T1	L1,L2,L5
T2	L2,L4
T3	L2,L3
T4	L1,L2,L4
T5	L1,L3
T6	L2,L3
T7	L1,L3
T8	L1,L2,L3,L5
T9	L1,L2,L3

List of transactions and items used

MINIMUM CONFIDENCE : 70%
MINIMUM SUPPORT : 2/9 = 22%

ITEMSET	SUP COUNT
L1	6
L2	7
L3	6
L4	2
L5	2

C1

ITEMSET	SUP COUNT
L1	6
L2	7
L3	6
L4	2
L5	2

L1

ITEMSET	SUP COUNT
L1,L2	4
L1,L3	4
L1,L4	1
L1,L5	2
L2,L3	4
L2,L4	2
L2,L5	2
L3,L4	0
L3,L5	1
L4,L5	0

C2

ITEMSET	SUP COUNT
L1,L2	4
L1,L3	4
L1,L5	2
L2,L3	4
L2,L4	2
L2,L5	2

L2

ITEMSET	SUP COUNT
L1,L2,L3	2
L1,L2,L5	2

C3

ITEMSET	SUP COUNT
L1,L2,L3	2
L1,L2,L5	2

L3

Minimum Confidence = 70%

Association rules

- L1.L2 -> L5
  - Confidence = sc{l1,l2,l5}/sc{l1,l2} = 2/4 = 50%
  - R1 is rejected
- L1.L5 -> L2
  - Confidence = sc{l1,l2,l5}/sc{l1,l5} = 2/2 = 100%
  - R2 is selected
- L2.L5 -> L1
  - Confidence = sc{l1,l2,l5}/sc{l2,l5} =4/4 = 100%
  - R3 is selected
- L1 -> L2.L5
  - Confidence = sc{l1,l2,l5}/sc{l1} = 2/6 = 33%
  - R4 is rejected
- L2 ->L1. L5
  - Confidence = sc{l1,l2,l5}/sc{l2} = 2/7 = 29%
  - R5 is rejected
- L5 -> L1.L2
  - Confidence = sc{l1,l2,l5}/sc{l5} = 2/2 = 100%
  - R6 is selected

FINAL RULES :

1. L1.L5 -> L2
2. L2.L5 -> L1
3. L5 ->L1.L2

Let sensitive item set H= {L1}. Now choose the representative rules containing the 'L1' in RHS. From the set of RR, one can find there are
L2.L5 -> L1
L5 ->L1.L2

Now delete the sensitive item ‘L1’ from all the transactions and the transactional data set will be modified.

TID	LIST OF ITEMS
T1	L2, L5
T2	L2, L4
T3	L2,L3
T4	L2, L4
T5	L3
T6	L2,L3
T7	L3
T8	L2,L3,L5
T9	L2,L3

Now choose the representative rules containing the ‘L1’ in LHS. From the set of RR, one can find there are:
L1.L5 -> L2
Now delete the sensitive item ‘L2’ from all the transactions which follow the above rule and the transactional data set will be modified

TID	LIST OF ITEMS
T1	L5
T2	L4
T3	L3
T4	L4
T5	L3
T6	L3
T7	L3
T8	L3,L5
T9	L3

6. Results

TID	LIST OF ITEMS
T1	L1,L2,L5
T2	L2,L4
T3	L2,L3
T4	L1,L2,L4
T5	L1,L3
T6	L2,L3
T7	L1,L3
T8	L1,L2,L3,L5
T9	L1,L2,L3

Original Dataset

TID	LIST OF ITEMS
T1	L5
T2	L4
T3	L3
T4	L4
T5	L3
T6	L3
T7	L3
T8	L3,L5
T9	L3

Modified Dataset

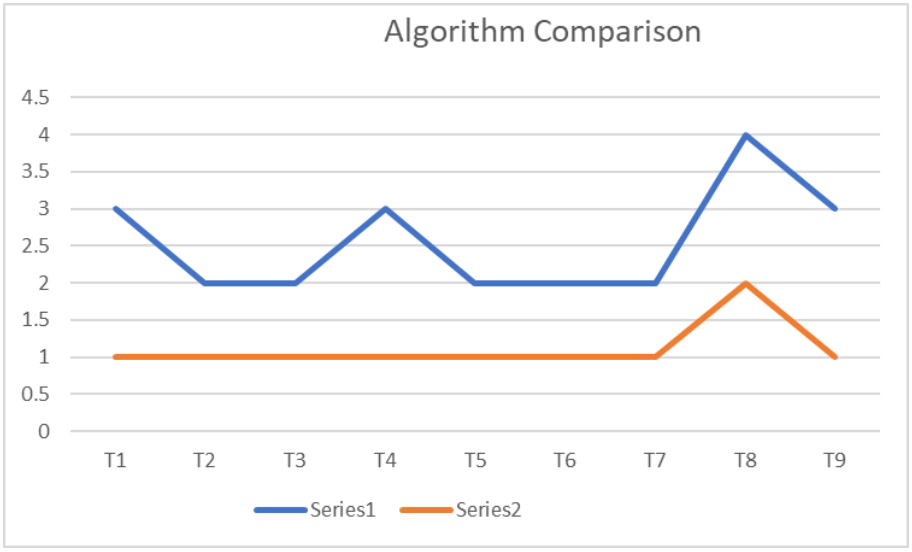

Fig: Comparison of the two dataset with the same utility

Since the same utility is being provided even after abstraction of sensitive items, our hybridized algorithm is successful in providing an adequate amount of data privacy with less loss of utility

7. Conclusion

The database privacy problems in data mining have been discussed and an algorithm for hiding sensitive data in association rules mining proposed. The proposed algorithm is hybrid of two existing algorithm ISL and DSR . Our algorithm prunes more number of hidden rules with same number of transactions scanned, less CPU time. In this approach there are no side effects of any other rule since just the addition of hiding of elements in total transaction have been done. Modifications have been made only according the minimum support and confidence in order to hide an element of a transaction.

8. References

[1] Shweta Taneja, Shashank Khanna, Sugandha Tilwalia, Ankita "A Hybrid C-Tree Algorithm for Privacy Preserving Data Mining " in proceedings of International Journal of Soft Computing and Engineering (IJSCE) , ISSN: 2231-2307, Volume4, Issue-ICCIN-2K14, March 2014

[2] Sridhar Mandapati, Dr Raveendra Babu Bhogapathi and Dr M.V.P.Chandra Sekhara Rao "Swarm Optimization Algorithm for Privacy Preserving in Data Mining" in proceedings of IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 2, No 3, March 2013

[3] Shweta Taneja, Shashank Khanna, Sugandha Tilwalia, Ankita "A Review on Privacy Preserving Data Mining :Techniques and Research Challenges " in proceedings of International Journal of Computer Science and Information Technologies, Vol. 5 (2) , 2014

[4] Imran Khan, Virendra Kumar, Savita Shiwani "An Efficient Technique Privacy Preserving Association Rule Data Mining using Modified Hybrid Algorithm " in proceedings of International Journal of Science, Engineering and Technology, Volume 02, Issue 06, July 2014

[5] Praveena Priyadarsini, M.L.Valarmathi, S.Sivakumari "Hybrid Perturbation Technique using Feature Selection Method for Privacy Preservation in Data Mining " in proceedings of International Journal of Computer Applications, Volume 58– No.2, November 2012

[6] Pingshui WANG, "Survey on Privacy Preserving Data Mining" in proceedings of International Journal of Digital Content Technology and its Applications Volume 4, Number 9, December 2010

[7] Kshitij Pathak, Narendra S Chaudhari, Aruna Tiwari " Privacy Preserving Association Rule Mining by Introducing Concept of Impact Factor" in proceedings of IEEE, 2011

[8] Praveena Priyadarsini, M.L.Valarmathi, S.Sivakumari "Hybrid Perturbation Technique using Feature Selection Method for Privacy Preservation in Data Mining " in proceedings of International Journal of Computer Applications, Volume 58– No.2, November 2012

[9] Dharmendra Thakur ,Prof. Hitesh Gupta "An Exemplary Study of Privacy Preserving Association Rule Mining Techniques " in proceedings of International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 11, November 2013

[10] Arvind Batham, Mr.Srikant Lade ,Mr. Deepak Patel "A Robust Data Preserving Technique by KAnonymity and hiding Association Rules " in proceedings of International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 1, January 2014

[11] AGGARWAL, C. C. AND YU, P. S. 2008b. On static and dynamic methods for condensationbased privacy preserving data mining. ACM Trans. Data b. Syst. 33, 1.

[12] AGGARWAL, C. C. AND YU, P. S. 2008c. Privacy-Preserving Data Mining: Models and Algorithms. Springer, Berlin.

[13] AGGARWAL, C. C. AND YU, P. S. 2007. On privacy-preservation of text and sparse binary data with sketches. In Proceedings of the SIAM International Conference on Data Mining (SDM).

[14] Amruta Mhatre, Durga Toshniwal, “Hiding Co-occurring Sensitive Patterns in Progressive Databases”, ACM, March 22,2010.

[15] Shikha Sharma & Pooja Jain, “A Novel Data Mining Approach for Information Hiding”, International Journal of Computers and Distributed Systems, Vol. No.1, Issue 3, October 2012.

[16] L.Sweeney, “Achieving k-Anonymity Privacy Protection Using Generalization and Suppression”, International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, vol.10, no.5

[17] Razali, A.M. and S. Ali, “Generating treatment plan in medicine: A data mining approach”. Am. J. Applied Sci., 6: pp. 345-351, 2009.

[18] Evfimievski, A., R. Srikant, R. Agrawal and J. Gehrke. “Privacy preserving mining of association rules”. Proceedings of the 8th ACM

[19] Saygin, Y., V.S. Verykios and A.K. Elmagarmid. “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, 2002.

[20] Vaidya, J., H. Yu and X. Jiang. “Privacy preserving SVM classification”. Knowl. Inform. Syst., pp. 161-178, 2008.