International Journal of Computer & Information Technologies (IJOCIT)
www.ijocit.ir & www.ijocit.org

ISSN = 2345-3877

A Decreasing Max-Min Approach for Hiding Informative Association Rules

Zohreh Rostamkhani[a*], M. T. Taghavifard[1]
[1] *Faculty of Management, Mehr Alborz university , Tehran, Iran*
[2] *Associate Professor, Faculty of Management, Allameh Tabatabai University, Tehran*
Email: Z_rostamkhany@yahoo.com

Abstract: There are some pieces of data containing informative information should be protected against unauthorized access. This protection confidence is a purpose for the database researchers as well as relative agencies. Recent technology advances in data mining is raised the leakage risks that one may encounter when sharing data to collaboration. An issue which is still not concentered enough is the need to balance the reliability of the disclosed data with the requirements of the data right users. Every restriction approaches affect, in some directions, right data reliabilities. In this paper, we present confidence issues of rules, the association rules mining. Accordingly, we present an approach for hiding a set of ARs, which is detected as informative by database administrators. One rule has been called as informative if its leakage risk is above a certain analyzer threshold. In some cases, informative rules must not be disclosed to the unauthorized corporations, since they are referring informative data which their disclosures may be utilized by company competitor's analyzers. We also evaluate the hiding process with a similar one in order to analyze their performance.

Keywords: Informative rule, privacy preserving, association rule hiding, big data.

1. Introduction

Many corporations as well as organizations in order to backing their short and long-term planning activities are searching for a way to collect, store, analyze, and report data about their conditions. Databases, therefore, contain confidential information, such as social security numbers, income, credit ratings, type of disease, customer purchases which should be protected in a right way.

Securing against unauthorized accesses is a long-term aim of the database security research group of companies. Solutions to these issues require combining several methods and approaches [1-3]. In an environment where data have a lot of informative levels, this data may be categorized at various levels and made it accessible just to subjects with an appropriate clearance. It is, however, well known that simply limiting access to informative data does not grantee informative data protection, completely.

For instance, informative, or in other words “high risk”, data may be mined from non-informative, or “low risk” data through some mining processes based on some inferences of the application the user has. Such an issue, known as the “mining issue” has been broadly investigated and available solutions have been detected. The proposed solutions address the problem of how to protect disclosure of informative data through the combination of mined rules with non-informative data as figure1 Example of mined rules is deductive rules, functional dependencies, or material implications [3]. Recent strategies in Data Mining (DM) approaches and related applications have, however, increased the security issues which one may confront companies’ database when sharing data. The inference of information that can be achieved by such approaches has been the focus of the Knowledge Discovery in Databases (KDD) researchers’ main goal for years and also by now it is a well understood issue [4].

Apart from the point, not until very recently, the impact on the data confidence originating by

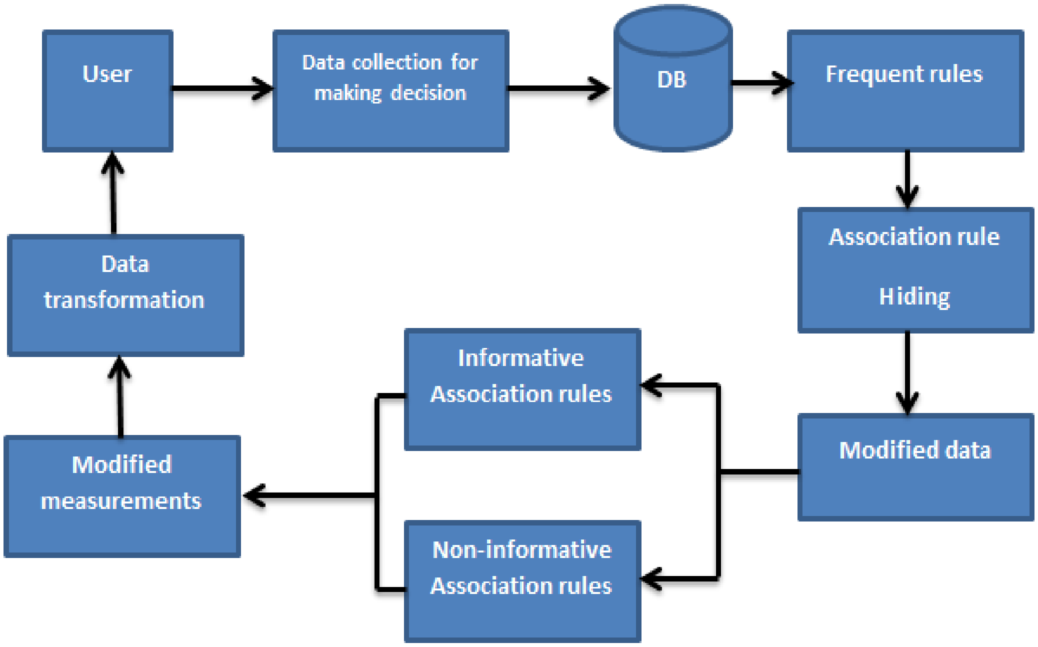
User
Data collection for making decision
DB
Frequent rules
Association rule Hiding
Modified data
Informative Association rules
Non-informative Association rules
Modified measurements
Data transformation

Figure 1. Main approach of association rule hiding [4].

These techniques has not been focused. The process of extracting hidden patterns from large databases was first showed as a threat to databases privacy by O' Leary [5].

Piatetsky-Shapiro, of GTE Laboratories, was the chair of a mini-symposium on knowledge discovery in databases and privacy, organized around the issues raised in O' Leary's scientific paper in 1991. The point mentioned by the panel was the limitation of disclosure of private data in many areas such as medical and socioeconomic fields; the aim is not to detect patterns in special cases but patterns about groups.

The detection risk in the confidence of informative data, that is not restricted to specific pattern, is another form of threat which is showed in a paper by Clifton, Mitre Corporation and Marks from Defense Department [6].

The authors demonstrated a scenario of how various data mining approaches can be applied in a business to catch at a high profit that we bring it below.

Suppose to have a deal with Dedtrees Paper Corporation, as purchasing provider of BigMart, a hypermarket. Dedtree proposes its products with a low price, in other hand if BigMart accept to give them access to its private relational database. In this case, BigMart allows the Dedtrees to begin mining data. By applying an association rule mining technique, Dedtree detects that people who purchase skim milk also purchase Green paper.

Dedtrees now runs a marketing strategy saying that "you can get 50 cents off skim milk with every purchase of a Dedtrees product." This strategy heavily reduces the volume of Green paper sales. Accordingly this strategy increases the prices to BigmMart, because of the lower volume of sales. Consequently, BigMart observes reduced competition; All in all, BigMart begins to lose their business rather to competitors, who are able to have a better negotiate with Green Paper Company.

Presented strategy of this paper, includes the need to avoid disclosure not only of confidential personal information from database, but also to prevent data mining strategies from detecting informative information which is not even mined to the database owners.

Presented paper proposes a new approach and a well-designed algorithm for hiding informative information from database. The hiding approach that this paper proposes is based on reducing confidence of rules that indicate how significant they are. By removing specific items from some transactions, they are modified on the hiding strategy.

Item Selection process in a rule to be hidden and the transaction selection that will be modified is a vital factor for catching the lowest information loss.

The rest of this paper is organized as follows: In Section 2, we present an overview of the current approaches to the problem of data mining and privacy preserving. Section 3 gives problem formalization, while proposed solution is presented in Section 4. Section 5 discusses results obtained from the algorithm performance. Concluding words is presented in Section 6.

2. Related works

The privacy aspect of data mining is mentioned in [6] and some probable strategies to the issue of detection of informative information in a data mining context are proposed. The suggested techniques contain fuzzifying and augmenting the source database and also limiting the access to the original database by releasing only samples of the source data. Clifton [7] used the last strategy as he studies the relation between the amount of released information and the impact of the patterns that are detected. Clifton also mentions how to discover the amount of data which data mining techniques could not extract informative data.

Clifton and Marks also mention different data mining techniques to increase the performance of any strategy which have the aim of leakage limitation of informative confidential information.

The strategy presented by Clifton in [7] is differ from any specific data mining strategy; similar papers [8], [9] propose techniques that prevent leakage of informative data for especial data mining strategies such as association rule mining classification methods.

Classification mining strategies apply informative information to rank goals; every group of goals have explanations mentioned by non-informative features. For Decision- Region-based techniques, the description space generated by each value of the sensitive attribute can be determined a priori. The paper presented in [8] proposes two aspects that can be applied to assess the classification inference system results and then authors apply these aspects in the Decision-Region based techniques, to alter the explanation of an informative goal, accordingly database owners can be sure that goals are not informative.

Agrawal and Srikant in [10] applied data modification strategies to change confidence of data values in the directions of data mining which they can be achieved from the altered database [10]. Authors used applications where the individual data values are confidential rather than the data mining results and concentrated on a specific data mining model, namely, the classification by decision trees. Agrawal et al. improve the data modifications techniques by applying expectation maximization for reconstructing the source data modification which is mainly applied to create the classification model [11].

Leakage restriction of informative information by data mining techniques, on the basis of recovering of mined association rules, has also been recently indicated [9]. They are proposed to avoid leakage of informative information by decreasing the confidence of the association rules utilizing heuristic methods that can be thought of as the antecedent of the heuristic model that we proposed in this research.

3. Problem formation

Mining of association rules contain two important aspects. First of all, it produces frequent item with considering a threshold for minimum support. In the hand, it develops association with a threshold of upper than minimum confidence by using first step item sets. For generating ARs, there are many directions like FP-Growth, FP-Tree, Eclat and Apriori.

Association rule mining steps mentioned before can be formulated as follows:

Given I= {i1, i2, i3, …, in} as a set of n items and II={i1i2, i1i3, i1i4, …, imin} called an item set, which X is a sub set of II. Relational database D contains D={m1, m2, m3, …,mn} which m is a transaction of transactional database. Also, X as an item set is supported by a transaction if transaction includes item set X. Count of items in each transactional data base is calculated as follows:

Support(X) =count(X)/n. (1)

Which n is the total number of items exists in database. Each database rules may be calculated by the following formula:

Confidence(X,Y)=Support(X,Y)/Support(X). (2)

Both should have minimum thresholds to control the number of item sets and consequently number of association rules mined from source database. Thresholds are user defined and can be varied user by user.

4. Proposed hiding strategy

The item sets meets the requirements of the proposed algorithm by choosing transactions that includes both the item sets that exist on either left hand side of the rule or right hand side of the informative rule.

According to the aim of informative association rule hiding algorithm, it hides informative rules defined by user only by decreasing the confidence of the right hand side of the rule up until informative rule confidence checker become minimum confidence threshold below.

The Max-Min approach is proposed if there exists more than one item on right hand side of the informative rule. Accordingly, the first step is sorting rule's consequent item sets by calculating support of each one. Choose the smallest support as the result of step one.

A step forward makes a selection from a range of candidate item sets having great value rather than the others. This step returns a proper item among informative rule's consequent. The support of selected item is decreasing through transactions that have been stood as a candidate before by removing its value.

As figure 2 shows the proposed algorithm, it begins by sorting transactions that fully support the informative rule on the basis of their backing items in increasing sorting model. As the result of Max-Min proposed strategy, the item that has the minimum impact on the database is selected and also is deleted form a number of transactions that contains informative rule item sets.

The support and consequently the confidence of the selected rules are calculated again as well as informative rule. As the last step, algorithm confidence checker checks major impact of informative rule. Accordingly, another informative rule is selected to modify its significance if its confidence be under user defined threshold and meet the user requirements.

The selection process among informative rules is on the basis of their support. Since the informative rules that have lower support are sensitive moderate other informative rules, the process selects them as hiding candidates. On the

```
Input:
(1) a source database D,
(2) a min_support,
(3) a min_confidence,
(4) a set of predicting items X
Output: a transformed database D0, where rules containing X on LHS will be hidden
Find large 1-item sets from D;
For each predicting item x member of X
    If x is not a large 1-itemset, then X = X - {x};
    If X is empty, and then EXIT;
    //no rule contains X in LHS
Find large 2-itemsets from D;
For each x member of X {
    For each large 2-itemset containing x {
        Compute confidence of rule U, where U is a rule like x → y;
        If confidence (U) < min_confidence, then
            Go to next large 2-itemset;
        Else {//Decrease Support of RHS
            Find TR = {t in D | fully support U};
            Sort TR in ascending order by the number of items;
            While {confidence (U) P min_confidence and TR is not empty}
                Select the minimum support item set From RHS;
                Select the item from item set That has less effect on others.
                Choose the first transaction t from TR;
                Modify t so that y is not supported;
                Compute support and confidence of U;
                Remove and save the first transaction t from TR;
            }; // end While
        }; // end if confidence (U) < min_confidence
        If TR is empty, then {
            Cannot hide x →y;
            Restore D;
            Go to next large-2 item set;
        } // end if TR is empty
    } // end of for each large 2-itemset
    Remove x from X;
} // end of for each x member of X
Output updated D, as the transformed D0;
```

Figure 2. Proposed hiding algorithm [10].

other hand, border rules should be considered not to be removed during hiding process. In this case, the rules that not only have the lowest support but even have a minimum side effect of bordering rules are focused of algorithm attention.

5. Results

The proposed algorithm has been run on a PC with Intel 2270 MHz cori3 processors and 2GB RAM running on Windows 7 operating system.

5.1. Data set

The transactional dataset is a market basket database in order to frequent item set mining (retail.dat) at http://fimi.ua.ac.be/data/. it includes the retail market basket data from an Belgian retail. The dataset was collected over three periods from the middle of December 1999 to the end of November 2000. The dataset contains 88,162 transactions and 16,469 product IDs where the first column of dataset is the transaction identification. Each transaction contains the coded items of retailer which were sold to a customer. The items are separated by a space in each transaction.

5.2. Experimental results

The proposed algorithm deals with two criteria of hiding process. The first one is side effects that occur during performing hiding approach, and another one is time requirement to complete the process. Side effect includes "lost rule", "new rule" and "hiding failure". The number of non-informative rules, which are not found after algorithm execution, called lost rules. If the algorithms produce new non-informative rules that cannot find in original database, the algorithm produced new association rules as side effect. Algorithm has hiding failure if there is not enough transaction to complete the hiding process.

The proposed algorithm results is compared with hiding approach presented by Verykios et al. [12] to measure its reliability. The performance of these strategies is illustrated in the following sentences.

Figure 3 presents the performance of the strategies in the lost rules decreasing as side effect. In this way, the proposed strategy has better results in decreasing lost rules compared to its similar approach. As the figure shows, the total number of lost rules is increasing by raising repetition of algorithm. Consequently, proposed technique causes fewer lost rules rather than similar. On the other hand, Verikios algorithm increases number of lost rules with facing above three informative rules.

Figure 4 shows that only a few new rules generate both strategies. What's important is proposed technique runs hiding process without any new rules, on the other hand, similar algorithm produces approximately 0.5% new rules. By the way, the percentage of new rules generated by both techniques is in very low range and generally, new rules generated by these algorithms are completely same in high repetition.

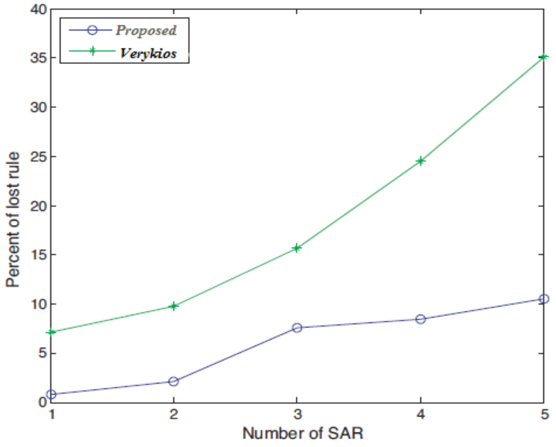

Figure 3. Lost rules percentage.

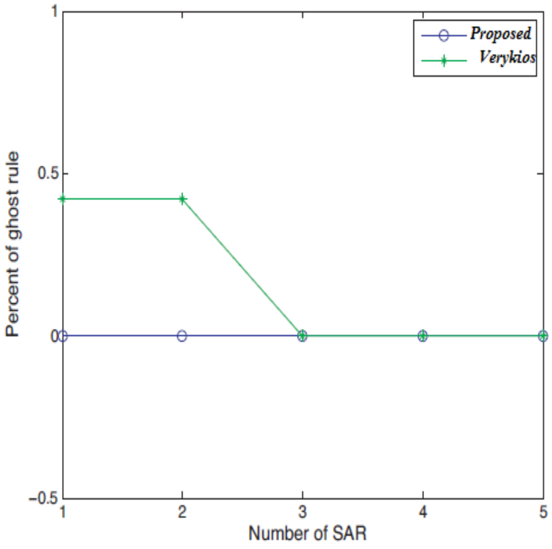

Figure 4. New rules percentage.

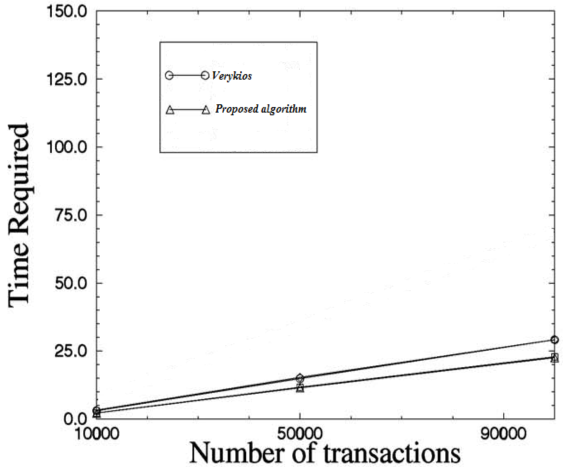

Figure 5. Time requirement.

Both proposed technique and Verykios do not have hiding failure during hiding process. In other definition, both algorithms catch the same reliability in hiding informative rules. Also, as figure 5 shows both hiding algorithms have approximately same execution time.

6. Conclusions

In this paper, we presented an improved approach in order to protect informative association rule with hiding technique.

According to the proposed algorithm, it reduces the importance of the association rules by decreasing their item support. Also, a Max-Min approach utilized to select the item set having low side effect in informative rule item set selection process. The process iterates up until the user defined confidence threshold meets the requirements of informative rule modified confidence.

Authors also prepared criteria to analyze the performance of the proposed technique with its similar. Although both techniques lost rules are rising by increasing the number of informative rules, the proposed algorithm has low pace compared to its similar, so its indicates that most of the non-informative rules are under control by applying proposed algorithm.

References

1. M. Dunne, R. Lusch, "Retailing strategy for hiding rules", (2005), Thompson South-Western, Mason, United State, pp. 166-173.
2. D. Corsten, N. Kumar, (2012), "Do suppliers benefit from collaborative relationships with large retailers, An empirical investigation of efficient consumer response adoption", Journal of Marketing Vol. 69, pp. 80–94.
3. Y. Yin, I. Kaku, J. Tang, J.M. Zhu, (2011), "Association Rules Mining in Inventory Database, Data Mining Concepts, Method and Application in Management and Engineering Designs", Springer, London, England, Vol. 3, pp. 9–23.
4. N. Kumar, A. Gangopadhyay, G. Karabatis, (2007), "Supporting mobile decision making with association rules and multi-layered caching", Decision Support Systems Vol. 43 pp. 16–30.
5. S. Anand, N. Viswanathan, 2009 The 21st Century Retail Supply Chain: Three Key Imperatives for Retailers, Aberdeen Group, Boston, Massachusetts, United State, pp. 1–28.
6. Y.J. Chen, (2010), "Knowledge integration and sharing for collaborative molding product design and process development", Computers in Industry, Vol. 61, pp. 659–675.
7. D.Y. Zhang, Y. Zeng, L. Wang, H. Li, Y. Geng, (2011), "Modeling and evaluating information leakage caused by inferences in supply chains", Computers in Industry, Vol. 62, pp. 351–363.
8. F. Biennier, J. Favrel, (2005), "Collaborative business and data privacy: toward a cybercontrol", Computers in Industry Vol. 56 pp. 361–370.
9. T.Y. Chen, (2008), "Knowledge sharing in virtual enterprises via an ontology-based access control approach", Computers in Industry, Vol. 59, pp. 502–519.
10. Y. Zeng, L. Wang, X. Deng, X. Cao, N. Khundker, (2012), "Secure collaboration in global design and supply chain environment: problem analysis and literature review", Computers in Industry, Vol. 63, pp. 545–556.
11. R. Gouriveau, D. Noyes, (2004), "Risk management dependability tools and case-based reasoning integration using the object

formalism", Computers in Industry, Vol. 55, pp. 255–267.

12. V. S. Verykios, E. Bertinio, Y. Saygin, E.Dasseni, (2004), "Association rules hiding", IEEE transactions on knowledge and data engineering , Vol. 16, pp. 434-447.