International Journal on Emerging Technologies (Special Issue on ICRIET-2016) **7**(2): 65-69(2016)

ISSN No. (Print) : 0975-8364
ISSN No. (Online) : 2249-3255

A Pandect on Association Rule Hiding Techniques

Dr. K. Prabha [*] ***and T. Suganya*** [**]

**Assistant Professor, Department of Computer Science, Periyar University PG Ext. Centre, Dharmapuri – 636705,Tamilnadu, India*
***Ph.D. Research scholar, Department of Computer Science, Periyar University PG Ext. Centre, Dharmapuri – 636705, Tamilnadu, India.*

(Corresponding author: T. Suganya)

(Received 16 September, 2016 Accepted 19 October, 2016)
(Published by Research Trend, Website: www.researchtrend.net)

ABSTRACT: The most useful technology to extract the information or knowledge from large database is the data mining. Data mining is used to deal with the huge size of the data stored in the database. It has various techniques for the extraction of data, association rule is the most effective data mining technique. There are many techniques used in Privacy preserving data mining to hide association rules and generated by association rule generation algorithms. Association rule hiding is the method of modifying original database to make the sensitive rules disappear. The main challenging issues are the security and the privacy. Applications of association rule include health insurance, fraudulent discovery and loss-leader analysis, telecommunication networks market and risk management, inventory control etc., Association rule mining aims at extraction, hidden relation and interesting associations between the existing items in a transactional database. In this paper, present a study of hiding techniques, applications, importance and the approaches.

Keywords: Association rule hiding, Data mining, Privacy preserving data mining

I. INTRODUCTION

Data mining is the process of analyzing data from different perspectives and summarizing it into useful information. Data mining software is one of a number of analytical tools for analyzing data. Data mining sometimes called data or knowledge discovery. Data mining, also known as knowledge discovery in databases, has been recognized as a new area for database research. The area can be defined as efficiently discovering interesting rules from large collections of data.

Besides extracting information or knowledge from raw data, there is also need for some technique or scheme that deal with security of that information, privacy preserving in data mining (PPDM) is the technique that deal with the security of the information that extracted by data mining techniques, PPDM allow to mine the information from large amount of data while protecting sensitive information defined by the data base owner, or the information that database owner do not want to disclose. The main aim of PPDM is to minimized the risk of misuse of data while does not affect the data mining techniques. Privacy preserving data mining is first introduced by Agrawal and Srikant[1].

A. Association Rule

It is defined as the implication X→Y like if / then statements. Where X and Y are the set of items.
Example:- "If a customer buys a dozen eggs, he is 80% likely to purchase milk".

B. Support

Measures of how the collection of items in an association occurs together as a percentage of all the transactions.

C. Confidence

Confidence of rule" X given Y" is a measure of how much more likely it is that Y occurs when A has occurred.
An Association rule has two parts:-

1. *Antecedent i.e. if part:* - An Antecedent is an item found in data.
2. *Consequent i.e. then part:* - A Consequent is an item found in combination with antecedent. e.g.- Association Rule X→ Y , X – An Antecedent Y – A Consequent.

D. Frequent Itemset

An itemset X is called frequent item set in the transaction database D if supp(X) ≥ minsupp. If X is frequent and no superset of X is frequent, X is denoted as a maximal item set.

E. Closed Itemset

A closed itemset is defined as an itemset X which has the property of being the same as its closure, i.e., X= cit(X). The minimal closed itemset containing an itemset Y is obtained by applying the closure operator cit to Y.

F. Optimal Rule Set

A rule set is optimal with respect to interestingness metric if it contains all the rules except those with no greater interestingness than one of its more general rules. An optimal rule set is a subset of a non redundant rule set.

Association rule mining is to find out association rules that satisfy the predefined minimum support and confidence from a given database. The problem is usually decomposed into two sub-problems [3].

1. Find those itemsets whose occurrences exceed a predefined threshold in the database; those itemsets are called frequent or large itemsets.

2. To generate association rules from those large itemsets with the constraints of minimal confidence.

II. ASSOCIATION RULE HIDING

The association rule hiding is one of the techniques that used in PPDM. The association rule hiding methodologies aim at sanitizing the original database in a way that at least one of the following goals is accomplished [4].

A. No rule that is considered as sensitive from the owner's perspective and can be mined from the original database at pre-specified thresholds of confidence and support can be also revealed from the sanitized database, when this database is mined at the same or at higher thresholds.

B. All the non sensitive rules that appear when mining the original database at pre-specified thresholds of confidence and support can be successfully mined from the sanitized database at the same thresholds or higher.

C. No rule that was not derived from the original database when the database was mined at pre specified thresholds of confidence and support can be derived from its sanitized counterpart when it is mined at the same or at higher threshold.

D. Association rule hiding process totally depend on support or confidence of the rule, there is two way to hide any rule ,either decrease support up to certain threshold or decrease confidence up certain threshold, so the mining algorithm, that works on support not able to mine sensitive rules.

1. By decreasing the numerator item support while keeping the support of denominator unchanged.

2. By increasing support of denominator items while keeping the support of numerator items unchanged.

Fig.1.1 is showing general framework for association rule hiding. However the modification on the database may cause some side effect that may lead to some disturbance in association rule mining, following are the some side effects that may occur in the process of rule hiding:

- ***Lost Rules*****:** The non sensitive association rules which are present in original database and can be mined by applying mining algorithm but cannot be mined after applying hiding algorithm from modified database.
- ***False rules*****:** The sensitive association rules which are not hidden by hiding algorithm and can be mine by applying mining algorithm on modified database.
- ***Ghost rules*****:** The rules which are not present in original database but generated after applying hiding algorithm.

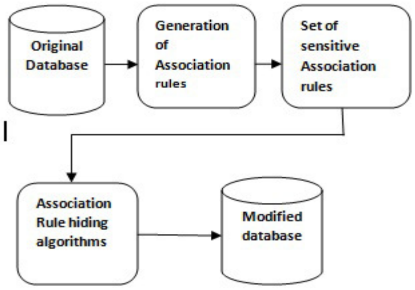
Original Database
Generation of Association rules
Set of sensitive Association rules
I
Association Rule hiding algorithms
Modified database

Fig. 1. General Framework for hiding sensitive association rule.

The association rules represent the associations between the data variables. An association rule is an implication of the form written below [2]. X □Y [Support= S%, Confidence=C%], where X, Y ⊂ I and X ∩ Y = Φ, and I is an Itemset. X is called as the Antecedent or body and Y is called as Consequent or head of the rule. Each rule has two measures of value support and confidence. The computation of support and confidence can be defined by the following equations: Support (X□Y) = P (XUY) Confidence (X□Y) =P(Y/X) = support_count (XUY) / support_count(X) Where support S is the probability that rule contains {X, Y} and confidence C is the conditional probability that specify the C% of the transaction of database considered must specify X□Y. Minimum support and Minimum confidence are needed to eliminate the unimportant association rules.

The association rule holds iff it has the support and confidence value greater than or equal to minimum support (min_sup) and minimum confidence (min_conf) threshold value. An example of calculating support and confidence for transactional database given in Table 1. is described below:

Table 1: Example of Transactional Database Customer.

S.No	Item Purchased Customer (A)	Item Purchased Customer (B)
1	Pizza	Coke
2	Burger	Sprite
3	Pizza	Sprite
4	French fries	Coffee

Table 1.1: If A is "purchased pizza" and B is "purchased soda" then Support = P (A and B) = ¼ Confidence = P (B / A) = ½

Confidence does not measure if the association between A and B is random or not.

III. ASSOCIATION RULE MINING

The main aim of association rule mining is to extract frequent item sets, correlation and association among different set of items in the transactional database, relational databases or other information repository. Association rule mining algorithm finds association rules in the form of: IF AB and CD then HELLO IF UV and XY then BYE Here AB, CD, UV and XY are different objects out of which if any person takes AB and CD then due to high probability, he will take HELLO. Similarly if he will choose UV and XY then he will choose BYE. In general, expressions which are in the form of A=>B, called association rules where A represents antecedent and B represents consequent. Association rules represent how many times B has occurred if A has already occurred depending on the chosen support and confidence value. Here support is nothing but the probability of items or item sets in the given database (like transactional or other) and confidence represents conditional probability.

A. Apriori Algorithm:

In general, Apriori algorithm [8] works on two phases – first phase is to choose minimum support value which is applied in the database to find frequent item sets while in second phase, these item sets and the minimum confidence constraints are used to generate rules.

The pseudo code for the Apriori algorithm are given as follows -

```
Step 1: let Cn be the candidate item set of size n.
Step 2: let Fn be the frequent item set of size n.
Step 3: F1 = {Frequent items}
Step 4: REPEAT
Step 5: Cn+1 = Candidates generated from Fk ;
Step 6: REPEAT for each transaction t in database
Step 7: increment the count of all candidates in Cn+1 that are
        contained in t.
Step 8: Fk+1 = Candidates in Cn+1 with minimum support.
Step 9: UNTIL (Fn not equal to □ )
Step 10: return Cn Fn
```

IV. ASSOCIATION RULE MINING PROCESS

Association rule mining is a two-step process:

A. Find All Frequent Item-sets: First find all the sets of items whose support count value is equal to or more than minimum support count value. All these item sets are termed as frequent itemsets.

B. Generate strong association rules from the frequent itemsets: Second, for each frequent itemsets generate the association rules that have confidence value more than or equal to minimum confidence value. Once the frequent itemsets from transactions in a database have been found, it is straightforward to generate the strong association rules from the frequent itemsets.

This can be done by using the equation for confidence. Based on this, association rules can be generated as follows: For each frequent itemset L, generate all nonempty subsets of L. For every non empty subset S of L, output the rule "S□ (L-S)" If (support_count(L))/(support_count(S)) >= min_conf Where min_conf is the minimum confidence threshold. Because the second step is much less costly than the first, the overall performance of algorithm for mining association rules is determined by the first step. Different methods follow the different approach to generate the frequent itemsets. Once the frequent itemsets are found, generation of association rules from frequent patterns is easier.

V. IMPORTANCE OF ASSOCIATION RULE MINING PROCESS

The importance of association rule mining is as follows: [7]

- The association rule mining helps in finding particular relationships between various data elements of the large database i.e. database having a large number of records (108- 1012 bytes).
- Association rule mining helps in the classification of data.
- It helps to search useful information and knowledge that can enhance the business or scientific operations.

- To provide better and efficient methods to analyze the data. It can handle data of high dimensionality.
- Increased competition for customers requires availability of information on demand.
- Association rule mining helps in finding such required useful information.
- Helps in finding the outlier entries, which may be useful in some cases such as fraud detection.

VI. ASSOCIATION RULE MINING APPROACH

Association rule mining approach can be divided into two classes: [6]

A. Bottom-Up Approach

Bottom Up Approach look for frequent itemsets from the given dataset that satisfy the predefined constraint. Bottom up approach gets large frequent itemsets through the combination and pruning of small frequent itemsets. The principle of the algorithm is: firstly calculates the support of all itemsets in candidate itemset Ck obtained by Lk-1, if the support of the itemset is greater than or equal to the minimum support, the candidate k-itemset is frequent k-itemset, that is Lk, then combines all frequent k-itemsets to a new candidate itemset Ck+1, level by level, until finds large frequent itemsets.

A major challenge in mining frequent itemsets from a large data set is the fact that such mining often generates a huge number of itemsets satisfying the minimum support threshold, especially when min sup is set low. This is because if an itemset is frequent, each of its subsets is frequent as well.It is useful to identify a small representative set of itemsets from which all other frequent itemsets can be derived such approach is known as Top-Down approach.

B. Top-Down Approach

Top-down Approach looks for more specific frequent itemsets rather than finding more general frequent itemsets. The number of frequent itemsets produced from a transaction data set can be very large. It is useful to identify a small representative set of itemsets from which all other frequent itemsets can be derived. Two such representations are presented in this section are

1) Maximal Frequent Itemset

2) Closed Frequent Itemset

An itemset X is a maximal frequent itemset (or max-itemset) in set S if X is frequent, and there exists no super-itemset Y such that $X \subseteq Y$ and Y is frequent in S[2].

An itemset X is closed in a data set S if there exists no proper super-itemset Y such that Y has the same support count as X in S. An itemset X is a closed frequent itemset in set S if X is both closed and frequent in S[2].

The precise definition of closed itemset, however, is based on Relations (1) and (2).Given the functions:

$f(T) = \{i \in I \mid \forall t \in T, i \in t\}$

Which returns all the itemset included in the set of transactions T, and

$g(I) = \{t \in T \mid \forall i \in I, i \in t\}$

Which returns the set of transactions supporting a given itemset I (its tid-list), the composite function fog is called Galois operator or closure operator.

- Generator is an itemset p is a generator of a closed itemset y if p is one of the itemsets (there may be more than one) that determines y using Galois Closure operator: h(p) = y.
- It is intuitive that the closure operator defines a set of equivalence classes over the lattice of frequent itemsets: two itemsets belongs to the same equivalence class if and only if they have the same closure, i.e. their support is the same and is given by the same set of transactions.

From the above definitions, the relationship between equivalence classes and closed itemsets is clear: the maximal itemsets of all equivalence classes are closed itemsets. Mining all these maximal elements means mining all closed itemsets.

VII. APPLICATIONS OF ASSOCIATION RULE MINING

To discover knowledge that was not known earlier is used by the association rule mining. It can be used in various domains. Here are some of the applications of ARM. Nestorov and Jukić [9] proposed a data mining framework that is coupled with data warehousing technology to capture co-occurrence patterns towards ad-hoc ARM on OLAP environment. Huang and Hu [10] employed roughest theory and AI technology to mine association rules for expert decision making. Boukerche and Samarah applied ARM in WSN for discovering temporal relationships between sensor nodes to improve Shyu et al.employed ARM to spatial data to discover spatial-temporal correlations. [8] extracted association rules from SVM classification trees. Couturier et al. [8] made HCI easy by visualizing extracted association rules. Pan et al. [9]To mine medical images in healthcare domain is the applied association rule mining. Privacy preserving and quantitative ARM are other useful applications of ARM.

VIII. CONCLUSION

Association rule mining has the two distinct phases such as the extracting frequent item sets from database and generation of association rules.

The technique for hiding sensitive information in database is the association rule hiding. The techniques used in PPDM is the association rule hiding. Support and confidence are the statistical measures used to guard the quality rules to be generated. This paper deals with the overall pandect on hiding techniques in association rules. The future work is to compare hiding techniques by using the association algorithms.

REFERENCES

[1]. Agarwal and Srikant,"Privacy-preserving data mining",In ACM SIGMOD, May 2000, pp.439-450.

[2]. Agarwal.R and Srikant.R Fast algorithms for mining association rules. In Proc.20thInt. Conf. Very large Data Bases, 487-499.

[3]. Jaiwei Han and Micheline Kamber " Data Mining Concepts and Techniques," Second Edition Morgan Kaufmann Publishers.

[4]. M. Atallah, E. Bertino, A. Elmagarmid, M. Ibrahim, and V.S.Verykios "Disclosure limitation of sensitive rules."In Proc.of the IEEE Knowledge and Data Engineering Exchange Workshop.

[5]. Mayank Agarwal, manuj Mishra, Shiv Pratap Singh Kushwah in International Journal of Soft Computing and Engineering (IJSCE) ISSN:2231-2307, Volume-5 Issue-1,March 2015.

[6]. Mihir R Patel, Dipak Dabhi in International Journal of Emerging Technology and Advanced Engineering (ISSN 2250-2459, ISO 9001:2008 Volume **5**, Issue 1, January 2015).

[7]. Rajdeep Kaur Aulakh in International Journal of Advanced Research in Computer Science and Software Engineering (ISSN:2277 128X Volume **5**, Issue 3, March 2015).

[8]. Shaoning Pang and Nik Kasabov. (2008). r-SVMT: Discovering the Knowledge of Association Rule over SVM Classification Trees. IEEE, p2486-2493.

[9]. Svetlozar Nestorov, Nenad Jukić(2002), Ad-Hoc Association-Rule Mining within the Data Warehouse,IEEE,0(0),p1-10.

[10]. ZHE AUANG, WN-QUAN HU (2003), Applying AI technology and rough set theory to mine association rules for supporting knowledge management.IEEE.p1820-1825.