International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)
Web Site: www.ijettcs.org Email: editor@ijettcs.org
Volume 3, Issue 6, November-December 2014 **ISSN 2278-6856**

Study of Mining and Hiding of Sensitive Association Rule

Shintre Sonali Sambhaji[1], Prof. Kalyankar Pravin P.[2]

[1] CSE Dept., Dr. B.A.M.U. University, Aurangabad, T.P.C.T's. College of Engineering, Osmanabad, India.

[2]Associate Professor, CSE Dept
Dr. B.A.M.U. University, Aurangabad, T.P.C.T's. College of Engineering, Osmanabad, India.

Abstract

This paper is organized into two approaches. In first approach an Improved Apriori algorithm is being presented that efficiently generates association rules. This reduces unnecessary database scan at the time of forming frequent large itemsets. In second approach we have tried to hide sensitive association rule by using an improved Apriori algorithm. In this paper, we have used an approach that strategically modifies fewer transactions in a transaction database to decrease support and confidence of a sensitive rule without producing any side effects. Thus, in the paper, we efficiently generate frequent itemset by applying Improved Apriori algorithm and generated association rules by applying minimum support and minimum confidence and then identify the sensitive association rules.

Index Terms:-Association rules, confidence, Data mining methods and Algorithm, Minimum Support Threshold (MST), Minimum Confidence Threshold (MCT), Rule hiding.

1. INTRODUCTION

The term 'Data Mining' indicates a wide range of tools and techniques to extract useful information, which can be sensitive from a large collection of data. The objective of this work is to propose a new strategy to avoid extraction of sensitive data. Data should be manipulated in such a sensitive way that information cannot be discovered through data mining techniques. While dealing with sensitive information it becomes very important to protect data against unauthorized access. The Apriori algorithm is based on modifying the database transaction. It is an important aspect in improving mining algorithm that deals with how to decrease candidate item sets in order to generate frequent item sets efficiently. In the classical Apriori algorithm, when candidate generations are generated, the algorithm needs to test their occurrence frequencies. Association rule is a common technique of data mining, which revealing useful hidden information from the dataset. This technique is popular for discovering behavior from large dataset. In this paper improved algorithm is proposed for mining association rules in generating frequent k- itemsets. Whether these candidates are frequent itemset after generating new candidates, this algorithm finds frequent itemset directly and removes the subset which is not frequent by selecting only those transactions for scanning which are a subset of frequent itemsets. Once frequent itemset are find out, the association rules are generated based on minimum confidence. The association rule whose confidence is greater than or equal to the minimum confidence threshold are identified as sensitive rule and such rule should hide without any side effect before releasing database. This Apriori algorithm used to modify transaction data and inserting new data in the database and removing data from the database. More specify that given a transaction database D, a minimum support, a minimum confidence, and a set of items S to be hidden [1], [2], [3]. While the support measure the frequency of association rules and the confidence is a measure of the strength of the relationship between a set of items in mining sensitive association rules that are greater than the minimum support threshold and minimum confidence threshold.

2.ASSOCIATION RULES

The efficiency of mining association rules is an important field of Data Mining technique. Association rule mining searches for interesting customer habits by looking at the association. Association rule mining finds interesting associations or correlations among a large set of data items. An e.g. of such rule might be that 98% of the customers that purchases keyboard also tend to buy mouse at the same time [4]. Association rules having certain measures such as: support and confidence.

• Support -It is a measure of frequency of a rule.
• Confidence -It is a measure of strength of the relation.

The association rules are written as X→Y means whenever X appears Y also tends to appear. X and Y may be single items or sets of items, but the same item does not appear in both. The 1% presence of Applications in marketing, store layout, customer segmentation, medicine, finance and many more.

- Suppose X and Y appear together in only 1% of the transactions, but whenever X appears there is an 80 % chance that Y also appears.
- The 1% presence of X and Y together is called the support (or prevalence) of the rule and 80% is called the confidence (or predictability) of the rule.
- Confidence denotes the strength of the association between X and Y. Support indicates the frequency of the pattern. A minimum support is necessary if an association is going to be of some business value.

3.IMPROVED APPRIORI ALGORITHM

The efficiency of mining association rules is an important field of Knowledge Discovery in Databases. The Apriori algorithm is a classical algorithm in mining association rules. Apriori employs an iterative approach known as a levelwise search, this paper presents an improved Apriori algorithm to increase the efficiency of generating association rules. This algorithm adopts a new method to reduce the redundant generation of sub-itemsets during pruning the candidate itemsets, which can form directly the set of frequent itemsets and eliminate candidates having a subset that is not frequent. Improved Apriori algorithm mines frequent itemsets. This algorithm adopts a new method to reduce the redundant generation of sub-itemsets during pruning the candidate itemsets, which can form directly the set of frequent itemsets and eliminate candidates having a subset that is not frequent. Let I={i1, i2, … , im} denote the set of items that are displayed in a store. Moreover, let D represent a set of transactions, where each transaction T is a set of items such that T⊆I. A unique identifier, namely TID, is associated with each transaction. Let A be a set of item in I, A transaction T is said to contains A if A⊆T. An association rule is an implication of the form A→B, where A⊂I, B⊂I and A∩B≠Φ. Support of rule, denoted by s, is the percentage of transactions in D that contain A also contain B, it can be computed by |AB|/|D|; Confidence of rule, denoted by c, is the ratio between the number of transactions containing both A and B and the number of transactions containing A, it can be computed as |AB|/|A|. In database D, a set of items is called itemset, and an itemset that contains k items is called k-itemset. Support count is the number of transactions containing an itemset, if an itemset whose support is greater than or equal to a minimum support threshold (called minsup), we name it the frequent itemset. There is a transaction database D, which has nine transactions, as shown in Table 1.

Table 1. Transaction database D

Database D

TID	Items
1	A,B,E
2	B,D
3	B,C
4	A,B,D
5	A,C
6	B,C
7	A,C
8	A,B,C,E
9	A,B,C

Let TID-set(A) denote the set of transaction TID which contain item A in D, so the amount of transaction which contain A in D is the amount of item of TID-set(A), more exactly, sup-count(A) can be computed by |TID-set(A)|. The transaction set in D, which have A and B, is the intersection of TID-set(A) and TID-set(B), and sup-count(A→B) can be computed as |TID-set(A)∩TID-set(B)|. The join and the pruning of Apriori algorithm is improved correspondingly: all non-empty subset of frequent itemset must be frequent; to produces Lk-candidates, we only join the set whose for k-2 item is same in Lk-1. So we propose the following improvement to the traditional Apriori algorithm:

(1) Compute minimum sup-count by min-sup*|D|.
(2) Scan the database once to produce L1-candidates, simultaneously construct TID-set(X1) for each item. After scanning, compute sup-count for each item and find the set of frequent items L1.
(3) Produces L2-candidates from L1*L1.Scanning L1, we can find TID-set(X2) and sup-count of each itemset, deletes the patterns whose frequencies don't satisfy the min support count, and find L2.
(4) In order to produce Lk(k≥3), join itemsets which satisfy the join rule. Scanning Lk-1, we can find TID-set(Xk) and compute sup-count of each itemset, then deletes the patterns whose frequencies do not satisfy the minimum support count, and find Lk.

Using improved Apriori Algorithm, first we produce L1-candidates and find the set of frequent items L1. Sele-join L1 to produce L2-candidates, scan L1, and achieve TID-set(X1) and sup-count, deletes the patterns whose frequencies do not satisfy the minimum support count,. and find L2. By joining the itemset which satisfied rule in L2, we can obtain {{A,B,C}, {A,B,E}, {B,C,D}, {B,C,E}, {A,C,E}, {B,D,E}}. However {A,D}, {C,D}, {C,E}, and {D,E} is not the frequent, and {B,C,D}, {B,C,E}, {A,C,E}, {B,D,E} don't belong to L3-candidates. Compute TID-set(X3) and the support count separately, L3 is composed by itemsets which have the minimum support count. The itemset which satisfies the condition in L3 sele-join, and obtains {{A,B,C,E}}. {B,C,E} is not the frequent itemset, so L4-candidates=ϕ, the algorithm finishes. The process is as shown in Figure 1.

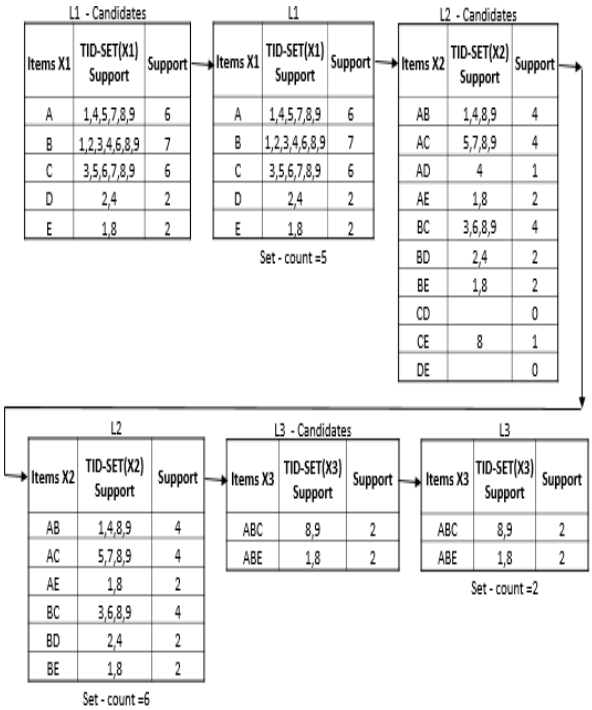

Fig 1. The operation process of the improved Apriori algorithm

In this paper, we have worked on generating strong association rule and hiding sensitive one. Finally, the improved algorithm is verified, the results show that the improved algorithm is reasonable and effective, can extract more value information.

4.METHODOLOGY

In this paper, we have used the two methods mining and hiding of sensitive association rule by using an improved Apriori algorithm with their description of problem formulation and properties rule hiding as below:

Mining of Association Rule:

The count of itemset I denoted as is the number of transactions containing I in the Database and |D| is the count transaction in DB. For two itemset X and Y, where X Y=Φ, X=>Y is called as a strong rule if both following conditions holds.

1. $Sup_{XUY} = C_{XUY} / |D| \geq MST$ and
2. $Conf_{X \rightarrow Y} = C_{XUY} / C_X \geq MCT$.

Mining association rule is one of main data mining research areas at present, and focuses particularly in finding the relation among different items in the database and tries to find frequency patterns that can be represented as knowledge. Mining of frequent itemsets is an important phase in association mining, which discovers frequent itemsets in transactions database. It is core in many tasks of data mining that try to find interesting patterns from datasets, such as association rules. Mining association rules mean finding frequent patterns, associations, correlations between sets of items in transaction databases. The discovery of association rules is divided into two phases [10], [11]: finding the frequent itemsets and generation of association rule. In the first phase, every set of items is called itemset, if they occurred together greater than the minimum support threshold [12], this itemset is called frequent itemset. Finding frequent itemsets is easy but costly, so this phase is more important than second phase. In the second phase, it can generate many rules from one itemset. Minimum support and confidence is defined by the user which represents constraint of the rules. So the support and confidence thresholds should be applied for all the rules to prune the rules in which its values less than threshold values. In association rule mining it firstly find out frequent itemsets that having minimum support and minimum confidence, and using these frequent itemsets association rules are generated.

The association rule mining is a two-step process:

- **Step1:** Find all the frequent item sets in the transaction database. If support of itemset X, support (X) ≥ minsup, then X is a frequent itemset. Otherwise, X is not a frequent itemset.
- **Step2:** Generate strong association rules from frequent itemsets.

The problem of mining association rules is to find all rules that are greater than the user-specified minimum support and minimum confidence [7], [8], [9].

- **Hiding of Sensitive Association Rules:**

Let D' be the database after applying a sequence of modification to D. A strong rule X=>Y in D will be hidden in the D' if one of the following conditions holds in:

1. $Sup_{XUY} < MST$ and
2. $Conf_{X \rightarrow Y} < MCT$

Association Rule Hiding for Data Mining is designed for researchers, professors and advanced-level students in computer science studying privacy preserving data mining, association rule mining, and data mining. Hiding the sensitive rules (sensitive rules are those rules that contain sensitive item(s)). Hiding sensitive rules by changing the support and the confidence of the association rule or frequent itemset as data mining mainly deals with generation of association rules. As it is known that association rule is an important entity, which may cause harm to the confidential information of any business, defence or organization and raises the need of hiding this information in the form of association rules. Saygin [5] and Wang [6] have proposed some algorithms which help in reducing the support and the confidence of the rules. Approach in hides only those rules, which has a sensitive item either in the right or in the left.

Properties for Rule Hiding:

Property 1. Let ∑XUY be the set of all transactions containing XUY. To hide X → Y by removing items in XUY from the transactions in ∑XUY, the minimal number of transactions that should be modified, called the minus support count, is computed as:

$$MSC(X \rightarrow Y) = C(XUY) - [|D| * MST] + 1$$

Property 2. To hide X → Y by removing items in Y from the transactions in ∑XUY, the minimal number of transactions that should be modified, called the minus consequent confidence count, is computed as:

$$MCCCX \rightarrow Y = C(XUY) - [Cx * MCT] + 1$$

Property 3. To hide X → Y by removing items in X from the transactions in ∑XUY, the minimal number of transactions that should be modified, called the minus precedent confidence count, is computed as:

$$MPCCX \rightarrow Y = [(C(XuY) - Cx * MCT) / (1 - MCT)] + 1$$

5.ASSOCIATION RULE HIDING ARCHITECTURE

This experiment shows that sensitive association rule hiding can be done efficiently and easily using improved Apriori algorithm. Following fig. 2 shows rule hiding the architecture in which we generate association rules from frequent large itemset, then we have to mine sensitive association rules, after rule mining process we have to hide sensitive association rules. When we hide the rule, association rules must be updated and the original database is modified. That modified database should be released. For rule hiding process firstly the items and transactions are

selected for modification that is we modify the original database. Then the modified database is released. If some sensitive rules are not hidden, the user can release as it is, release nothing, or relax the constraint to hide more sensitive rules. In the process some rules will be generated that is sensitive rules and nonsensitive rules. Then select the sensitive association rules whose confidence are greater than or equal to the minimum confidence threshold (MCT). The remaining rule will be nonsensitive it will It will generate association Then this database mines the sensitive and nonsensitive rules. After that we have to hide sensitive association rule, but at a time only one rule can hide. While the sensitive rules are hidden as many as possible. We propose an approach that strategically modifies the database by using the modification scheme to decrease the supports or confidences of the rules.

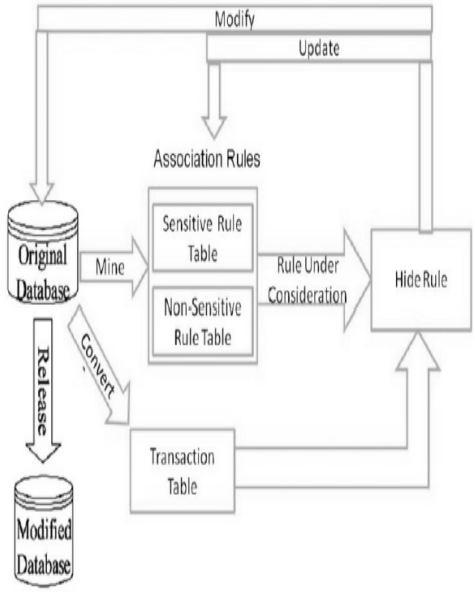

Fig 2. Architecture of Rule Hiding

When we hide the rule, association rules must be updated and the original database is modified. That modified database should be released. For rule hiding process firstly the items and transactions are selected for modification means we modify the original database. Then the modified database is released. If some sensitive rules are not hidden, the user can release as it is, release nothing, or relax the constraint to hide more sensitive rules.

6.EXPERIMENTAL RESULT

In this paper a unique approach is used which has two parts in the first part of this paper, it contains mining of association rule and in second part contains hiding of sensitive association rule by Improved apriori algorithm. For mining of sensitive association rules we have to generate association rules, after generating association rules mining process mines the sensitive association rules. Then we have to mark sensitive association rule. Remaining will be nonsensitive rules. We have to hide sensitive association rule. At a time only one rule we can hide This algorithm increases the efficiency of generation of association rule. This Improved apriori algorithm mines frequent item sets without new candidate generation. Firstly, we have taken a database and mine the association rules from that database with MST=30% and MCT=60%, then we have got rules and from these rules mark milk=>umbrella as sensitive whose confidence is 70% are shown in following fig.3

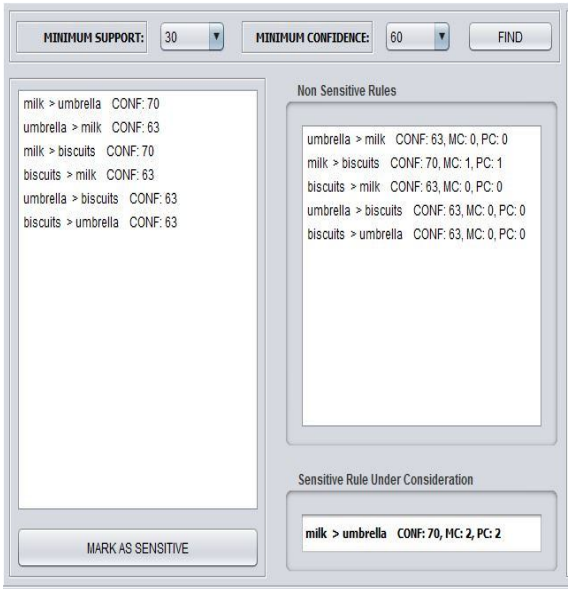

Fig. 3 Mining of Association Rule

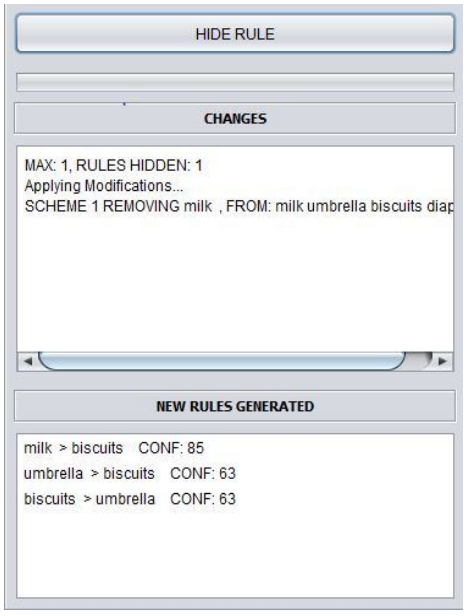

Fig. 4 Hiding of Sensitive Association Rule

After mining the rules we have to hide the sensitive rule. The marked sensitive rule milk=>umbrella is hidden. Above Fig 4 removes milk from milk, umbrella, biscuits and diaper. New rules are generated as milk=>biscuits, umbrella=> biscuits and biscuits=> umbrella. Hide a particular rule for three methods, hide using No False

Rule, hide using No Lost Rules and hide using both methods. The hiding failure side effect measures the number of sensitive association rules that cannot be hidden. The support or the confidence of the large itemsets or the association rule is changed which helps in hiding them. The hiding strategies proposed are based on reducing the support and confidence of rules. In order to achieve this, transactions are modified by removing some items, or inserting some new items depending on the hiding strategy. Each sensitive rule based on the minus count and plus count, this approach can determine the minimal number of transactions that should be modified. In the similar way, the maximal number of transactions that can be modified without hiding a non-sensitive rule can be estimated.

7.CONCLUSION

Apriori algorithm is the study of theory and extensive literature, based on the number of items for mining frequent itemsets put forward an improved algorithm. In this paper, we presented an improved Aprirori Algorithm, in which Association rule mining is an important data mining task that finds interesting association among a large set of data item. It may disclose pattern and various kinds of sensitive information. Such information may be protected against unauthorized access. The main aim of this work is to propose a new method to hide the sensitive association rules. Data will be distorted in such a sensitive way that information cannot be discovered through data mining techniques. The proposed work analyses the existing techniques and gives their limitations. This method will hide all the rules containing the sensitive items without producing any side effect, i.e. no false rules or no lost rules should be generated which maintains the integrity of the database. Finally, the experimental results show that the improved algorithm indeed a higher efficiency, improved performance of the algorithm. Also the strong rules which are generated by the Improved Apriori algorithm by decreasing support and confidence of this algorithm which can reduce the number of database scanning and the redundancy.

ACKNOWLEDGEMENT

Every project big or small is successful largely due to the effort of a number of wonderful people who have always given their valuable advice or lent a helping hand. I sincerely appreciate the inspiration; support and guidance of all those people who have been instrumental in making this project a success. At this juncture, I feel deeply honored in expressing my sincere thanks to all my respective college department members with special thanks to my guide Prof. Pravin P. Kalyankar for making the resources available at the right time and providing valuable insights leading to the successful completion of my project without which this project would not have been possible. Last but not the least I place a deep sense of gratitude to my family members and my friends who have been a constant source of inspiration during the preparation of this project work.

REFERENCES

[1] Guanling Lee, Chien-Yu Chang, Arbee L.P Chen" Hiding Sensitive Patterns in Association Rules Mining", Proceedings of the 28th Annual International Computer Software and Applications Conference (COMPSAC'04), 2004.

[2] Shyue-Liang Wang, Yu-Huei Lee, Steven Billis, Ayat Jafari" Hiding Sensitive Items in Privacy Preserving Association Rule Mining", IEEE International Conference on Systems, Man and Cybernetics, 2004

[3] Xiaoming Zhang, Xi Qiao" New Approach for Sensitive Association Rule Hiding", 2008 International Workshop on Education Technology and Training, IEEE computer society, 2008

[4] S. Chai, J. Yang, Y. Chang, "The Research of Improved Apriori Algorithm for Mining Association Rule", In IEEE International Conference, 2007.

[5] V. S. Verykios,, Ahmed K. Elmagermld, Elina Bertino, Yucel Saygin, Elena Dasseni, "Association Rule Hiding," IEEE Transactions on knowledge and data engineering, Vol. 6, no. 4, (2004)

[6] S-L. Wang, Yu-Huei Lee, S. Billis and A. Jafari, "Hiding sensitive items in privacy preserving association rule mining," IEEE International Conference on Systems, Man and Cybernetics, vol. 4, pp. 3239 – 3244, (2004)

[7] Shyue-Liang Wang, Bhavesh Parikh, Ayat Jafari "Hiding informative association rule sets" Science direct. 2006.

[8] Chih-Chia Weng, Shan-Tai Chen, Hung-Che Lo" A Novel Algorithm for Completely Hiding Sensitive Association Rules" Eighth International Conference on Intelligent Systems Design and Applications, Vol. 2, pp. 202-208,2008.

[9] Yogendra Kumar Jain , Vinod Kumar Yadav, Geetika S. Panday" An Efficient Association Rule Hiding Algorithm for Privacy Preserving Data Mining" International Journal on Computer Science and Engineering (IJCSE), Vol. 3, pp. 2792- 2798,2011.

[10] R. Agrawal, T. Imielinski, and A. Swami, "Mining association rules between sets of items in large databases," in ACM SIGMOD Record, vol. 22, pp. 207-216, 1993

[11] R. Srikant, "Fast algorithms for mining association rules and sequential patterns," UNIVERSITY OF WISCONSIN, 1996.

[12] T. C. Corporation, "Introduction to Data Mining and Knowledge Discovery", Two Crows Corporation, Book, 1999.

AUTHOR PROFILE

Sonali Sambhaji Shintre, is a M.E student of Computer Science & Engineering from T.P.C.T.'s College of Engineering, Osmanabad India. She graduated in Information Technology from BAMU University, Aurangabad. Her current research work focuses on Study of Mining and Hiding of Sensitive Association Rules.

Associate Professor Pravin P. Kalyankar, had completed his Master of Computer Science & Engineering, India with Graduate degree in Computer Engineering. Since more than a decade he has been the faculty of Computer science & technology in T.P.C.T.'s College of Engineering, Osmanabad, India where he is currently working has a Head of Department for MCA Department.