# AN UPDATED HYBRID TECHNIQUE FOR PRESERVING PRIVACY IN DATA MINING USING ASSOCIATION RULE HIDING

[1]Arpit S/o Rameshwar Rathod, [2]Prof. Rahul Moriwal

AITR, Indore

*Abstract:* **In this paper, we provide here an overview of the new and rapidly emerging research area of privacy preserving data mining. Privacy preserving in data mining is a very popular research topic. A large number of researchers are working on improving security in data mining. Also a detailed review of the work accomplished in this area is also given along with the coordinates of each work to the classification hierarchy. The critical review of some modern data hiding approaches is also performed. A upgraded hybrid technique is also proposed for preserving privacy in data mining using association rule hiding. The proposed updated method is taking less number of data base scans to hide an association rule**

## Introduction:

Data mining is the non-trivial process of identifying valid and potentially useful patterns in data. Many governmental organization, businesses etc are finding a way to collect, analyze and report data about individuals ,households or businesses, in order to support (short and long term) planning activities. Information system contains private or confidential information like their social security number, income of employees, purchasing of customer etc, that should be properly secured.

Privacy preserving data mining [4,8] is a new investigation in data mining and statistical databases [1]. In PPDM data mining algorithms are analyzed for side effects obtain in data privacy. There is a two fold consideration in privacy preserving data mining. The first is sensitive raw data that are kept secure from unauthorized access like identifiers, names ,addresses should be modified from original database. The second one is sensitive knowledge is excluded that can be mined from a database by using data mining algorithms as such type of knowledge compromises data privacy.

The problem for finding an optimal sanitization to a database against association rule analysis has been proven to be NP-Hard [2]. In [3], authors presented three algorithms 1.a, 1.b and 2.a for hiding sensitive association rules. The work done in [5] contains two algorithms are built based on blocking for rule hiding. The first one focuses on hiding the rules by reducing the minimum support of the itemsets that generates these rules. The second algorithm focuses on reducing the minimum confidence of the sensitive rules. In [6] and [7] algorithms based on blocking technique are proposed and analyzed. The work in [9] proposed a hybrid method to hide a rule by decreasing either its support or its confidence.

## Problem Statement

The problem of sensitive rule hiding is described as follows: Given a transaction database, MST, MCT, a set of strong rules, and a set of sensitive items, how can we modify the database such that using the same MST and MCT, the set of strong rules in the modified database satisfies all the constraints: 1) no sensitive rule, 2) no lost rule, and 3) no false rule?

Let D be the database of transactions and J = {J1, ..., Jn} be the set of items. A transaction T includes one or more items in J . An association rule has the form X → Y , where X and Y are non-empty sets of items (i.e. X and Y are subsets of J) such that X ∩ Y = Null. A set of items is called an itemset, while X is called the antecedent. The support of an item (or itemset) x is the percentage of transactions from D in which that item or itemset occurs in the database The confidence or strength c for an association rule X → Y is the ratio of the number of transactions that contain X or Y to the number of transactions that contain X.

The problem of mining association rule is to find all rules that have support and confidence greater then user specified minimum support threshold (MST) and minimum confidence threshold (MCT).

As an example, for a given database in following table, a minimum support of 33% and a minimum confidence of 70%, nine association rules can be found as follows: B=>A(66%, 100%), C=>A (66%, 100%), B=>C (50%, 75%),C=>B (50%, 75%), AB=>C

(50%, 75%), AC=>B (50%,75%), BC=>A (50%, 100%), C=>AB (50%, 75%),B=>AC (50%, 75%).

**Table 1**

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

The objective of privacy preserving data mining is to hide certain sensitive information so that sensitive information can not be discovered through data mining techniques. Given a transaction database, a minimum support threshold and minimum confidence threshold and set of sensitive items X, the objective is to modify database in such a way that no predictive association rule containing X on the left hand side will be discovered. So if in above example element A is sensitive then rules AB=>C (50%, 75%), AC=>B (50%, 75%) should not be discovered by data mining algorithm.

**Proposed Algorithm**

- **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 5**
- **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**

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 T1, T2, T3, T4 with A = B = 1. Put 0 for item A in all the four transactions. After this modification, we get Table 2 as the modified table.

**Table 2**

| TID | ABC |
|---|---|
| T1 | 011 |
| T2 | 011 |
| T3 | 011 |
| T4 | 010 |
| T5 | 100 |
| T6 | 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 3**

| TID | ABC |
|---|---|
| T1 | 011 |
| T2 | 011 |
| T3 | 011 |
| T4 | 010 |
| T5 | 100 |
| T6 | 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 3 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 removes this problem of hybrid algorithm.

**Conclusion:**

In this paper, we have presented a survey various privacy preserving data mining algorithms. The critical review, of privacy preserving in data mining techniques, done in this paper will help the researchers to overcome the drawbacks of existing algorithms and make an efficient algorithm. This paper also contains an updated hybrid technique for privacy preserving in data mining using association rule hiding. It is taking less number of data base scans to hide association rules containing sensitive data.

**References**

[1] ArisGkoulalas–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

[2] M. Atallah, E. Bertino, A. Elmagarmid, M. Ibrahim, and V. S. Verykios "Disclosure limitation of sensitive rules,".*In Proc. of the 1999 IEEE Knowledge and Data Engineering Exchange Workshop (KDEX'99)*, pp. 45–52, 1999.

[3] 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.

[4] Stanley R. M. Oliveira; Osmar R. Za¨_ane, "Privacy Preserving Frequent Itemset Mining", IEEE International Conference on Data Mining Workshop on Privacy, Security, and Data Mining, Maebashi City, Japan. Conferences in Research and Practice in Information
Technology, Vol. 14.2002

[5] Y.Saygin, V. S. Verykios, and C. Clifton, "Using Unknowns to Prevent Discovery of Association Rules," *ACM SIGMOD*, vol.30(4), pp. 45–54, Dec. 2001.

[6] Y. Saygin, V. S. Verykios, and A. K. Elmagarmid, "Privacy preserving association rule mining," *In Proc. International Workshop on Research Issues in Data Engineering (RIDE 2002),* 2002,pp. 151–163.

[7] E. Pontikakis, Y. Theodoridis, A. Tsitsonis, L. Chang, and V. S. Verykios."A quantitative and qualitative analysis of blocking in association rule hiding". In Proceedings of the 2004 ACM Workshop on Privacy in the Electronic Society (WPES), pages 29–30, 2004.

[8] A. Gkoulalas-Divanis and V.S. Verykios, "An Integer Programming Approach for Frequent Itemset Hiding," *In Proc. ACM Conf. Information and Knowledge Management (CIKM '06)*, Nov. 2006.

[9]Belwal, Varsheney, Khan, Sharma, Bhattacharya. *Hiding sensitive association rules efficiently by introducing new variable hiding counter*.Pages 130-134, 978-2008, IEEE.