A Hybrid Technique for Hiding Sensitive Association Rules and Maintaining Database Quality

Nikunj H. Domadiya and Udai Pratap Rao

Abstract In this digital world, data mining is a decisive for innovation and better services for user, but it raises the issues about individual privacy. Privacy can be achieved by hiding sensitive or private information in database before publishing it for innovation. This paper presents a hybrid technique for hiding sensitive association rules, which combines two heuristic based techniques viz. Decrease Support and Decrease Confidence of sensitive rule for selection and modification of items from the transactions. The proposed hybrid technique combines advantages of both algorithms to maintain the quality of the database and preserve the privacy of database. From the experimental results it is observed that proposed algorithm is competent to maintain privacy and database quality.

Keywords Data mining · Association rule · Sensitive pattern · Privacy preserving data mining and association rule hiding

1 Introduction

Association rule mining is very useful technique to find the hidden relationship among data in large database. It can be used for many applications as healthcare, business policy and marketing, e-commerce etc. As a part of business improvement or decision, many organizations share their database for mutual benefits. In many case, due to privacy law or some policy, organization don't want to disclose the private information stored in database. It raises the problem of securing this private data from the adversary to maintain the privacy. The privacy is major constrain during the mining of database which contains private information. Here privacy is

N.H. Domadiya (✉) · U.P. Rao
Sardar Vallabhbhai National Institute of Technology, Surat 395007,
Gujarat, India
e-mail: domadiyanikunj002@gmail.com

U.P. Rao
e-mail: udaiprataprao@gmail.com

S.C. Satapathy and S. Das (eds.), *Proceedings of First International Conference on Information and Communication Technology for Intelligent Systems: Volume 2*,
Smart Innovation, Systems and Technologies 51, DOI 10.1007/978-3-319-30927-9_35

to hide private information which organization don't want to disclose with other, to solve this issue PPDM technique is very useful to enhance the security of database. In a centralize database, PPDM techniques are used to transform the database such that private information or sensitive pattern cannot be mined from the data mining techniques and maintain the quality of final result. In 1996, Clifton et al. [1, 2] had discussed the privacy in data mining. They also discussed the hiding of association rule to maintain the privacy. In 1999 Atallah et al. [3] proposed heuristic approach to prevent disclosure of sensitive patterns.

1.1 Problem Description

The problem of association rule hiding can be describe as: Transform the original database such that data mining techniques will results only non sensitive rules and all sensitive rules must not mined from transformed database. This transformed database is known as sanitized database.

In general the problem can be defined as:

Given transnational database D, Minimum support threshold, Minimum confidence threshold. Association rules R can be generated as result of data mining technique from D. Sensitive rules (SR, SR $\subset$ R) selected from given set of rules R by database owner. The problem is to transform database D into D′ in such way that, data mining results from D′ will hide all sensitive rules.

The objective of proposed algorithm is to achieve the following conditions

1. Transformed database D′ must hides all sensitive rules.
2. Mining of transformed database D′ must results in all non sensitive rules.
3. Transformed database must not introduce any artificial rules, which are not present in D.

The problem of finding an optimized sanitized database, which satisfies all these conditions has been proved as NP-hard in [3].

The structure of remaining paper is as follows: Sect. 2 presents theoretical background and related work. In Sect. 3, proposed hybrid technique for hiding sensitive association rule is discussed. In Sect. 4, we analyze the algorithm with some evaluation parameter of database quality. Finally, last Sect. 5 concludes our work and gives the future direction.

2 Theoretical Background and Related Work

2.1 Association Rule Mining

Association rule mining [4] with given minimum support threshold (MST) and minimum confidence threshold (MCT) is defined as follows: let $\mathrm{I} = \{\mathrm{i}_1, \ldots \mathrm{i}_N\}$ be distinct literals called items. Given a database D = {$T_1 \ldots T_m$} is a set of transaction where each transaction T is a set of items as $T_i \subset I$ (1 = i = m).

A rule X → Y is mined from database if *support*($X \rightarrow Y$) ≥ MST and *confidence* ($X \rightarrow Y$) ≥ MCT.

Researchers have proposed different approaches for hiding the sensitive association rule to preserve the privacy of sensitive information and these approaches can be classified in Heuristic based approach, Reconstruction based approach and Cryptography based approach. Here we propose a hybrid algorithm for sensitive rule hiding, which is based on heuristic approach.

2.2 Heuristic Based Approach for Sensitive Rule Hiding

The heuristic based approach is used to hide association rules as many as possible by modifying the transactions, while minimizing side effects which can be generated by hiding process. Following side effects can be generated by hiding process.

1. Some sensitive rules which can be mined from sanitized database, called *hiding failure* (HF) effect.
2. Some non-sensitive rules are hidden accidently in sanitized database, called lost rule or missing cost (MC) effect.
3. Some rules are newly created, called ghost rule or artificial rule (AP) effect.

In this type of approach, no algorithm can satisfy all condition of association rule hiding. It produces some side effect as described above. In [3], author proved that finding an optimal solution of association rule hiding algorithm is NP-hard. Algorithms included in this approach, which hide sensitive knowledge by sanitizing selected transactions from the database to decrease the support or confidence of sensitive rule and tries to minimize the side effect. These approaches use either *Data Perturbation* which permanently remove some items from selected transaction of database or *Data Blocking* which replace some items by unknown value (ex. ?).

Next, we see some existing efforts to solve association rule hiding based on data perturbation approach.

Oliveira et al. [5] proposed a graph base sanitization approach to hide sensitive rules. The author also presents the forward inference and backward inference attack on sanitize database. They maintain 2-itemset paring set which is all possible subsets of sensitive frequent itemset. Then algorithm deletes one by one 2-itemset to hide sensitive frequent items set. This algorithm soles both above attack and also minimizes the side effect factors.

Wu and Wang [6] proposed algorithm that compare all three techniques for hiding sensitive rule in terms of number of transaction modification. The algorithm selects the one with the lowest number of modifications of transaction in database to maintain database quality.

Oliveira et al. [7] proposed different flexible algorithm which considers the disclosure threshold for each sensitive rules. Secondly they focus on discovery of maximum non-sensitive rules after database sanitization. This algorithm is not on memory based so it is more suitable for large database size.

Domadiya and Rao [8] modifies the algorithm of Modi et al. [9] to reduce the side effect on sanitized database. They select the items from the R.H.S parts of the sensitive rules based on their frequency. This algorithm modify the original database in such way that total number of modification are reduced and hide all the sensitive rules to maintain the privacy.

In context of heuristic based approach, there is another area called *data blocking approach*, where researchers have worked to address NP-hardness of PPDM.

All the above techniques comes under wither D_SUPP or D_CONF1. we can calculate total number of modification require in database to hide any sensitive rule using following equation.

$$N_{supp} = suppcount(Rule) - Minsuppcount + 1$$
$$N_{conf} = suppcount(Rule) - \lceil Min\_Conf * Supp\_count(L.H.S.\ of\ Rule) \rceil + 1$$

From the analysis, we had observed that, Some sensitive association rules had $N_{supp} < N_{conf}$ and other had $N_{supp} > N_{conf}$. We can find the better technique for hiding any sensitive rule based on the relation of $N_{supp}\ and\ N_{conf}$. For any sensitive rule if

- $N_{supp} < N_{conf}$ then D_SUPP technique hide sensitive rules with less number of modification.
- $N_{supp} > N_{conf}$ then D_CONF1 technique hides sensitive association rule with less number of modification.

In all existing research, only single D_SUPP or D_CONF1 technique is used for hiding all sensitive rules.

In this paper, we propose a hybrid technique to combine the advantage of both D_CONF1 and D_SUPP techniques. We do not include D_CONF2 technique to reduce the artificial pattern (AP) as it insert some new items in database. So we have not used this technique. In next section we describe the proposed hybrid technique for hiding sensitive rule.

3 Proposed Hybrid Technique for Hiding Sensitive Rules

The framework of proposed hybrid technique is shown in Fig. 1. It starts with applying the apriori algorithm [4] to mine the association rules from original database D. Association rules which disclose some private information are

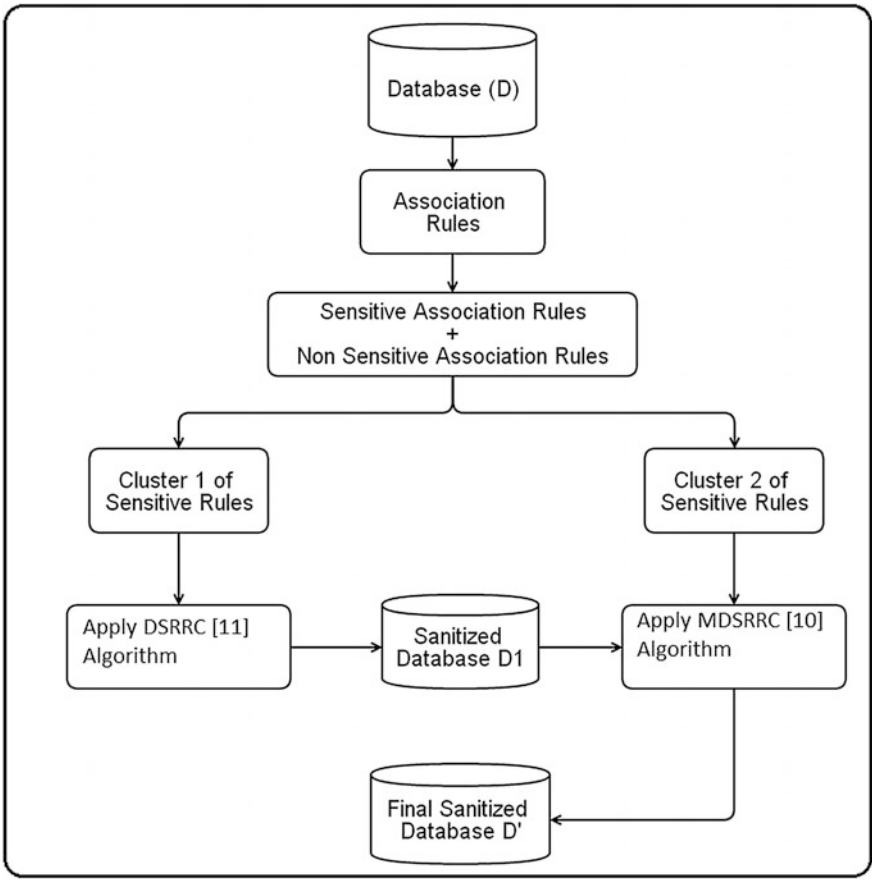
Database (D)
Association Rules
Sensitive Association Rules
+
Non Sensitive Association Rules
Cluster 1 of Sensitive Rules
Cluster 2 of Sensitive Rules
Apply DSRRC [11] Algorithm
Sanitized Database D1
Apply MDSRRC [10] Algorithm
Final Sanitized Database D'

Fig. 1 A framework of proposed hybrid technique for sensitive rule

consider as sensitive rules. Then algorithm divide the sensitive rules based on N_{supp} = number of modification require to reduce the support below min_supp and N_{conf} = reduce the confidence below min_conf. Now we have two clusters of sensitive rules. Cluster 1 with all sensitive rule having $N_{supp} \leq N_{conf}$ and cluster 2 having $N_{supp} > N_{conf}$. Now for hiding sensitive rules in cluster 1, apply DSRRC [9] algorithm based on decreasing support below min_supp. For cluster 2, apply our MDSRRC [8] algorithm based on decreasing confidence below min_conf. It generates the final sanitized database. Sanitized database maintain privacy by hiding all sensitive rules [10].

Proposed Hybrid Technique for Hiding Sensitive Rules

```
1.  Apply association rule mining algorithm (apriori [5]) on database D. It will results in set of association rules R.
2.  Select set of private rules $SR \subset R$ as set of sensitive rules. (Selected by Database Owner)
3.  While (SR not empty)
4.  {
5.    Remove one sensitive rule from SR and Calculate $N_{supp}$ and $N_{conf}$
6.    If ($N_{supp} \leq N_{conf}$)
        Add this sensitive rule in Cluster 1.
      Else
        Add this sensitive rule in Cluster 2.
7.  }
8.  For cluster 1 apply DSRRC [6] algorithm on original database D and generate sanitized database D1.
9.  For cluster 2 apply our MDSRRC [4] algorithm on sanitized database D1 and generate final sanitized database D'.
10. Update all modified transaction in original database and generate final sanitized database D'
```

The proposed hybrid algorithm as shown above starts with mining the association rules (R) using Apriori algorithm [4] in database D. Then user specifies some private rules as set of sensitive rules (SR). Then algorithm generates two clusters of sensitive rules from SR by comparing N_{conf} and N_{supp}. Then apply DSRRC [9] algorithm for cluster 1 and proposed MDSRRC algorithm for cluster 2 to hide sensitive rules SR. At last it updates the modified transaction to original database and generates final sanitized database D′. Sanitized database maintains privacy by hiding the sensitive rules and maintain database quality by minimum number of modification in database using hybrid technique.

4 Experimental Results and Analysis of Proposed Algorithm

For performance analysis of proposed algorithm, we have compared proposed algorithm with MDSRRC [8] algorithm using a retail database with total 88,162 transactions given in [11]. We have applied apriori algorithm [4] with MST = 5 % or MST (in count) = 4408 and MCT = 10 % to generate all possible association rules. In our experiment, we choose three rules ($\{32, 39\} \rightarrow \{48\}$, $\{32, 48\} \rightarrow \{39\}$, $\{39\} \rightarrow \{38, 48\}$) as sensitive rules from all possible rules. All sensitive rules with its MST, MCT and N_{supp} and N_{conf} values are shown in Table 1. As shown in Table 1, first two sensitive rules have $N_{supp} < N_{conf}$ and for last one $N_{supp} > N_{conf}$. After applying hybrid algorithm on retail database given in [11], we have evaluated our proposed Hybrid algorithm by considering the following evaluation parameter discussed in [12]. (a) HF (hiding failure), (b) MC (misses cost), (c) AP (artificial patterns), (d) DISS (dissimilarity) and (e) SEF (side effect factor). Experimental results show that proposed hybrid algorithm works better compared to existing MDSRRC algorithm [8] based on any single basic techniques (D_SUPP, D_CONF1,

Table 1 Sensitive rules with MST, MCT, N_{supp}, N_{conf}

Rule no.	Sensitive rules	MST (in count)	MCT (%)	N_{supp}	N_{conf}
1	{32, 39} → {48}	5402	63.89	995	4557
2	{32, 48} → {39}	5402	67.24	995	4599
3	{39} → {38, 48}	6102	12.04	700	140

Table 2 Performance results of MDSRRC [8] and proposed algorithm

Evaluation parameter	MDSRRC [8] (%)	Proposed hybrid algorithm (%)
HF (hiding failure)	0	0
MC (missing cost)	26.66	13.33
AP (artificial pattern)	0	0
DISS (D, D′)	2.77	0.78
SEF (side effect factor)	26.66	13.33

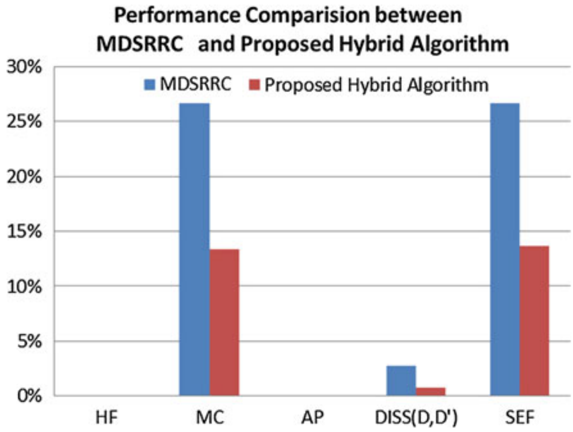

Fig. 2 Performance results of MDSRRC [8] and proposed algorithm

D_CONF2). It hides all sensitive rules (HF = 0 %) without generating any artificial rules (AP = 0 %) and maintain database quality by minimizing the modification on database. Performance results in terms of evaluation parameters are shown in Table 2 (Fig. 2).

5 Conclusion and Future Scope

This paper presented a hybrid technique to hide the sensitive rule by combining the advantage of both D_SUPP and D_CONF1 approaches. The proposed algorithm tries to remove the limitation of these both algorithms by comparing the number of modifications N_{supp} and N_{conf} to minimize the modification on the database.

We have demonstrated our proposed algorithm using a sample database. The proposed hybrid technique modifies a minimum number of transactions to maintain database quality. An artificial pattern (AP) gives the wrong direction to analysts, so the proposed algorithm does not insert any new items in transactions and maintains an artificial pattern (AP) 0 %. In future, the proposed hybrid technique can be improved in terms of missing cost (MC) and side effect factor (SEF).

References

1. Clifton, C., Kantarcioglu, M., Vaidya, J.: Defining privacy for data mining. In: National Science Foundation Workshop on Next Generation Data Mining, vol. 1, p. 1. Baltimore, MD (2002)
2. Clifton, C., Marks, D.: Security and privacy implications of data mining. In: ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery, pp. 15–19 (1996)
3. Atallah, M., Elmagarmid, A., Ibrahim, M., Bertino, E., Verykios, V.: Disclosure limitation of sensitive rules. In: Proceedings of the 1999 Workshop on Knowledge and Data Engineering Exchange, KDEX'99, pp. 45–52, Washington, DC, USA. IEEE Computer Society (1999)
4. Han, J.: Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers Inc., San Francisco (2005)
5. Oliveira, S.R.M., Zaane, O.R., Saygin, Y.: Secure association rule sharing. In: Dai, H., Srikant, R., Zhang, C. (eds.) Advances in Knowledge Discovery and Data Mining, 8th Pacific-Asia Conference, PAKDD 2004, Sydney, Australia, 26–28 May 2004. Proceedings, series Lecture Notes in Computer Science, vol. 3056, pp. 74–85. Springer, Berlin (2004)
6. Wu, S., Wang, H.: Research on the privacy preserving algorithm of association rule mining in centralized database. In: Proceedings of the 2008 International Symposiums on Information Processing, ISIP'08, pp. 131–134, Washington, DC, USA. IEEE Computer Society (2008)
7. Oliveira, S.R.M., Zaane, O.R.: Protecting sensitive knowledge by data sanitization. In: Proceedings of the Third IEEE International Conference on Data Mining, ICDM'03, pp. 613–618, Washington, DC, USA. IEEE Computer Society (2003)
8. Domadiya, N.H., Rao, U.P.: Hiding sensitive association rules to maintain privacy and data quality in database. In: 2013 IEEE 3rd International Conference on Advance Computing Conference (IACC), pp. 1306–1310 (2013)
9. Modi, C.N., Rao, U.P., Patel, D.R.: Maintaining privacy and data quality in privacy preserving association rule mining. In: 2010 International Conference on Computing Communication and Networking Technologies (ICCCNT), pp. 1–6, July 2010
10. Wu, Y.H., Chiang, C.M., Chen, A.L.P.. Hiding sensitive association rules with limited side effects. IEEE Trans. Knowl. Data Eng. **19**(1), 29–42 (2007)
11. Retail database. http://fimi.ua.ac.be/data/retail.dat
12. Verykios, V.S., Gkoulalas-Divanis, A.: A survey of association rule hiding methods for privacy, vol. 34. In: Advances in Database Systems, pp. 267–289. Springer, Berlin (2008)
13. Verykios, V., Elmagarmid, A., Bertino, E., Saygin, Y., Dasseni, E.: Association rule hiding. IEEE Trans. Knowl. Data Eng. **16**(4), 434–447 (2004)
14. Saygin, Y., Verykios, V.S., Clifton, C.: Using unknowns to prevent discovery of association rules. SIGMOD Rec. **30**(4), 45–54 (2001)
15. Saygin, Y., Verykios, V., Elmagarmid, A.: Privacy preserving association rule mining. In: Twelfth International Workshop on Research Issues in Data Engineering: Engineering E-Commerce/E-Business Systems. RIDE-2EC 2002. Proceedings, pp. 151–158 (2002)

16. Wang, S.L., Jafari, A.: Using unknowns for hiding sensitive predictive association rules. In: IEEE International Conference on Information Reuse and Integration, IRI, pp. 223–228, Aug 2005
17. Modi, C., Rao, U.P., Patel, D.R.: An efficient approach for preventing disclosure of sensitive association rules in databases. In: Arabnia, H.R., Hashemi, R.R., Vert, G., Chennamaneni, A., Solo, A.M.G. (eds.) Proceedings of the 2010 International Conference on Information and Knowledge Engineering, IKE 2010, 12–15 July 2010, Las Vegas Nevada, USA, CSREA Press, 2010, pp. 303–309