NATIONAL CONFERENCE ON RECENT RESEARCH IN ENGINEERING AND TECHNOLOGY (NCRRET-2015)
INTERNATIONAL JOURNAL OF ADVANCE ENGINEERING AND RESEARCH DEVELOPMENT (IJAERD)
E-ISSN: 2348 - 4470 , PRINT-ISSN:2348-6406

A Survey of Association Rule Hiding Approaches For Privacy Preservation Data Mining

Divya C. kalariya,
Research Scholar
G.H. Patel College of Engineering and Technology.
Vallabh Vidyanagar, India
divya2patel@gmail.com

Vinita Shah,
Asst. Prof., IT Dept.
G.H. Patel College of Engineering and Technology.
Vallabh Vidyanagar, India
vinitashah@gcet.ac.in

Jay Vala
Asst. Prof., IT Dept.
G.H. Patel College of Engineering and Technology.
Vallabh Vidyanagar, India
jayvala@gcet.ac.in

Abstract—Data mining algorithm is used to extract the useful knowledge or information from database. Privacy preservation data mining is novel research area where data mining algorithms are analyzed for their side-effects they done on data privacy. Privacy preservation data mining (PPDM) deals with the problem of hiding the sensitive information while analyzing data. Many techniques are available for PPDM like data distortion, data hiding, rule hiding, data modification etc. Association rule hiding is one of the technique of PPDM. It hides sensitive rules which are generated by association rule generation algorithm before releasing database. This paper discusses different approaches of association rule hiding technique.

Index Terms—Association Rule Mining, Association rule hiding, Confidence, Data mining, Privacy Preservation Data Mining (PPDM), Support.

1 Introduction

Data mining aims to extract hidden information from data warehouses. In data mining, different type of algorithms are used to extract different useful information from large amount of data. Algorithms are analyzed for their side effect which incur the data privacy. For example, using data mining algorithm on database, anyone can extract sensitive information like frequent pattern, association rules, unclassified data etc. It means data mining poses a threat to information privacy. To solve that problem, privacy preservation data mining concept is used in data mining and database security field.

Different techniques are used to solve PPDM. Association rule hiding is one technique of PPDM to hide sensitive rules which is generated by rule generation algorithm. Association rule generation algorithm is based on frequent items occurring in database. Frequent items mean the set of item which occurring together in a transaction. Finding that frequent items using different algorithm like apriori, FP growth tree. Generated rules are input in rule hiding algorithm, when applying rule hiding. Result of rule hiding algorithm is sanitized database which is not containing sensitive rules.

To understand the requirement of association rule hiding in PPDM, Here take one example which includes one cancer researcher and two hospitals. When researcher want to do survey on database of cancer hospital so researcher requests for the dataset of two hospitals i.e. A and B for review purpose. Both hospital A and B wants to hide the treatment related information based on symptoms from each other and also from researcher. So before giving database to researcher, both hospital use rule hiding technique to hide sensitive rules i.e. frequent symptoms→treatment 1. Output of association rule hiding algorithm is sanitized database which never generates the sensitive rule define by hospital. And this sanitized database is given to researcher for review approach.

Problem statement is defined in next section. Association rule mining concepts are described in section 3. Association rule hiding approaches & different algorithms are discussed in section 4. Section 5 describes process flow of association rule hiding. Output performance parameters are explained in section 6. Section 7 presents the conclusion.

2 Problem Definition

Dataset D is our input then AR is Association rule which is generated from input database D. If user want to hide some sensitive association rule (SR) selected by user then SR can be hidden by applying different rule hiding approaches which are discussed in section 4. Using approaches sanitized database D′ can be generated. D′ contains only the rules which are not present in SR (AR - SR). Rule hiding approach should try to maintain data quality of D′ so dissimilarity between D′ and D (D - D′) should be as possible as lesser.

3 Association Rule Mining

Association rule mining firstly proposed by Agrawal et al in 1993[1]. An association rule is an implication expression of the form X→ Y, where X and Y are disjoint item sets, i.e., $X \cap Y = \emptyset$.[1] Support and confidence are two basic parameter of association rule mining. Definition of support and confidence is defined below [2]:

Support is percentage of transactions in dataset that contain XUY.

$$Support(XY) = \frac{Total\ no\ of(XY)}{Total\ no\ of\ transaction\ in\ D} \quad (1)$$

Confidence is the percentage of transactions in dataset containing X that also contain Y. Confidence show the conditional probability.

$$Confidance(XY) = \frac{support\ (XY)}{\text{support X}} \quad (2)$$

Based on Minimum Support Threshold (MST) and Minimum Confidence Threshold (MCT) value, frequent item set and association rules are generated using different algorithm like apriori, FP growth.

If user want to hide the rules then he should try to decrease the confidence value of that rule compare to MCT. User can do this by decreasing the value of confidence by increasing the value of denominator or by decreasing the value of numerator. And the value of denominator and numerator can be changed by altering the value of support count of Item sets. Altering the values of support count are based on different approach which are explained in next section.

Result of association rule hiding algorithm is based on some parameter like accuracy, completeness, consistency of sanitized database.

4 ASSOCIATION RULE HIDING APPROCHES & DIFFERENT ALGORITHMS

The concept of privacy preservation data mining has been recently proposed in response to the concerns of preserving privacy information from data mining algorithms. [3] Basically there are two type of privacy related to data mining which are output privacy and input privacy. Output privacy, means the data is minimally altered so that the mining result will not disclose certain privacy. [4] For output privacy, many technique are developed i.e. heuristic approach based technique like perturbation, blocking, swapping etc. Input privacy, is that the data is manipulated so that the mining result is not affected or minimally affected. [4] For input privacy, many techniques are developed i.e. cryptographic approach based technique like secure multiparty computation etc.

There are main five types of different approaches of association rule hiding which are related to input/output privacies are discussed below.

4.1 Heuristic Based Approach

This approach involves efficient, fast and scalable algorithms that selectively sanitize a set of transactions from the original database to hide the sensitive association rules [5].It is divided further in to two types that are Distortion techniques and blocking technique.

Distortion technique delete items by replacing 1-values to 0-values for reducing the confidence of rules or this technique add items by replacing 0-values to 1- values for reducing the support of rules. Sensitive rules are being hidden based on modification in database due to deleting or adding the items. Different algorithm are available for this approach. In [6], authors have presented three algorithms 1.a, 1.b and 2.a for hiding sensitive association rules. Algorithm 1.a inserts the items in transaction therefore increases the support value of L.H.S. side items in rules so confidence of that rule will be decreased automatically. Side effect of insertion new items in database is generation of new association rules. Algorithm 1.b and 2.a deletes the R.H.S. items of rules so confidence will decreased. Sometimes algorithm 1.b & 2.a affect the non-sensitive rules also. Two algorithms Increase Support of L.H.S (ISL) and Decrease Support of R.H.S (DSR) are proposed in [5]. In [7] Algorithm DCIS (Decrease Confidence by Increase Support) and DCDS (Decrease Confidence by Decrease Support) are proposed. ISL and DCIS based on item adding approach while DSR and DCDS is based on Item deleting approach. DSRRC (Decrease Support of R.H.S. item of Rule Clusters) is given in [8], which provides privacy for sensitive rules at certain level while ensuring data quality. Proposed DSRRC algorithm clusters the sensitive association rules based on R.H.S. of rules and hides as many as possible rules at a time by modifying fewer transactions. Algorithm DSRRC cannot hide rules having multiple RHS items. To solve the disadvantages of DSRRC algorithm, MDSRRC algorithm are proposed in [9]. Table 1 describe the list of different algorithms of distortion technique..

TABLE1

LIST OF ALGORITHM

Approach	Algorithm	Conclusion
Insertion Based Algorithm(L.H.S.)	Algorithm 1.a	Large number of new rule generation and less number of rules are lost.
	ISL	
	DCIS	
Deletion Based Algorithms(R.H.S.)	Algorithm 2.a	Large number of rules are lost and less number of new rule generation.
	Algorithm 2.b	
	DSR	
	DCDS	
	DSRRC	
	MDSRRC	

Blocking technique replaces an existing value to "?". This technique inserts unknown values in the data to fuzzify the rules. Sometimes providing wrong information to other is not acceptable. Adversary can easily find out the unknown value in sanitized dataset.

4.2 Border Based Approach

It hides sensitive association rule by modifying the borders in the lattice of the frequent and the infrequent item sets of the original database. The item sets which are at the position of the borderline separating the frequent and infrequent item sets forms the borders. Border based approach is unable to identify optimal hiding solution but still dependent on heuristic to decide upon the item modification [10].

4.3 Exact Approach

It formulates the hiding process as constraints satisfaction problem or an optimization problem which is solved by integer programming. It tries to minimize the distance between the original database and its sanitized version. It takes high times complexity because of binary integer programing. Authors of [12] introduced the exact methodology to perform sensitive frequent item set hiding based on the notion of a hybrid database generation.

4.4 Reconstruction Based Approach

In this approach, inverse frequent item set mining algorithm is used. Step of reconstruction based approach is discussed below. First select the database D as input and generate the frequent item set (FS). Now convert FS to sanitized FS; FS′ doesn't generate the sensitive rules. Now apply inverse frequent dataset mining algorithm are used to convert FS′ to sanitized database D′. The open problem of this approach is to restrict the number of transactions in the new database.

4.5 Cryptography Based Approach

It is used for multiparty computation, when database is distributed among several sites. Multiple parties may wish to share their private data, without leaking any sensitive information at their end. This approach is divided in two types: vertically partitioned distributed data and horizontally partitioned distributed data. Authors in [13], proposed a secure approach for sharing association rules when data are vertically partitioned. In terms of communication cost this approach is very effective, but it is very expensive for large amount of datasets. The authors in [14] addressed the secure mining of association rules over horizontal partitioned data.

5 PROCESS FLOW OF ASSOCIATION RULE HIDING

In this section, steps of association rules hiding process are discussed. Figure 1 represents the flow of process.

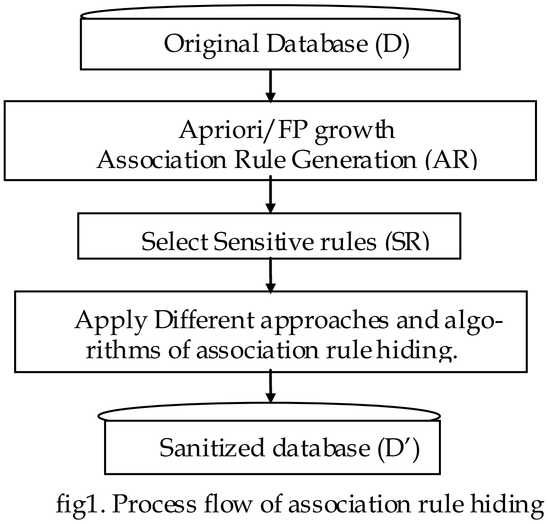
Original Database (D)
Apriori/FP growth
Association Rule Generation (AR)
Select Sensitive rules (SR)
Apply Different approaches and algorithms of association rule hiding.
Sanitized database (D′)

fig1. Process flow of association rule hiding

Database (D) is taken as input. Then apply apriori/FP growth algorithm on D. It generates frequent items and using it, association rule generation algorithm generates rules (AR).User select the sensitive association rules (SR) from AR. Then apply association rule hiding algorithm and generates sanitized database (D′).when applying rule generation algorithm on D′ then D′ never generates the SR. Output is (AR-SR).

6 PERFORMANCE PARAMETERS

Evolution parameter are used to evaluating the performance of association rule hiding algorithms. Detailed description of parameter are discussed below. [15]

6.1 Efficiency

Space requirements, CPU-time and communication required for hiding are used to measured efficiency. Good performance in terms of resources allocated.

6.2 Scalability

It is measured in terms of good performance for varying sizes of input datasets. Hiding failure: It is the percentage of the portion of information that fails to be hidden. It is derived by, $HF = |Rs(D')| / |Rs(D)|$ where, $|Rs(D')|$ are the number of sensitive rules appearing in the sanitized database and $|Rs(D)|$ are the number of sensitive rules in the original database.

6.3 Data quality

Data quality parameters are accuracy measure, completeness, consistency which is in relationship to preservation of original data values and of data mining results.

6.4 Ghost Rules

It shows percentages of rules that are not present in the original database but can be derived from sanitized database.

6.5 Privacy level

It measures the degree of uncertainty according to which the protected information can still be predicted.

6.6 Dissimilarity

It shows difference between original database and sanitized database.

6.7 Lost Rules cost

It measures the number of no sensitive association rules found in the original database but not in sanitized database.

7 CONCLUSION

In this survey we have discussed the requirement of privacy preservation in data mining. We have also described the steps of association rule hiding techniques in PPDM. We briefly discussed the various approaches and algorithms of association rule hiding. We conclude that performance parameters described in section 6 are used to analyze the different algorithms and approaches. Based on survey papers, it has been found that Heuristic Approach is the most widely used approach for rule hiding purpose. We may develop hybrid approach by combining two or more approaches for accurate result.

REFERENCES

[1] Agrawal, R., Imielinski, T., and Swami, A. N.1993. Mining association rules between sets of items in large databases. In Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data , P. Buneman and S. Jajodia, Eds. Washington, D.C., 207 216

[2] J. Han and M. Kamber, Data Mining: Concepts and Techniuqes. Morgan Kaufmann Publishers, San Francisco, CA, 2001, pp. 227–245.

[3] Agrawal et al., 2000; Brankovic & Estivill-Castro, 1999; Clifton &Marks, 1996; Lindell &Pinkas, 2000; O' Leary, 1991;Verykios et al., 2004

[4] Shyue-Liang Wang , Bhavesh Parikh, Ayat Jafari "Hiding informative association rule sets" Expert Systems with Applications 33, ELSEVIER, 2007, pp. 316-323.

[5] Aris Gkoulalas–Divanis;Vassilios S. Verykios "Association Rule Hiding For Data Mining" Springer, DOI 10.1007/978-1-4419-6569-1, Springer Science + Business Media, LLC 2010

[6] Vassilios S. Verykios, A.K. Elmagarmid, E. Bertino, Y. Saygin, and E. Dasseni, "Association Rule Hiding," IEEE Transactions on Knowledge and Data Engineering, vol. 16, no. 4, pp. 434-447, 2004.

[7] Shyue-LiangWang ;Dipen Patel ;Ayat Jafari ;Tzung-Pei Hong, "Hiding collaborative recommendation association rules", Published online: 30 January 2007, Springer Science+Business Media, LLC 2007

[8] Chirag N. Modi, Udai Pratap Rao,Dhiren R. Patel "Maintaining Privacy and Data Quality in Privacy Preserving Association Rule Mining" Second International conference on Computing, Communication and Networking Technologies, IEEE, 2010, pp.1-6.

[9] Nikunj H. Domadiya,Udai Pratap Rao "Hiding Sensitive Association Rules to Maintain Privacy and Data Quality in Database" International Advance Computing Conference (IACC), IEEE, 2013, pp.1306-1310.

[10] Vikram Garg, Anju Singh & Divakar Singh "A Survey of Association Rule Hiding Algorithms" Fourth International Conference on Communication Systems and Network Technologies, IEEE, 2014, pp. 404-407.

[11] X. Sun, and P. Yu, "A Border-Based Approach for Hiding Sensitive Frequent Itemsets," In: Proc. Fifth IEEE Int'l. Conf. Data Mining (ICDM 2005), pp. 426–433, 2005.

[12] A. Gkoulalas-Divanis, and V. S. Verykios, "Exact knowledge hiding through database extension" IEEE Trans Knowledge Data Eng 2009, pp. 699–713.

[13] J. Vaidya, and C. Clifton, "Privacy preserving association rule mining in vertically partitioned data," In proc. Int'l Conf. Knowledge Discovery and Data Mining, pp. 639–644, July 2002.

[14] M. Kantarcioglu, and C. Clifton, "Privacy-preserving distributed mining of association rules on horizontally partitioned data," IEEE Transactions on Knowledge and Data Engineering, Vol. 16, No. 9, pp. 1026-1037, Sept. 2004.

[15] Khyati B. Jadav, Jignesh Vania & Dhiren R. Patel "A Survey on Association Rule Hiding Methods" International Journal of Computer Applications (0975 – 8887) Volume 82 – No 13, November 2013, pp. 20-25