ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

*International Journal of Advanced Research in Computer and Communication Engineering*
*Vol. 1, Issue 9, November 2012*

# An Efficient Algorithm for Privacy Preserving Data Mining Using Heuristic Approach

[1]M.Mahendran, [2]Dr.R.Sugumar, [3]K.Anbazhagan, [4]R.Natarajan

Research Scholar, Dept. of CSE, CMJ University, Shillong, Meghalaya, India

Associate Professor, Dept. of CSE, VelMultitech Dr. RR Dr.SR Engineering College, Chennai, India

Research Scholar, Dept. of CSE, CMJ University, Shillong, Meghalaya, India

Research Scholar, Dept. of CSE, CMJ University, Shillong, Meghalaya, India

**Abstract: P**rivacy preserving data mining is an important topic on which lot of researchers going on last years. There are many approaches to hide association rule. In this paper Efficient Heuristic approach method is proposed which is more effective to hide association rule. The objective of this algorithm is to extract relevant knowledge from large amount of data, while protecting at the time sensitive information. The proposed method focused on hiding set of frequent items containing highly sensitive knowledge that only remove information from transactional database with no hiding failure.

**Keywords:** Minimum Confidence, Minimum Support, Itemset, Association rules.

## I. INTRODUCTION

Privacy preserving data mining (PPDM) is a novel research direction in Data Mining (DM), where DM algorithms are analysed for the side-effects they incur in data privacy. The main objective of PPDM is to develop algorithms for modifying the original data in some way, so that the private data and private knowledge remain private even after the mining process [1]. In DM, the users are provided with the data and not the association rules and are free to use their own tools; So, the restriction for privacy has to be applied on the data itself before the mining phase. For this reason, we need to develop mechanisms that can lead to new privacy control systems to convert a given database into a new one in such a way to preserve the general rules mined from the original database. The procedure of transforming the source database into a new database that hides some sensitive patterns or rules is called the *sanitization process*[2]. To do so, a small number of transactions have to be modified by deleting one or more items from them or even adding noise to the data by turning some items from 0 to 1 in some transactions. The released database is called the sanitized database. On one hand, this approach slightly modifies some data, but this is perfectly acceptable in some real applications[3, 4].

This study mainly focus on the task of minimizing the impact on the source database by reducing the number of removed items from the source database with only one scan of the database. Section-2 briefly summarizes the previous work done by various researchers; In Section-3 preliminaries are given. Section-4 states some basic definitions and of which definition 5 is framed by us which is used in the proposed heuristic based algorithm. In Section-5 the proposed algorithm is presented with illustration and example. As the detailed analysis of the experimental results on large databases is under process, only the basic measures of effectiveness is presented in this paper, after testing the algorithm for a sample generated database.

## II.ASSOCIATION RULE MINING

Let I = {i1,…., in} be a set of items. Let D be a database which contains set of transactions. Each transaction t _ D is an item set such that t is a proper subset of I. As transaction t supports X, a set of items in I, if X is a proper subset of t. Assume that the items in a transaction or an item set are sorted in lexicographic order. An association rule is an implication of the form X_Y, where X and Y are subsets of I and X_Y= Ø. The support of rule X_Y can be calculated by the following equation: Support(X_Y) = |X_Y| / |D|, where |X_Y| denotes the

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

*International Journal of Advanced Research in Computer and Communication Engineering*
*Vol. 1, Issue 9, November 2012*

number of transactions containing the itemset XY in the database, |D| denotes the number of the transactions in the database D. The confidence of rule is computed by Confidence(X_Y) = |X_Y|/|X|, where |X| is number of transactions in database D that contains itemset X. A rule X_Y is strong if support(X_Y) _ min_support and confidence(X_Y) _ min_confidence, where min_support and min_confidence are two given minimum thresholds.

Association rule mining algorithms calculate the support and confidence of the rules. The rules having support and confidence higher than the user specified minimum support and confidence are retrieved. Association rule hiding algorithms prevents the sensitive rules from being revealed out. The problem can be declared as follows "Database D, minimum confidence, minimum support are given and a set R of rules are mined from database D. A subset SR of R is denoted as set of sensitive association rules.SR is to be hidden. The objective is to modify D into a database D' from which no association rule in SR will be mined and all non sensitive rules in R could still be mined from D'.

## III. APPROACHES OF ASSOCIATION RULE HIDING ALGORITHMS

Association rule hiding algorithms can be divided into three distinct approaches. They are heuristic approaches, border-revision approaches and exact approaches.

*A.Heuristic Approach*

Heuristic approaches can be further categorized into distortion based schemes and blocking based schemes. To hide sensitive item sets, distortion based scheme changes certain items in selected transactions from 1's to 0's and vice versa. Blocking based scheme replaces certain items in selected transactions with unknowns. These approaches have been getting focus of attention for majority of the researchers due to their efficiency, scalability and quick responses.

*B.Border Revision Approach*

Border revision approach modifies borders in the lattice of the frequent and infrequent item sets to hide sensitive association rules. This approach tracks the border of the non sensitive frequent item sets and greedily applies data modification that may have minimal impact on the quality to accommodate the hiding sensitive rules. Researchers proposed many border revision approach algorithms such as BBA (Border Based Approach), Max– Min1 and Max-Min2 to hide sensitive association rules. The algorithms uses different techniques such as deleting specific sensitive items and also attempt to minimize the number of non sensitive item sets that may be lost while sanitization is performed over the original database in order to protect sensitive rules.

*C. Exact Approach*

Third class of approach is non heuristic algorithm called exact, which conceive hiding process as constraint satisfaction problem. These problems are solved by integer programming. This approach can be concerned as descendant of border based methodology.

## IV. PRELIMINARIES

*A.Transactional Database:*

A transactional database is a relation consisting of transactions in which each transaction *t* is characterized by an ordered pair, defined as *t* = **<** *Tid, list-of-elements***>** , where *Tid* is a unique transaction identifier number and *list-of-elements* represents a list of items making up the transactions. For instance, in market basket data, a transactional database is composed of business transactions in which the *listof- elements* represents items purchased in a store.

*B. Basics of Association Rules:*

One of the most studied problems in data mining is the process of discovering association rules from large databases. Most of the existing algorithms for association rules rely on the support confidence framework introduced in [8]. Formally, association rules are defined as follows: Let $I = \{i1,...,in\}$ be a set of literals, called items. Let *D* be a database of transactions, where each transaction *t* is an itemset such that . A unique identifier,

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

*International Journal of Advanced Research in Computer and Communication Engineering*
*Vol. 1, Issue 9, November 2012*

called *Tid*, is associated with each transaction. A transaction *t* supports *X*, a set of items in *I*, if . An association rule is an implication of the form , where , and . Thus, we say that a rule holds in the database *D* with *support* if , where *N* is the number of transactions in *D*.Similarly, we say that a rule holds in the database *D* with *confidence* ) if , where is the number of occurrences of the set of items *A* in the set of transactions *D*. While the support is a measure of the frequency of a rule, the confidence is a measure of the strength of the relation between sets of items. Association rule mining algorithms rely on the two attributes, *minimum Support(minSup )* and *minimum Confidence(minConf )*. The problem of mining association rules have been first proposed in 1993[8].

*C. Frequent Pattern:*
A pattern X is called a frequent pattern if *Sup(X)* ≥ *minSup* or if the *absolute support* of X satisfies the corresponding *minimum support count* threshold. [pattern is an itemset; in this article, both terms are used synonymously]. All association rules can directly be derived from the set of frequent patterns[8, 9]. The conventions followed here are o *Apriori property[10]: all non empty subsets of a frequent itemsets(patterns) must also be frequent.* o *Antimonotone property: if a set cannot pass a test, then all of its supersets will fail the same test as well.*

## V. PROPOSED ALGORITHM

In order to hide an association rule, X → Y, we can either decrease its support or its confidence to be smaller than user-specified minimum support transaction (MST) and minimum confidence transaction (MCT). To decrease the confidence of a rule, we can either (1) increase the support o of X, the left hand side of the rule, but not support of X → Y, or (2) decrease the support of the item set X →Y .For the second case, if we only decrease the support of Y, the right hand side of the rule, it would reduce the confidence faster than simply reducing the support of X → Y. To decrease support of an item, we will modify one item at a time by changing from 1 to 0 or from 0 to 1 in a selected transaction.

Based on these two concepts, we propose a new association rule hiding algorithm for hiding sensitive items in association rules. In our algorithm, a rule X → Y is hidden by decreasing the support value of X →Y and increasing the support value of X. That can increase and decrease the support of the LHS and RHS item of the rule correspondingly. This algorithm first tries to hide the rules in which item to be hidden i.e., X is in right hand side and then tries to hide the rules in which X is in left hand side. For this algorithm t is a transaction, T is a set of transactions, R is used for rule, RHS (R) is Right Hand Side of rule R, LHS (R) is the left hand side of the rule R, Confidence (R) is the confidence of the rule R, a set of items H to be hidden.

**ALGORITHM:**

**INPUT:** A source database D, A minimum support min_support (MST), a minimum confidence min_confidence (MCT), a set of hidden items X.

**OUTPUT:** The sanitized database D, where rules containing X on Left Hand Side (LHS) or Right Hand Side (RHS) will be hidden.

**Steps of algorithm:**
1. Begin

2. Generate all possible rule from given items X;

3. Compute confidence of all the rules for each hidden item H, compute confidence of rule R.

4. For each rule R in which H is in RHS

    4.1 If confidence (R) < MCT, then

    Go to next 2-itemset;

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

*International Journal of Advanced Research in Computer and Communication Engineering*
*Vol. 1, Issue 9, November 2012*

Else go to step 5

5. Decrease Support of RHS item H.

5.1 Find T=t in D fully support R;

5.2 While (T is not empty)

5.3 Choose the first transaction t from T;

5.4 Modify t by putting 0 instead of 1 for RHS item;

5.5 Remove and save the first transaction t from T; End While

6. Compute confidence of R;

7. If T is empty, then H cannot be hidden;

8. For each rule R in which is in LHS

9. Increase Support of LHS;

10. Find T=t in D| t does not support R;

11. While (T is not empty)

12. Modify t by putting 1 instead of 0 for LHS item;

13. Remove and save the first transaction t from T; End While

14. Compute confidence of R;

15. If T is empty, then H cannot be hidden;

End For;

End Else;

End For;

16. Output update D, as the transformed D;

The framework of the proposed approach is shown in figure 1.

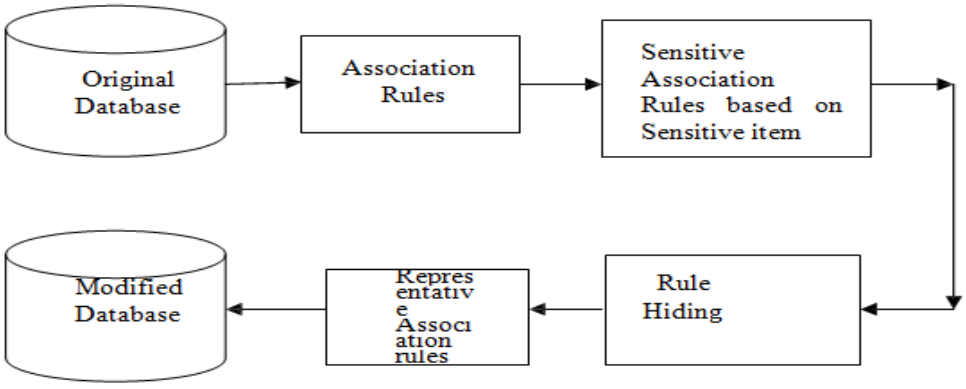


**Figure 1: Association Rule Hiding Framework**

## VI. PERFORMANCE EVALUATION

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

*International Journal of Advanced Research in Computer and Communication Engineering*
*Vol. 1, Issue 9, November 2012*

Association rule mining over woman's clothing store [3] is considered a basic knowledge discovery activity. For discovering correlations among items, Association rule mining provides a useful mechanism belonging to customer transactions in a woman's clothing store database. Let D be the database of transactions and I= {I1, ..., In} be the set of items. A transaction T includes one or more items in I .An association rule has the form A→ B, where A and B are non-empty sets of items (i.e. A and B are subsets of I) such that A ∩ B = Null. A set of items is called an itemset, while A 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 A → B is the ratio of the number of transactions that contain A or B to the number of transactions that contain A.

*A.SOLUTION BY PROPOSED METHOD*

We take an example of woman's clothing store in which we are having four items {Jeans, T-shirt, Skirt, Shoes} and five transactions [4]. We assume minimum support threshold (MST) of 60% and minimum confidence threshold (MCT) of 70% .

| **TID** | **ITEMS** |
|---|---|
| T1 | JEANS,T SHIRT ,SHOES |
| T2 | TSHIRT |
| T3 | JEANS,SKIRT,SHOES |
| T4 | JEANS,TSHIRT |
| T5 | JEANS,TSHIRT,SHOES |

Table 1: Transactions for Table I

One has also given a MST of 60% and a MCT of 70%. One can see four association rules can be found as below- JEANS->TSHIRT (60%, 75%) TSHIRT->JEANS (60%, 75%) JEANS->SHOES (60%, 75%) SHOES->JEANS (60%, 100%) Now there is a need to hide TSHIRT and SHOES as it is sensitive.

Table 2: Initial Association Rule Constraints Data Table

| | SUPPORT | CONFIDENCE | SV |
|---|---|---|---|
| JEANS->TSHIRT | 60% | 75% | 0 |
| TSHIRT->JEANS | 60% | 75% | 0 |
| JEANS->SHOES | 69% | 75% | 0 |
| SHOES->JEANS | 60% | 100% | 0 |

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

*International Journal of Advanced Research in Computer and Communication Engineering*
*Vol. 1, Issue 9, November 2012*

*B.Approach to hide TSHIRT*

Table 3: Transactions for Table II

| TID | ITEMS |
|---|---|
| T1 | JEANS,TSHIRT,SHOES |
| T2 | TSHIRT |
| T3 | JEANS,SKIRT,SHOES |
| T4 | JEANS,SKIRT,SHOES |
| T5 | JEANS,TSHIRT,SHOES |

TABLE 4: Data Table for hiding TSHIRT

| | SUPPORT | CONFIDENCE | SV |
|---|---|---|---|
| JEANS->TSHIRT | 60% | 75% | 0 |
| TSHIRT->JEANS | 48% | 58% | 1 |
| JEANS->SHOES | 60% | 75% | 0 |
| SHOES->JEANS | 60% | 100% | 0 |

*C. Approach to hide SHOES*

Table 5: Transactions for Table III

| TID | ITEMS |
|---|---|
| T1 | JEANS,TSHIRT,SHOES |
| T2 | TSHIRT |
| T3 | JEANS,SKIRT,SHOES |
| T4 | JEANS,SKIRT,SHOES |
| T5 | JEANS,TSHIRT,SHOES |

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

*International Journal of Advanced Research in Computer and Communication Engineering*
*Vol. 1, Issue 9, November 2012*

| | |
|---|---|

The above table contains five transactions. In this approach we hide shoes(using LHS) using hiding algorithm, so for hiding shoes hidden counter runs two times. So we get the values of support and confidence below minimum support threshold and minimum confidence threshold. So by our approach the rule for jelly is hidden as shoes is sensitive element.

TABLE 6: Data Table for hiding SHOES

| | SUPPORT | CONFIDENCE | SV |
|---|---|---|---|
| JEANS->TSHIRT | 60% | 75% | 0 |
| TSHIRT->JEANS | 48% | 58% | 1 |
| JEANS->SHOES | 60% | 75% | 0 |
| SHOES->JEANS | 42% | 58% | 2 |

## VII. CONCLUSIONS

In this paper, the database privacy problems are addressed and a new technique for privacy preservation is proposed. Association rule hiding techniques are used to hide sensitive association rules. A new heuristic method to hide the sensitive association rules is proposed. Data distortion technique is applied so that sensitive information cannot be discovered through data mining techniques. Confidence of the rules is represented as representative rules. Confidence of the rule is recomputed and compared with threshold level. The confidence of the sensitive rules might be reduced while maintaining the support. From the experimental results, it is observed that all the rules containing sensitive items are hidden. The algorithm is implemented and numerical example is shown. Further research is in progress to evolve a method which can avoid the computational overhead associated with confidence of the rules.

## VIII. REFERENCES

[1] Alberto Trombetta and Wei Jiang (2011), 'Privacy-Preserving Updates to Anonymous and Confidential Databases', IEEE Transactions on Knowledge and Data Engineering, Vol. 22, pp. 578-568.

[2] Gayatri Nayak and Swagatika Devi (2011), 'A Survey On Privacy Preserving Data Mining: Approaches And Techniques', International Journal of Engineering Science and Technology, pp.2127-2133

[3] Guang Li and Yadong Wang (2011), 'Privacy-Preserving Data Mining Based on Sample Selection and Singular Value Decomposition', Proceedings of the IEEE International Conference on Internet Computing and Information Services , pp.298-301

[4]Jain Y.K. (2011), 'An Efficient Association Rule Hiding Algorithm for Privacy Preserving Data Mining', International Journal of Computer Science and Engineering, pp.96-104

[5]Kyriakos Mouratidis and Man Lung Yiu (2011), 'Anonymous Query Processing in Road Networks', IEEE Transactions on Knowledge and Data Engineering, Vol. 22, pp. 2-16.

[6] Baris Yildiz and Belgin Ergenç (2010), 'Hiding Sensitive Predictive Frequent Itemsets', Proceedings of the International Multi Conference of Engineers and Computer Scientists, Hong Kong, Vol.1, pp.572-576

[7] Bo Peng and Xingyu Geng (2010), 'Combined Data Distortion Strategies for Privacy-Preserving Data Mining', Proceedings of the 3rd International Conference on Advanced Computer Theory and Engineering, pp. 241-253.

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

*International Journal of Advanced Research in Computer and Communication Engineering*
*Vol. 1, Issue 9, November 2012*

[8] Brian, Loh and Patrick (2010), 'Ontology-Enhanced Interactive Anonymization in Domain-Driven Data Mining Outsourcing', Proceedings of the 2nd International Symposium on Data, Privacy, and Ecommerce, pp.9-15

[9] Du W. and Zhan Z (2010), 'Using randomized response techniques for privacy-preserving data mining', Proceedings of the 9th ACM International Conference on Knowledge Discovery and Data Mining, Washington, pp. 142-157

[10]Gkolalas Divanis and Verykios V. (2010), 'Exact knowledge hiding through database extension', IEEE Transactions on Knowledge and Data Engineering, Vol. 21, pp. 699-713.

[11]Islam M. and Brankovic L. (2010), 'Noise Addition for Protecting Privacy in Data Mining', Proceedings of the 6th Engineering Conference on Mathematics and Applications, pp.207-219

[12]Jaideep Vaidya and Chris Clifton (2010), 'Leveraging the multi in Secure Multiparty Computation', Proceedings of the 5th IEEE International Conference on Data Mining , pp120-128

[14]Krish Srinivasa Rao and Chiranjeevi (2010), 'Distortion Based Algorithms For Privacy Preserving Frequent Item Set Mining', International Journal of Data Mining and Knowledge Management Process, Vol.1, No.4, pp. 1033-1045

[15] Dr.Sugumar Rajendran, Dr.Rengarajan Alwar, Dr.Saravanakumar Selvaraj, "Determining the Existence of Quantitative Association Rule Hiding in Privacy Preserving Data Mining", International Journal of Advanced Research in Computer and Communication EngineeringVol.1,Issue 2,pp.104-109.

[16] Mark Shaneck and Yongdae Kim (2010), 'Efficient Cryptographic Primitives for Private Data Mining', Proceedings of the 43rd Hawaii International Conference on System Sciences, pp.77-89.