Optimized Support Balance Model For Sensitive Rule Hiding

Dr Geetha Mary A, Attunuru Mohana Samara Simha Reddy, Keshav Agarwal, Sumegha Sawa

Abstract: Data products are designed to generate information for public or business policy, and research and public information [1]. When data mining and database techniques are advanced, the security of classifying sensitive data in a database becomes a primary issue when providing information to data snoopers. Association rule hiding [2] investigation is a ground-breaking and mainstream apparatus for finding connections covered up in expansive informational indexes. The relationships can be shown as random item-sets or association rules. Rules are categorized as sensitive and non-sensitive depending upon the type of information it can reveal to an adversary. The rule is liable if its probability of disclosure exceeds a given threshold. PPDM is an important tool that can be applied in different fields such as healthcare, e-commerce, product research. [3]The rule is liable if the probability of disclosure exceeds the given threshold. (PPDM)privacy preserving Data Mining is the mainframe that can be applied in different fields, for example, e-healthcare, product research, e-commerce etc. There is a need to hide sensitive rules of the association. The main approach is to decrease the rules' support or confidence. This is achieved by altering the details of the transaction. This generates side effects such as the generation of new rules and non-sensitive rules are falsely hidden [4]. This article proposes an efficient algorithm to hide association rules using the new support formula. This formula has been tested on the value of range 20 to 500. Whenever the algorithm is applied, the range is checked and the corresponding value is entered [5].

Index Terms: Association rules, DSR, ISL, Privacy Preserving Data mining, minimum confidence, minimum support, sensitive rules.

1 INTRODUCTION

Data rules the world. It is important to keep personal data safe before publishing the data on the public domain. Data hiding techniques such as anonymization, association rule hiding techniques, clustering and many more play a major role to hide sensitive data. When we say about anonymization- k-anonymity, l diversity, t closeness comes in the role. When we think about clustering graph-based data hiding is solved. When we talk about association rule data hiding techniques ISL(increase shift left) and DSR (decrease shift right) come in the role. Through this paper, our motive is to thoroughly study the existing association rule hiding techniques and find a better solution. When we look at existing algorithms, we came to know that ISL deals with hiding the sensitive rule by increasing the left side of the rule. The problem with this rule was that it generates new rules that are never listed [6].Whereas DSR does the same by decreasing the right side of rule resulting in hiding some of the non-sensitive rules. Thus we came to advance version of ISL which can solve both of these problems and can lead to optimizing association rule hiding techniqueAfter surfing through a large number of research papers and articles, it is found that applying ISL or DSR is not very effective in hiding sensitive rules but also hides the rules that are of knowledge. This research work is an experiment that is to modify ISL such that only sensitive rules which are to be hidden will have the confidence less than the minimum confidence.

2 BACKGROUND

We find the roots back in machine learning about Association rule mining. It is the method or a technique that is being used for discovering [7] the relations between values. They have been used drastically in every filed from medicines to entertainment.

- *Dr Geetha Mary A- Head of Department (Data Science) at Vellore Institute of Technology, Vellore, Email-geethamary@vit.ac.in*
- *Attunuru Mohana Samara Simha Reddy, Keshav Agarwal and Sumegha Sawa currently pursuing B.Tech(CSE with specialization in Information Security) at Vellore Institute of Technology, Vellore ,India*

Let us understand the concept behind the Association rule mining. While going to super-markets one can observe that the bread is always kept near to the milk booth. Another example can be baby powder is kept next to the diapers [8]. These are all analyzed using association rule mining and then kept together Person who wants to purchase bread, may end up purchasing milk as well. It will help to increase the purchase as well as it will be easier for the customers to get all the required products [9]. Two key problems with Association Rule mining are the high cost of generating association rules and a large number of rules that are generated [10].There are some models that propose a sensitive rule hiding model to hide sensitive fuzzy association rules [11]. The use of a triangular and trapezoidal membership function is done to get fuzzified information [12]. Data mining is considered as the group of methodology that we apply to the data. Big data can also be considered for deducing the important information by using data mining [13]. In Increase Shift Left, when there is a need to hide any association rule, either the support value is decreased to the less than given support or to be less than pre-defined minimum confidence value. Confidence can be decreased by increasing the support count [13].For the DSR, the transactions containing P as well as Q, if the support count of Q is decreased, then the rule's right side will decrease the confidence at a high rate. To minimize the support count of an item, it is preferred to decrease the right side to 0 leaving the left side as 1. Plausible amount of research has been done in the field of association rule hiding [14]. It can be seen that in data mining, data is modified or perturbed in such a way that it cannot be identified by the data mining algorithms [15]. Data anonymization can be divided into three groups cryptographic methods, random anonymization, and tokenization. Under encryption, encrypted data provides high protection and low utility. It is not in a readable form. In this case, protection is 1 and utility is 0. Random anonymization is to modify the data. There is a loss of some utility for the sake of privacy [14]. Tokenisation is the process to convert sensitive data into tokens that have no connections with the original data.

3 ASSOCIATION RULE HIDING

To explore the Association rule [15] mining, we have to go in the details of transaction and item sets. For an example, there be I item sets and each transactions T in which each of it is a set of items that is a subset of I. In a supermarket, we can easily see that where the teacups are kept, there you can also find saucer plates in the next shelf. This is because the maximum number of customers around 80% who will buy teacups may also by saucer plates. Here 80% will be the confidence. [16] By this, one can get to know that the confidence gives the degree of measure of the correlation between the teacups and the saucer plates. Support gives the significance of the correlation between items. The rules generated from frequent item-sets that have their support and confidence greater than the minimum support and minimum confidence are called Association rules. This can be seen from the example given belowSupport gives the significance of the correlation between items. The association rules that will be generated from the prominent item-sets having the confidence, as well as support greater than the minimum confidence, are termed as Association rules [17].

TABLE 1
DATA SET

ID	Transaction
1	000101
2	010111
3	101110
4	101011
5	011000

For this data set minimum support is specified as 60% and minimum confidence is specified as 60%

The existing algorithm sensitive rules 6→4 and 6→5 hidden but with this rule, 3→6 is also generated.

4 PROPOSED MODEL

In this section, the authors are proposing a formula for finding the minimum support. They are also providing the algorithm for applying methods for hiding association rules.
s support, m is minimum support, v is the given value and i is the row increased.

$$s = m - (v * i)$$

This formula has been tested on the value of range 20 to 500. Whenever the algorithm is applied, the range is checked and the corresponding value is entered.

TABLE 4
VALUES BY WHICH SUPPORT SHOULD DECREASE VALUES BY

Range	Value
20-30	4
31-40	2.5
41-50	2
51-60	1.5
61-70	1.25
71-90	1
91-110	0.85
111-130	0.7
131-150	0.6
151-200	0.5
201-300	0.4
301-400	0.26
401-500	0.2

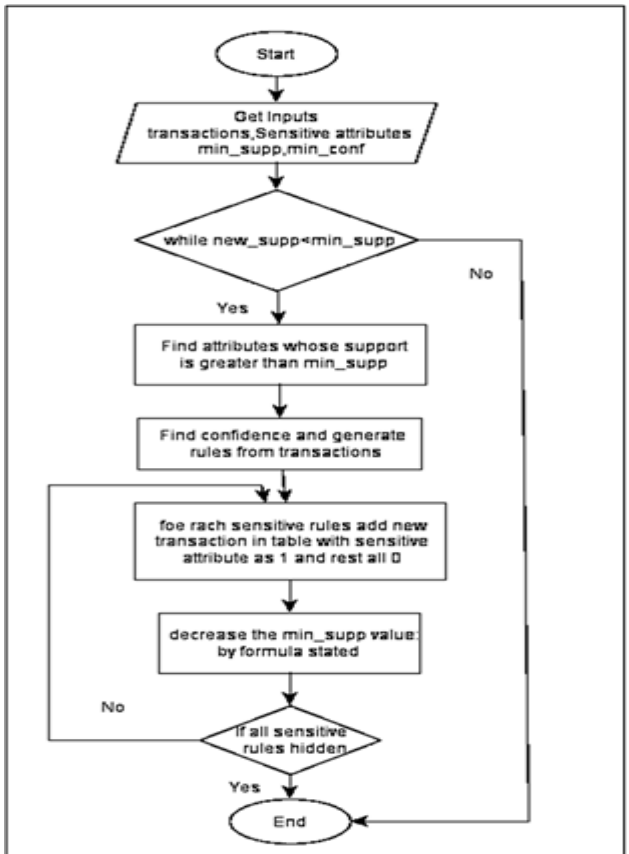

Fig. 1. flow-chart of the overall system of the Proposed Model

The following are the notations used in our proposed Algorithm:

TABLE 5

ABBREVIATIONS USED IN THE ALGORITHM

DS	data set
Min_Sup	minimum support
Min_Conf	minimum confidence
MDS	Modified Data set
New_min_supp	The support generated by the formula
Tn	Total number of Transactions
Att	Total no. of attributes in the table
S	Sensitive Rules
S_var	Sensitive variables
P=>Q	Rules generated
i, j	looping variable

```
Input:DSt,Min_Sup,Min_Conf,
Output:New_min_Supp,MDS with hidden sensittve rules

1.begin
2.  get inputs
3.   while New_Min_Sup<Min_Sup or init==1
4.   do
5.       New_Min_Sup=Min_Sup
6.       for each j: Att
7.            for each i: Tn
8.                Find the frequency of each Att
9.            end for
10.           Storing the frequency and Support of each
    Att in array
11.      end for
12.      do
13.        if Support>Min_Sup
14.           Store in list
15.        end
16.      for each Att stored generate all possible rules
17.           Check the min_Conf of the rules
18.           Generate Rules
19.       end for
20.      for each S_var:
21.          Adding Tn with S_var  as 1 and others Att 0
22.       end for
23.       Display new Tn
24.       New_Min_Sup=Min_Sup-((rows added in
Tn)*Tabular Value);
25.      Display New_Min_Sup
26.    go line 3.
27.end
```

Fig. 2. Algorithm for optimizing the support Balance

There are six attributes, |I|=6, having twenty transactions, |D|=20, in the considered database

TABLE 6

DATA SET FOR SOLVING THE QUESTION USING THE PROPOSED ALGORITHM

ID	Transactions
1	101110
2	000101
3	010111
4	101011
5	011000
6	100111
7	001011
8	101101
9	100111
10	100110
11	101100
12	010111
13	101001
14	011010
15	111100
16	100101
17	000011
18	101010
19	001111
20	101000

From the above dataset, the two sensitive attributes found to be 1 and 4 where minimum confidence [18] is 66% and minimum support [18] is 60%.


```
Input:DSt,Min_Sup,Min_Conf,
Output:New_min_Supp,MDS with hidden sensittve rules

1.begin
2.  get inputs
3.   while New_Min_Sup<Min_Sup or init==1
4.   do
5.       New_Min_Sup=Min_Sup
6.       for each j: Att
7.            for each i: Tn
8.                Find the frequency of each Att
9.            end for
10.           Storing the frequency and Support of each
    Att in array
11.      end for
12.      do
13.        if Support>Min_Sup
14.           Store in list
15.        end
16.      for each Att stored generate all possible rules
17.           Check the min_Conf of the rules
18.           Generate Rules
19.       end for
20.      for each S_var:
21.          Adding Tn with S_var  as 1 and others Att 0
22.       end for
23.       Display new Tn
24.       New_Min_Sup=Min_Sup-((rows added in
Tn)*Tabular Value);
25.      Display New_Min_Sup
26.    go line 3.
27.end
```

Fig. 2. Algorithm for optimizing the support Balance

TABLE 8
TOTAL TRANSACTIONS

Transaction ID	Transaction
1	101110
2	000101
3	010111
4	101011
5	011000
6	100111
7	001011
8	101101
9	100111
10	100110
11	101100
12	010111
13	101001
14	011010
15	111100
16	100101
17	000011
18	101010
19	001111
20	101000
21	100000
22	100000
23	000100
24	000100

TABLE 9
RULES AFTER HIDING THE SENSITIVE ATTRIBUTES

Rules	Confidence
3→1	66.66%
5→6	66.66%
6→4	66.66%
6→5	66.66%

5 RESULT ANALYSIS

After applying the algorithm, it is observed that the rules 1→3, 1→4, 4→1 and 4→6 are hidden and no new rules are generated. Minimum support comes out to be 44% for the given dataset. Confidence for the rules that are not to be hidden is given in Table 8. Also, Table 7 shows the total transactions with increased transactions, that is 4 rows increased in our case

6 CONCLUSION AND FUTURE SCOPE

The proposed model limits all the side effects and hides all given sensitive rules. In this example, the rules of the type A->BC or AB->C do not cross the minimum confidence value. Authors are successful to establish all the possible relation of form A->B and were successful to hide the sensitive rules without generating any ghost rules and without hiding existing rules. But the rules which have more than two attributes are still the areas where they have to work on. The authors hope that the proposed algorithm could promote the motivation for finding the alterations in the given algorithm and making it more efficient.

REFERENCES

[1] M. Atallah, E. Bertino, A. Elmagarmid, M. Ibrahim and V. Verykios, "Disclosure limitation of sensitive rules," in Institute of Electrical and Electronics Engineers Inc, Chicago, 1999.

[2] V. S. Verykios and A. G. Divanis, "A Survey of Association Rule Hiding Methods for Privacy," in Advances in Database Systems, vol. 34, Boston, Springer,boston,MA, 2008, pp. 267-287.

[3] K. I. o. I. Systems. [Online].

[4] P. Fournier-Viger. [Online].

[5] B. J. and M. d. Mare, " One-Way Accumulators: A Decentralized Alternative to Digital Signatures," in Advances in Cryptology, Berlin, Springer, 1993.

[6] Y.-H. Wu, C. Chiag, L. P. Arbee, W. YI-Hung and G. M. A.

[7] S.-L. Wang, K.-W. Huang, T.-C. Wang and T.-P. Hong, "Maintenance of discovered informative rule sets: incremental deletion," in ieee international conference on systems, man and cybernetics., 2005.

[8] cirworld.org. [Online].

[9] S. V. Mohan and T. A. Mohan, "Association Rule Hiding in Privacy Preserving," International Journal of Information Security and Privacy, 2018.

[10] J. Li, H. Shen and R. Topor, "Mining the Smallest Association Rule Set for Predictions," in IEEE International Conference on Data Mining, 2001.

[11] "www.inderscience.com," [Online].

[12] G. M. A., D. P. Acharjya and N. C. Sriman, "Privacy preservation in fuzzy association rules using rough set on intuitionistic fuzzy approximation spaces and DSR," Int.J.Autonomous and Adaptive Communications Systems, pp. 67-86, 2017.

[13] K. Venkataram and G. M. A., "Review on Big Data & Analytics – Concepts, Philosophy, Process and Applications," Cybernetics and Information Technologies, vol. 17, p. 27, 2017.

[14] S. J. Rizvi and J. R. Haritsa, "Maintaining data privacy in association rule mining," in Internation Conference on Very Large Data Bases, 2002.

[15] V. Garg, A. Singh and D. Singh, "A Survey of Association Rule Hiding Algorithms," in International Conference on Communication Systems and Network Technologies (CSNT), 2014.

[16] "ccsenet.org," [Online].

[17] epdf.pub, "epdf.pub," [Online].

[18] "www.bridgeport.edu," [Online].

[19] V. S. Verykios., "State-of-the-art in privacy preserving

data mining," no. ACM SIGMOD Record, 2004.

[20] I. S., A. Z. and E. , "Comprehensive Survey on Privacy Preserving Association Rule Mining: Models, Approaches,Techniques and Algorithms," International Journal on Artificial Intelligence Tools, 2014.

[21] N. R. Adam and J. C. Wortmann, "Security-Control Methods for Statistical Databases:A Comparitive Study," Association for Computing Machinery, vol. 21, 1989.