Discretion Conserving Mining of Association Rules

[1]Peng ping, [2]Qiang
[1, 2]M.Tech. Students, Dept. of Computer Science and Engineering,
Nanchang Hangkong University, china

Abstract

Association rules mining is one of the most important techniques of data mining that are used to extract the association patterns from large databases. Association rules are one of the most important assets of any organization that can be used for business development and profitability increase. Association rules contain delicate information that threatens the discretion of its publication and they should be hidden before publishing the database. The aim of beating association rules is to delete delicate association rules from the published database so that possible side effects are reduced. In this paper, we present a heuristic algorithm DCR to hide delicate association rules. In the proposed algorithm, two collecting operations are performed on the delicate association rules and finally, a bunch of smaller collections is chosen to hide. A selection of a smaller bunch of collections reduces the changes in the database and side effects. The results of performing experiments on real databases, shows the impact of the proposed algorithm on missing rules reduction.

Keywords: *Data Mining, Association Procedures, Frequent Item-sets, Isolation Conserving Data Mining, Collecting.*

I. INTRODUCTION

The massive quantity of data formed by organizations; nevertheless, maximum of these organizations are confronted with deficiency of knowledge. By expending data mining tools, concealed knowledge in the data can be removed. Nowadays, data mining has extensive requests in numerous fields such as advertising, medical analysis, and business. Removed data with data mining tools assist individuals and organizations in taking better decisions and development of business processes. Association rule mining is one of the most extensively used data mining tools which extract the dependence designs from large databases removed. Association rule presents the links among items in the database. Association rule mining contains of two stages: in the first stage, frequent item sets, by using association rule mining algorithms such as Apriority Algorithm, are extracted from the large volumes of data, then in the second stage, association rules are extracted from the set of frequent items.

For example, a rule with the support 70% shows that the customers at a supermarket to buy cheese will also buy bread. The support of a rule is intended by using the formula 1:

$$\text{Support } (X, Y) = | XUY| / |D|$$

|XUY| displays the amount of communications that contains |X| and Y and |D | is the amount of transaction in the database. The rule assurance is 100%. That resources 70% of transactions includes cheese and bread. Assured measurement is calculated as follows:

$$\text{Confidence } (X, Y) \mid XUY| /| X|$$

|X| is the numbers of connections that consist X. Association rules extracted from a database are divided into two groups of weak and strong association rules. If the confidence of an association rule is below the assurance inception, it will be named as a weak association rule, whilst the resilient association rule assurance is equal or above the confidence threshold which has been defined by the user. The strong association rules will be categorized in two types of complex and non- complex. Complex association rules contain significant information and decorations which revelation of those could jeopardize the owners of information. So, the delicate suggestion instructions must be hidden before sharing them. Beating delicate association plays an energetic significant role in defensive delicate knowledge in distribution. The aim of beating suggestion rules is to remove delicate ones in published database. There are two approaches in beating delicate association rules:

- LHS sustenance increase
- RHS sustenance decrease

Association rules beating algorithms can be separated into three highest methods border-based, exact, and empirical. In equally frame base and exact method, in order to decrease the side effects of smacking procedure, confident boundary of frequent items is transformed. Though these two methods in beating delicate suggestion items are operative, in some

spread, it does not function practically in beating some association rules.

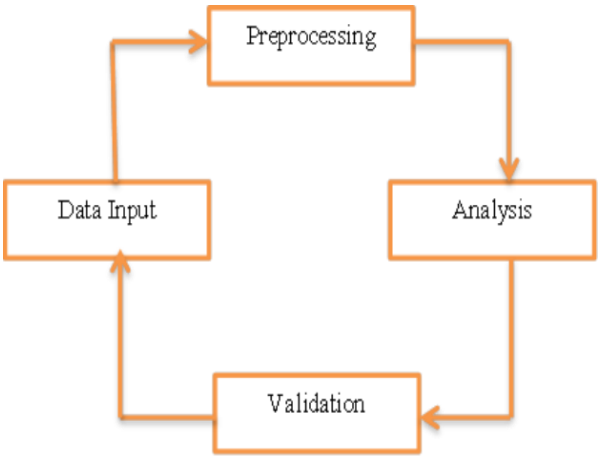

Fig 1 Data Mining Process

Information sharing is often beneficial for database owners, however, in some cases, it may disclose personal information. Discretion conserving techniques in data mining prevent unauthorized access to information. In this paper our focus is to hide delicate association rules. In this section, algorithms to hide the association rules that have been introduced in recent years will be evaluated. In the year of 2001, Saygm et al, proposed two algorithms to hide delicate association rules. The first one focuses on beating the rules by reducing the minimum support of the item-sets that generated these rules. The second one focuses on reducing the minimum confidence of the rules.

Each algorithm chooses the delicate communications to disinfect grounded on degree of battle. Naive Algorithm eliminates all items of designated transaction excluding for the article with the maximum occurrence in the database. The MinFIA algorithm chooses item with the minimum sustenance in the design as a target item and it eliminates the victim item from the penetrating communications. Unlike the MinFIA, algorithm MaxFIA chooses the item with the supreme support in the preventive pattern as a victim item. Algorithm IGA groups controlled patterns in groups of patterns distribution the similar item- sets so that all delicate decorations in the group will be hidden in one step.

Algorithm A hides association rules by increasing the sustenance of the rule's precursor until the rule sureness reductions below the minimum confidence inception.

Algorithm B hides delicate rules by lessening the occurrence of the consequential until moreover the confidence or the sustenance of the rule is below the threshold. Algorithm C reductions the support of the delicate rules until also their sureness is under the minimum confidence threshold or their sustenance is below the minimum support threshold.

1. An algorithm great number of new frequent item-sets are presented and, therefore, an accumulative number of novel rules are produced. Algorithm
2. An disturbs number of none delicate procedures in database due to elimination of substances from transaction.

The DCIS algorithm try to growth the support of left hand side of the rule and algorithms DCDS effort to reduction the support of the right hand side of the rule.

The algorithm can totally hide given subtle suggestion rule by scanning database only once, which suggestively condensed the implementation time. In this algorithm associations between the delicate connotation rules and each transaction in the original database are examined which can successfully choice the correct item to adjust.

II. RULE BEATING FOR DISCRETION CONSERVATION

The association rule beating technique is to eliminate the delicate instructions from the transactional database through suggestion rule mining. ARB method defends delicate data substances by secreting the delicate rules from miners and discloses all the non-complex rules to the miners. Data trepidation is used by Discretion Conserving Data Mining method grosses single-level trust on data miners. The method creates the vagueness concerning separate standards than the data unconfined to the third parties for data mining dedications.

In particular trust level possibility, a data proprietor generates troubled copy of its data with an amount of uncertainty. This assumption is restricted in many purposes where a data owner beliefs the data miners at numerous stages. An advanced element of Multi-Level Trust comprises new difficulties for agitation based PPDM. In inconsistency to the single-level faith condition where only one disconcerted copy is unconfined and numerous perturbed copies of the comparable data is obtainable for the data miners at numerous important stages. The additional trust in data miner occasioned in the fewer disconcerted copy access.

It also includes the access to the troubled imitations happen at inferior confidence periods.

Moreover, data miners access multiple disturbed duplicates in actions. With collection conserved done disturbed copies, the data miner on the other hand formed a detailed reinstatement of the inventive data than allowable by the data possessor. It is predictable as the collection attack. It contains the combining incidence condition where opponents join their copies to raise an attack. It also includes the situation where an opponent uses public data to perform the attack by themselves. Averting diversity attacks is the significant issue in solving the MLT-PPDM problem.

A compressed pre large GA-based algorithm is intended in to achieve beating process of the delicate item collections though removing transaction. The designed algorithm explains the problems of the evolutionary process by executing both the compact GA-based (CGA) mechanism and the pre-large thought. A suitability function that was flexible in nature was organized using three adaptable weights to recognize suitable businesses deleted to securitize the complex item sets with minimal side properties of beating failure, absent cost and simulated cost. A GA algorithm decreases the memory wants by not taking the crossover and alteration processes but simulator the concerts of traditional GAs.

A. Association Rule Beating Techniques with Minimal Side Effects

The corporate method of disinfect the database for thrashing the information that is delicate. A novel beating- missing- artificial utility (BMAU) algorithm is designed in to hide delicate item sets during transaction deletion. The transaction completes the developed ratio of delicate to non- delicate one is selected to delete. In instruction to hide thoughtful item sets, three side possessions were measured recognized as beating failures, missing article sets and artificial article sets. Data cleansing is used to hide the delicate information from disclose in PPDM. To decrease the side possessions, minimal alteration of the databases is compulsory.

The transactions with any of the delicate article set are calculated to trace the minimal HMAU standards between transactions. The transaction with minimal HMAU value is directly occupied away from the database. The procedure develops restated till all searching item sets are concealed. To avoid revealing concealed delicate article sets, the smallest amount is modernized in the deletion process.

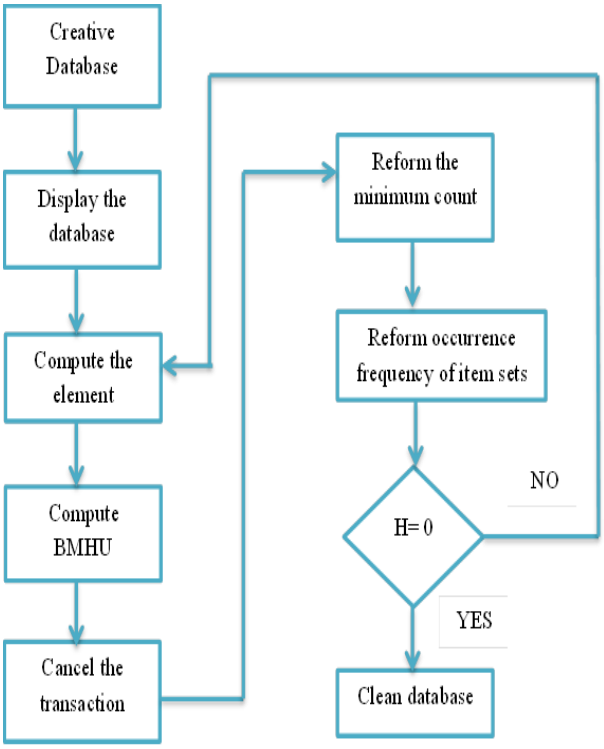

Fig 2 HMAU Algorithms

III. PROPOSED ALGORITHM

In this paper, we proposed DCR (Double Collecting Rules) to hide connotation rules. DCR use collecting to minimize side effects such as beating failure and misses cost. Collecting delicate rules and beating collections, instead of beating rules individually, reduces the changes in the database in which it minimizes the side effects. In process of collecting it should be noticed that the delicate rules structure remarkably influences the number of generated collections. For instance, consider these delicate rules as $c \rightarrow b$, $d \rightarrow a$, $c \rightarrow a$ and $b \rightarrow a$. If these rules are collected based on similar RHS, the collections will be at Table 1. So that, two collections have been generated that in fact by deleting items of "a" and "b" as enough, these four delicate rules will be hidden. Now consider $b \rightarrow a$, $b \rightarrow c$, $b \rightarrow d$ and $c \rightarrow e$. If they are collected based on similar RHS, the collections will be at Table 2.

So, four collections will be generated which it is necessary to delete "a", "c", "d", and "e" from the database as enough in order to hide these delicate rules; this is clear that collecting based on similar RHS generates four collections, while collecting based on similar LHS generated only two collections.

Collection_RHS	Rules
a	b→a, c→a, d →a
b	c→b

Table 1 Collecting based on RHS

Collection_RHS	Rules
a	b→a
c	b→c
d	b→d
e	c→e

Table 2 Collecting based on RHS

Collection_LHS	Rules
b	b→a, b→c, b→d
c	c→e

Table 3 Collecting based on LHS

In the proposed algorithm, two procedures of accumulating will be done. That resources it assortments established on equally comparable RHS and LHS and then the least collection will be designated. If the numbers of two assortments are identical, for a reduction of failures cost, assortments created on similar LHS will be designated. By execution two assembling procedures accurately, structural delicate rules properties in gathering and beating process have been reduced.

A. DCR Framework

Some significant model used in proposed algorithm is as follows:

- **Delicate item**: If there is an item in delicate rules is called a delicate item.
- **Item weight**: Number of iterations of any delicate item
- **Transaction weight**: The total weight of items in a transaction.
- **Heavy transaction**: Heavy transaction is the one that is greater than zero.
- **Light transaction**: Light transaction is the one that is equal to zero.

B. DCR Algorithm

1) Input

Original database D, Minimum Support Threshold (MST), and Minimum Confidence Threshold (MCT).

2) Output

Clean database D.

3) Prime requirements

1. Measurement of delicate items (Number of restatements of any delicate item will be calculated).
2. Heavy transaction will be arranged created on their weight in descendent order (In complaint of weight correspondence, they will be sorted created on their length in ascending order).
3. Light transaction will be particular and created on their distance they will be sorted
4. Delicate rules will be collected based on similar RHS and then the set of RHS will be produced. The set of RHS consists delicate rules' RHS.
5. Delicate rules will be collected based on similar LHS and then the set of LHS will be generated. The set of LHS consists delicate rules' LHS.
6. If RHS set is lesser than LHS set (with less numbers), RHS will be selected for beating; otherwise LHS set will be selected.
7. If RHS set has been selected for beating, delicate rules' sustenance will be reduced as follow:

7.1. Though all delicate rules are not concealed, it removes RHS item-sets from heavy transactions.

8. If LHS set requires been designated for beating, delicate rules' assurance with the aggregate LHS sustenance of delicate rules will be reduced as follows:

8.1. While all delicate rules are not hidden, it adds LHS item-sets in light dealings.

8.2. If all delicate rules are not hidden (it capacity occur that there would be inadequate light contract to add LHS item-sets) it shifts to step 4 Next collecting unhidden delicate association rules based on similar RHS it keeps on the process from step 7.

IV. PERFORMANCE ANALYSIS

In this paper, it has been trying to use DSRRC, ADSRRC, and MDSRRC to assess the proposed algorithm due to like operation in beating association rules. All four algorithms have been inspected on PC with Core i3 CPU, 4 GB Ram, and Windows 7 operating system. The selective database for testing these algorithms is Pizza and Baby Corn. Properties of two databases are as follows:

Database Name	Number of Transaction	Number of Item	Status
Pizza	3196	75	Dense
Baby Corn	8124	119	Sparse

Table 4 Database properties

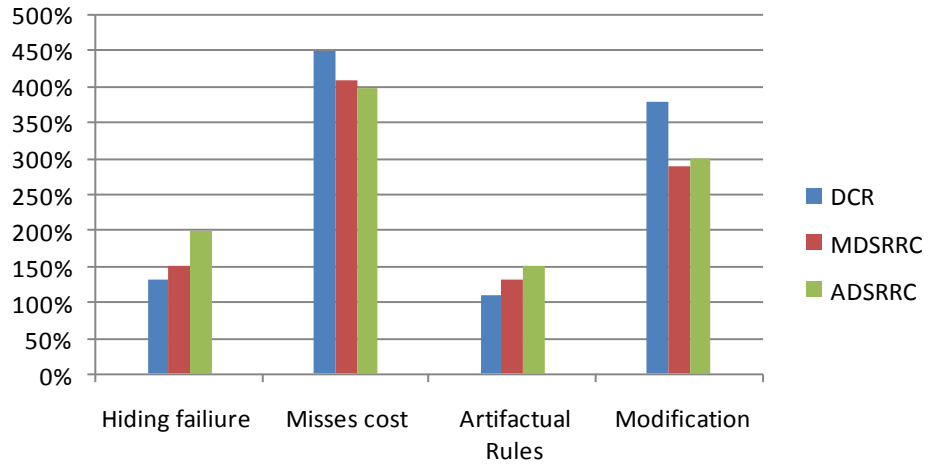

FIG 3 Examination Result of Mushroom Database

V. CONCLUSION

In this paper DCR procedure has been proposed. The aim of this algorithm is to reduce beating process side effects, especially beating failure and Misses cost. Collecting is the method that is used for beating association rules in this algorithm. Delicate rules will be hidden in collections instead of separately, which reduces misses cost. In this algorithm in order to enhance the outcome of collecting and eliminate the influences of the delicate rules on collections, two collecting procedures are performed on the delicate rules based on similar RHS and LHS. After selection of smaller collection beating procedure will be done. If collections based on similar RHS are selected, support of delicate rules by removing RHS items from heavy transactions below the threshold, will be decreased and the delicate rules will be hidden. Else collection based on the LHS is selected, Confidence of delicate rules by inserting LHS items in light transactions will be decreased below the threshold and they will be hidden. It has been tried to and due to dual collecting process and selection of smallest collection, it operates more efficiently than DSRRC, ADSRRC, and MDSRRC in reduction of misses cost and modifications in the database.

REFERENCES

[1] Yi-Hung Wu, Chia-Ming Chiang, and Arbee L.P. Chen, Senior Member, IEEE Computer Society, "Beating Sensitive Association Rules with Limited Side Effects", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 19, NO. 1, JANUARY 2007.

[2] Chun-Wei Lin, Binbin Zhang, Kuo-Tung Yang and Tzung-Pei Hong, "Efficiently Hiding Sensitive Itemsets with Transaction Deletion Based on Genetic Algorithms", Hindawi Publishing Corporation, the Scientific World Journal, Volume 2014, Article ID 398269 September 2014.

[3] FoscaGiannotti, Laks V. S. Lakshmanan, Anna Monreale, Dino Pedreschi, and Hui (Wendy) Wang, "Privacy-Preserving Mining of Association Rules From Outsourced Transaction Databases", Ieee Systems Journal, Vol. 7, No. 3, September 2013.

[4] TamirTassa, "Secure Mining of Association Rules in Horizontally Distributed Databases", Ieee Transactions On Knowledge And Data Engineering, Vol. 26, No. 4, April 2014.

[5] Yaping Li, Minghua Chen, Qiwei Li, and Wei Zhang, "Enabling Multilevel Trust in Privacy Preserving Data Mining", Ieee Transactions On Knowledge And Data Engineering, Vol. 24, No. 9, September 2012.

[6] Chun-Wei Lin, Tzung-Pei Hong and Hung-Chuan Hsu, " Reducing Side Effects of Hiding Sensitive Itemsets in Privacy Preserving Data Mining", Hindawi Publishing Corporation, the Scientific World Journal, Volume 2014, Article ID 235837 April 2014.

[7] VineetRichhariya., and PrateekChourey., "A Robust Technique for Privacy Preservation of Outsourced Transaction Database" International Journal of Research in Engineering & Technology (IJRET), Vol. 2, Issue 6, Jun 2014, 51-58.

[8] Gabriel Ghinita, Member, IEEE, PanosKalnis, and Yufei Tao, "Anonymous Publication of Sensitive Transactional Data", Ieee Transactions On Knowledge And Data Engineering, Vol. 23, No. 2, February 2011.

[9] Sara Hajian and Josep Domingo-Ferrer, Fellow, IEEE, "A Methodology for Direct and Indirect Discrimination Prevention in Data Mining", Ieee Transactions On Knowledge And Data Engineering, Vol. 25, No. 7, July 2013.

[10] DimitriosKarapiperis and Vassilios S. Verykios, Member, IEEE, "An LSH-Based Blocking Approach with a Homomorphic Matching Technique for Privacy-Preserving Record Linkage", Ieee Transactions On Knowledge And Data Engineering, Vol. 27, No. 4, April 2015.