Modified Hybrid Algorithm for Protecting Sensitive Information

Tanuj Kashyap

Abstract---Providing security to sensitive data against unauthorized access has been a long term goal for the database security research community and for the government statistical agencies. Recent advances in data mining technologies have increased the disclosure risks of sensitive data. Hence, the security issue has become, recently, a much more important area of research. Therefore, in recent years, privacy-preserving data mining has been studied extensively. A number of algorithmic techniques have been designed for privacy-preserving data mining. Most methods use some form of transformation on the data in order to perform the privacy preservation. Such methods reduce the impact of data. This reduction impact of data results in some loss of effectiveness of mining algorithms. A lot of research has already been done to solve the problem. In this paper, we will introduce an efficient algorithm to protect sensitive information.

I. INTRODUCTION:

[1]Data mining is a technique that helps to extract important data from a large database. Recent developments in computing and automation technologies have resulted in computerizing business and scientific applications in various areas. Turing the massive amounts of accumulated information into knowledge is attracting researchers in numerous domains as well as databases, machine learning, statistics, and so on. From the views of information researchers, the stress is on discovering meaningful patterns hidden in the massive data sets. Hence, a central issue for knowledge discovery in databases, additionally the main focus of this thesis, is to develop economical and scalable mining algorithms as integrated tools for management systems.

Data mining, that is additionally cited as knowledge discovery in databases, has been recognized because the method of extracting non-trivial, implicit, antecedently unknown, and probably helpful data from knowledge in databases. The information employed in the mining method usually contains massive amounts of knowledge collected by computerized applications. As an example, bar-code readers in retail stores, digital sensors in scientific experiments, and alternative automation tools in engineering typically generate tremendous knowledge into databases in no time. Not to mention the natively computing- centric environments like internet access logs in net applications. These databases therefore work as rich and reliable sources for information generation and verification. Meanwhile, the massive databases give challenges for effective approaches for information discovery.

The discovered information will be utilized in many ways in corresponding applications. For instance, distinctive the oft times appeared sets of things in a very retail info will be used to improve the choice creating of merchandise placement or commercial. Discovering patterns of client browsing and buying (from either client records or net traversals) could assist the modeling of user behaviors for client retention or customized services. Given the specified databases, whether relational, transactional, spatial, temporal, or transmission ones, we have a tendency to could get helpful info once the information discovery method if acceptable mining techniques square measure used.

II. RELATED WORK:

There is a large amount of work related to association rule hiding. Maximum researchers have worked on the basis of reducing the support and confidence of sensitive association rules [2, 3, and 4]. ISL and DSR are the common approaches used to hide the sensitive rules.

The work in [5] proposed a hybrid method to hide a rule by decreasing either its support or its confidence. This method uses features of both ISL & DSR algorithms. This is done by decreasing the support or the confidence n units at a time by modifying the values of transactions.

In 2008, Belwal et al[6] presented an algorithm. In this method, if one wants to hide any specified association rule $X \rightarrow Y$ our algorithm works on the basis of confidence ($X \rightarrow Y$) and support ($X \rightarrow Y$).To hide the rule $X \rightarrow Y$ (containing sensitive element X on LHS),our algorithm increases the special variable of the rule $X \rightarrow Y$ until confidence ($X \rightarrow Y$) goes below a minimum specified threshold confidence (MCT).As the confidence ($X \rightarrow Y$) goes below MCT(minimum specified confidence threshold),rule $X \rightarrow Y$ is hidden i.e. it will not be discovered through data mining algorithm

III. MODIFIED HYBRID ALGORITHM

To hide any specified association rule $X \rightarrow Y$ this algorithm works on the basis of confidence ($X \rightarrow Y$) and support ($X \rightarrow Y$). To hide any sensitive rule $X \rightarrow Y$, this algorithm first finds all those rules in which Y is in RHS then it finds all those transactions in which Y is 1 and the LHS is also 1. Then in all those transactions it makes Y = 0. In this algorithm, we are not required to hide the rules in which Y is in RHS, because these rules are already hidden by the previous step.

- Step 1: Transaction Data Base, Rule Data Base, MCT (Minimum Confidence Threshold) are the inputs.
- Step 2: Enter the sensitive element
- Step 3: Find all those rules in the rule data base which contains sensitive element on the RHS & whose confidence is greater than the MCT.
- Step 4: For each rule which contains a sensitive item on RHS Repeat step 4
- Step 5: While the data set is not empty
  - Find all those transactions where Sensitive item = 1 and LHS = 1
  - Then put sensitive item = 0 in all those transactions. In this way, the confidence will become less than the MCT (Minimum Confidence Threshold)
- Step 6: Exit

IV. RESULT ANALYSIS

– A Data Set [6]

Let us consider a transaction data base as follows:

Table 1: A Data set [6]

TID	Items
1	ABD
2	B
3	ACD
4	AB
5	ABD

One has also given a MST of 60% and a MCT of 70%. One can see four association rules can be found as below

Table 2: Rule Table

L	R	Supp	Conf
A	B	60%	75%
B	A	60%	75%
A	D	60%	75%
D	A	60%	100%

Now there is a need to hide D and B.

V. PREVIOUS METHODS:

One can see that by simple **ISL** algorithm if someone want to hide D and B, then he can check it by modifying the transaction T2 from B to BD (i.e. from 0100 to 0101).but still ISL cannot hide the rule D → A. Let us see by following example

Table 3: Bitmap of Data Set

TID	Items	Bit Map
1	ABD	1101
2	B	0100
3	ACD	1011
4	AB	1100
5	ABD	1101

Table 4: Hiding D → A by ISL approach

TID	Items	Bit Map
1	ABD	1101
2	B	***0101***
3	ACD	1011
4	AB	1100
5	ABD	1101

So by above explanation it is clear that rule D → A can not be hidden by ISL approach because by modifying 2 from B to BD (i.e. from 0100 to 0101) rule D → A will have support and confidence 60% and 75% respectively.

– By DSR approach:

Table 5: Hiding D → A by DSR approach

TID	Items	Bit Map
1	ABD	***0101***
2	B	0100
3	ACD	1011
4	AB	1100
5	ABD	1101

By DSR approach rule D → A is hidden as its support and confidence is now 40% and 66% respectively, but as a side effect the rule A → D is also hidden.

– By Hiding Counter Approach:

Rule	support	confidence	special variable
A → B	60%	75%	0
B → A	60%	75%	0
A → D	60%	75%	0
D → A	60%	100%	0

– *First to hide B*

Rule	support	confidence	special variable
A → B	60%	75%	0
B → A	50%	60%	1(Rule is hidden)
A → D	60%	75%	0
D → A	60%	100%	0

– *Now to hide D*

Rule	support	confidence	special variable
A → B	60%	75%	0
B → A	50%	60%	1**(Rule is hidden)**
A → D	60%	75%	0
D → A	43%	60%	2

VI. BY HYBRID APPROACH AND PROPOSED ALGORITHM :

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 1, 2, 3, 4 with A = B = 1. Put 0 for item A in all the four transactions. After this modification, we get Table 3 as the modified table.

Table 6: A Data Set [5]

TID	ABC
1	011
2	011
3	011
4	010
5	100
6	101

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.

Table 7: Modified Data Set After Hiding A

TID	ABC
1	011
2	011
3	011
4	010
5	100
6	001

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. So it is clear that the hybrid algorithm unnecessarily scans the database. Because it scans the data

base to find the same sensitive item A in LHS and it doesn't make any difference because item A is already hidden in the data base. Proposed algorithm 2 removes this problem of hybrid algorithm.

VII. RESULT COMPARISON:

The comparison table is as follows:

Table 7: Comparison of Algorithms

Algorithm	No of Rules Pruned	No. of Database Scans
Hybrid	6	6
Proposed Algorithm	6	3

VIII. CONCLUSION:

In this paper, we have proposed a modified algorithm for privacy preserving in data mining. We have also presented a review of the related work. The experimental results have shown that the modified hybrid algorithm takes less data base scans than the previous hybrid algorithm.

REFERENCES:

[1] A K. Pujari. Data Mining Techniques (book), 2001. University Press (India) limited.
[2] V. S. Verykios, A. K. Elmagarmid, B. Elisa, Y. Saygin, and D. Elena. Association rule hiding. In IEEE Transactions on Knowledge and Data Engineering, volume 16(4), pages 434–447, Los Alamitos, CA, USA, April 2004. IEEE Computer Society.
[3] M. Atzori, F. Bonchi, F. Giannotti, and D. Pedreschi. Blocking anonymity threats raised by frequent itemset mining. In Proceedings of the 5th IEEE International Conference on Data Mining (ICDM'05), pages 561–564, Houston, Texa, USA, November 2005. IEEE Computer Society.
[4] Vi-Hung Wu, Chia-Ming Chiang, and Arbee L.P. Chen, Senior Member, IEEE Computer Society Hiding Sensitive Association Rules with Limited Side Effects , VOL. 19, NO.1, JANUARY 2007. IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING,
[5] Ila Chandrakar, Manasa, Usha Rani, and Renuka. Hybrid Algorithm for Association Rule mining. Journal of Computer Science 6(12), pages 1494-1498, 2010
[6] Belwal, Varsheney, Khan, Sharma, Bhattacharya. Hiding sensitive association rules efficiently by introducing new variable hiding counter. Pages 130-134, 978-2008, IEEE.

