International Journal of Research, Science, Technology & Management
ISSN Online: 2455-2240, Volume 12 Issue 2, August 2018

AN EFFICIENT TECHNIQUE PRIVACY PRESERVING ASSOCIATION RULE DATA MINING USING MODIFIED HYBRID ALGORITHM

Kunal Sharma, Shakti Kumar

IJRSTM

ABSTRACT

Many organization and businesses wants to secure their data from illegitimate access. In that priority is given to data privacy and security concerns. It must be necessary for them to share their information about data for the sake of getting satisfactory results. The problem is how these individuals or parties can compare their data or share it without disclosing the sensitive data to each other. It is also always supposed that the multiple parties who want to compare or share data. The concept of data mining has been with for long time, but it took novel computing technology and software of last decades to develop effective tool recent days. Data mining is a powerful tool but like all powerful things is subject to abuse, misuse and ethical considerations. To ensure the integrity of its use and therefore the confidence of the users, the research must be adequately regulated to privacy issues. Failure to do so will increase the hesitation of individuals as well as organizations from releasing or exchanging data which will affect the performance of these organizations and limit their ability to take steps for the future. It will create its own set of problems. When we hide sensitive rules by existing algorithm then it meets with the limitation such as Rule loss, false rules generated and also hides non-sensitive rules. To overcome this limitation there is a need to introduce a new algorithm

International Journal of Research, Science, Technology & Management
ISSN Online: 2455-2240, Volume 12 Issue 2, August 2018

INDEX TERMS: Association rule, Confidence, Data mining, Privacy Preserving, Support

1. INTRODUCTION

Association rule was working of that form X -> Y where X,Y subset of I are the sets of items called Item sets and X ∩Y = Φ. Association rules show attributes value conditions that appear frequently together in a transaction data base. A mostly example are used of association rule data mining is Market Basket Analysis [2]. The set of items is-

I = {Milk, Bread, Butter, Beer}

A rule for the shopping market could be **{Butter, Bread} => {Milk}** meaning that if butter and bread are bought, customers also buy milk. **Association rules** [1, 2] provide information on the basis of "if-then" statements. These rules are computed from the data and, unlike the if-then rules of logic, the association rules are probabilistic. If 90% of transactions that purchase bread and butter, then also purchase milk.

Antecedent: bread and butter

Consequent: milk

Confidence factor: 90%

In addition to the antecedent (the "if" part) and the consequent (the "then" part), an association rule has two numbers that express the degree of uncertainty about the rule. Associations rule analysis the collection of antecedent and consequent are sets of items (called item sets) that are also known as disjoint. It means that they do not have any item in common.

Support for an association rule X->Y is the percentage of transaction in database that contains X U Y. The second big parameter is called the **Confidence** of the rule. Strength for an association rule X U Y is the ratio of number of transactions that contains X U Y to number of transaction that contains X.

Support X=>Y

$$= \frac{\text{Common item in any giving table}}{\text{Total no transaction in any table}}$$

Confidence X=>Y

$$= \frac{\text{Total Support in Number (A U B)}}{\text{Total support in Number (A)}}$$

Association Rule Hiding

The problem of association rule hiding was first probed in 1999.After that, many approaches were proposed. They are categorized as - data sanitization data modification approaches and knowledge

IJRSTM

sanitization data reconstruction approaches.

The data modification approaches [6, 7] are also the so-called data sanitization. They generally hide sensitive association rules by directly modifying sanitizing the original data D, to the database D' directly from D. As the sanitization is performed on data level, data modification approaches cannot control the hiding effects intuitively. Moreover, it is found that the data sanitization can produce a lot of I/O operations.

2. 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 [3,4 and 5]. ISL and DSR are the common approaches used to hide the sensitive rules.

The work in [8] 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 [9] presented an algorithm. In this method, if one wants to hide any specified association rule X → Y our algorithm works on the basis of confidence (X → Y) and support (X → Y).To hide the rule X → Y (containing sensitive element X on LHS),our algorithm increases the special variable of the rule X → Y until confidence (X → Y) goes below a minimum specified threshold confidence (MCT).As the confidence (X → Y) goes below MCT(minimum specified confidence threshold),rule X → Y is hidden i.e. it will not be discovered through data mining algorithm.

3. PROPOSED WORK

To hide any specified association rule X → Y this algorithm works on the basis of confidence (X → Y) and support (X → Y). To hide any sensitive rule X → 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. The complete procedure is as follows:

IJRSTM

INPUT:

I.

database of transactions

II.

database of rules

III.

set of sensitive items X

IV.

minimum support threshold (MST)

V.

minimum confidence threshold (MCT)

OUTPUT:

A transformed database of transactions where rules containing X will be hidden.

PROCEDURE:

I check for all sensitive elements.

For each x in X where x belongs to X

{

// Now check all the rules containing sensitive element x.

For each rule R which contain x on RHS

{

II Check whether Modified confidence of the rule

I go below MCT or not.

While (D is not empty)

II decrease the confidence of rule

{

Find all those transactions where x = 1 and LHS = 1

Then put x = 0 in all those transactions

}

}

}

End of procedure

3.1. A DATA SET

Suppose there is a database of transactions as below:

Table 1

TID	Items
ABC	T1
ABC	T2
ABC	T3
AB	T4
A	T5
AC	T6

Fig 4.2: A Data Set

Suppose MCT is 50%.

Table 2

ABC	TID
111	T1
111	T2
111	T3
110	T4
100	T5
101	T6

IJRSTM

The all possible rules with confidences are:

A–>B (66.66%),

A–>C (66.66%),

B–>A (100%),

B–>C (75%),

C–>A (100%),

C–>B (75%),

3.2. BY HYBRID APPROACH AND PROPOSED ALGORITHM 2

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. Choose B–>A and search all transactions with B = A = 1.There are four transactions T1, T2, T3, T4 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 3

ABC	TID
011	T1
011	T2
011	T3
010	T4
100	T5
101	T6

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 4

ABC	TID
011	T1
011	T2
011	T3
010	T4
100	T5
001	T6

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.

IJRSTM

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.

Table 3.3: Comparison Table

Algorithm	No of Rules Pruned	No. of Database Scans
Hybrid Algorithm	**6**	**6**
Proposed Algorithm 2	**6**	**3**

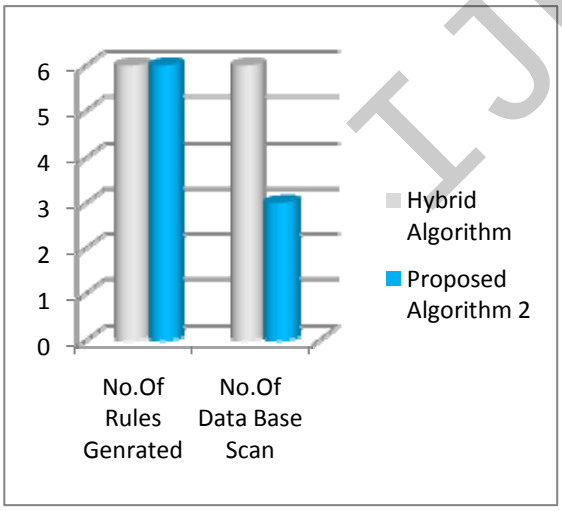

Chart 3.4 Pictorial representation of Comparison table

4. CONCLUSIONS AND FUTURE WORK

We presented a fundamental approach in order to protect sensitive rules from display. The approach reduces the importance of the rules by of large item sets until it is below a user-specified threshold, so that no rules can be derived from the selected item sets. We also measured the performance of the proposed algorithms according to two criteria: the time that is required by the hiding process and the side effects that are produced. As side effects, we considered loss of information and generation of false information. We lose information whenever some rules, originally mined from the database, cannot be retrieved after the privacy preserving process. We are adding some information whenever anyone rules that could not be fetch before the hiding can be mined from the released database.

IJRSTM

REFERENCES

[01] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In Proceedings of the ACM SIGMOD Conference on Management of Data, pages 207–216, New York, NY, USA, May 1993. ACM Press.

[02] A. K. Pujari. Data Mining Techniques (book), 2001. University Press (India) limited.

[03] R. Chen, K. Sivakumar, and H. Kargupta. Distributed web mining using Bayesian networks from multiple data streams. In N. Cercone, T. Young Lin, and X. Wu, editors, Proceedings of the 2001 IEEE International Conference on Data Mining (ICDM'01), pages 75–82, San Jose, California, USA, November 2001. IEEE Computer Society.

[04] S. Goldwasser. Multi-party computations: Past and present. In Proceedings of the 16th Annual ACM Symposium on the Principles of Distributed Computing, pages 1–6, USA, 1997. ACM Press.

[05] M. Atallah, E. Bertino, A. Elmagarmid, M. Ibrahim, and V. Verykios. Disclosure limitation of sensitive rules. In Proceedings of 1999 IEEE Knowledge and Data Engineering Exchange Workshop (KDEX'99 pages 45–52, Chicago, Illinois USA, November 1999. IEEE Computer Society.

[06] E. Dasseni, V. S. Verykios, A. K. Elmagarmid, and E. Bertino. Hiding association rules by using confidence and support. In I. S. Moskowitz, editor, Proceedings of the 4th Information Hiding Workshop, volume 2137, pages 369–383, 2001. Springer Veralg Lecture Notes in Computer Science.

[07] S. R. M. Oliveira and O. R. Zaiane. Algorithms for balancing privacy and knowledge discovery in association rule mining. In Proceedings of the 7th International Database Engineering and Applications Symposium (IDEAS'03), pages 54–65, China, 2003. IEEE Computer Society.

IJRSTM