ISSN: 2321-7782 (Online)
Volume 3, Issue 7, July 2015
International Journal of Advance Research in Computer Science and Management Studies
Research Article / Survey Paper / Case Study
Available online at: www.ijarcsms.com

Maintaining Data Privacy in Association Rule Mining

M. Suryapraba[1]
Research Scholar,
Department of Computer Science,
H.H. The Rajahs College (Autonomous), India

K. Karpagam[2]
Assistant Professor,
Department of Computer Science,
H.H. The Rajahs College (Autonomous), India

Abstract: ***Association rule mining is an important data mining technique that finds inter association among a large set of data items. In proposed work, Dual Clustering Rule (DCR) algorithm uses clustering to minimize side effects such as hiding failure using and misses cost. In this paper, we perform experiments on real databases that show the impact of the proposed algorithm on missing rules reduction.***

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

I. INTRODUCTION

Data mining is the exploration and analysis of large quantities of data in order to discover valid, novel, potentially useful, and ultimately understandable patterns in data. It is used for real-time enterprise, systemic learning, security, and homeland defense. Data mining is a well-known technique for automatically and intelligently extracting information or knowledge from a large amount of data. It may also disclose some of the sensitive information about individuals compromising the individual's right to privacy. Association rule mining is a kind of technique in data mining.

Privacy preserving has becoming an increasingly important field of research. It is one of the important challenges in data mining. For example in medical database it is useful to share information about diseases but at the same time it is required to preserve patient's identity. The main aim of all association rules hiding algorithm is to minimally modify the original database and see that no sensitive association rule is derived from it.

In this paper, in section II we discuss association rule hiding related works in detail. Section III, presents the proposed algorithm for sensitive rule hiding. In section 4, Experimental results of proposed algorithm are shown.

II. RELATED WORK

[Assaf Schuster et al] The algorithm is asynchronous, involves no global communication patterns, and dynamically adjusts to changes in the data or to the failure and recovery of resources. To the best of our knowledge, this is the first privacy-preserving mining algorithm to possess these features. Simulations of thousands of resources prove that our algorithm quickly converges to the correct result while using reasonable communication. The simulations also prove that the effect of the privacy parameter on both the convergence time and the number of messages is logarithmic [1].

[Alexandre Evfimievski et al] derive formulae for an unbiased support estimator and its variance, which allows recovering item set supports from randomized datasets, and show how to incorporate these formulae into mining algorithms. Finally, we present experimental results that validate the algorithm by applying it on real datasets. We have presented three key contributions towards mining association rules while preserving privacy. First, we pointed out the problem of privacy breaches, presented their formal definitions and proposed a natural solution. Second, we gave a sound mathematical treatment for a class of randomization algorithms and derived formulae for support and variance prediction, and showed how to incorporate these formulae into mining algorithms. Finally, we presented experimental results that validated the algorithm in practice by applying it to two real datasets from different domains. [2]

III. PROPOSED WORK

The expression data mining indicates a wide range of tools and techniques to exact useful information which can be sensitive form a large collection of data. Data should be manipulated or distorted in such a way that information cannot be discovered through data mining techniques.

Association rule is said to be "interesting" if it is support and confidence are greater than user defined thresholds sup_{min} and $conf_{min}$, respectively and the objective of the mining process is to find all such interesting rules Therefore, the mining objective is, in essence, to efficiently discover all frequent that are present in databases. The proposed approach is based on modifying the database transaction.

3.1 Apriori algorithm

The Apriori algorithm is an influential algorithm for mining frequent item set for Boolean association rules. Apriori is designed to operate on database containing transaction. Let B= {b1,b2…….bn} be a set of items. Let D be a set of transaction or database. Each transaction t€D is an item set such that is a proper subset of B. An association rule is an implication of the form c→e, so that c €B, e€B and c∩e=□ c and d is asset of items called item set. The support of the rule c→e is the percentage of transactions in T that contain c∩e. It determines how frequent the rule is applicable to the transaction set T. The support of the rule is represented by the formula

$$Support = \frac{C \cup E}{D}$$

The confidence of a rule describes the percentage of transactions containing c which also contain d. It is given by

$$Confidence = \frac{C \cup E}{C}$$

3.2 Algorithm used: Dual Clustering

There are two domains in dual clustering. One domain refers to the optimization domain and the other refers to the constraint domain. Attribute of the optimization domain are those involved in the optimization of the objective function, while those on the constrain domain specify the application dependent constraints.

DCR algorithm

Input: Source database E, minimum support threshold (MST), and minimum confidence threshold (MCT).

Output: The sanitized database D'

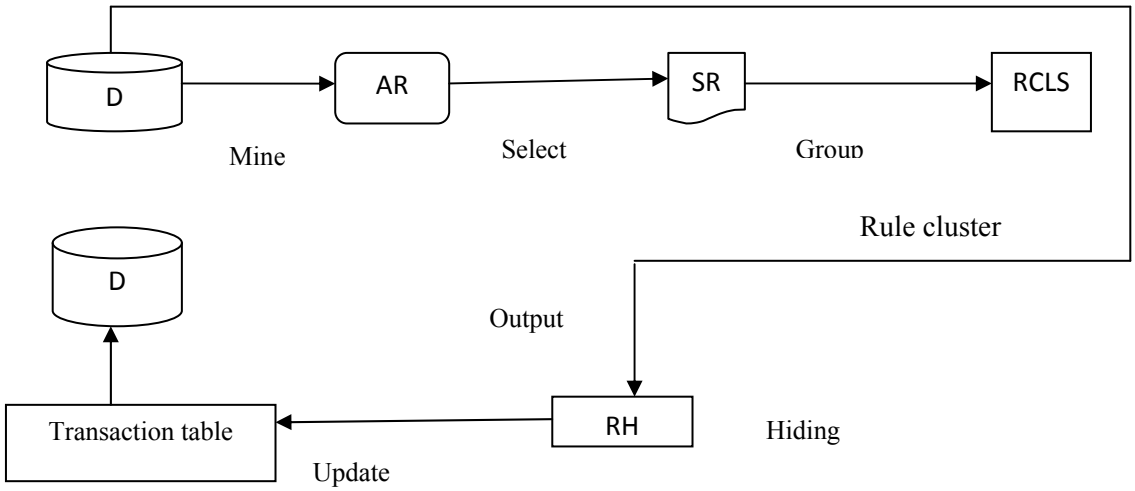
1. Begin
2. Generate association rules.
3. Selecting the sensitive rule set RH with single antecedent and consequent eg. c→e.
4. Clustering-based on common item in R.H.S of the selected rules.
5. Find sensitivity of each item in each cluster.
6. Find the sensitivity of each rule in each cluster.
7. Find the sensitivity of each cluster.
8. Index the sensitive transactions for each cluster.
9. Sort generated clusters in decreasing order of their sensitivity.
10. For the first cluster, sort selected transaction in decreasing order of their sensitivity.

11. For each cluster c€ C.
12. {
13. While(all the sensitive rules € c are not hidden)
14. {
15. Take first transaction for cluster c.
16. Delete common R.H.S item from the transaction.
17. Update the sensitivity of deleted item for modified transaction in other cluster and sort it.
18. For i=1 to no. of rule R h € c.
19. {
20. Update support and confidence of the rule r € c.
21. if (support of r<MST or confidence of r<MCT).
22. {
23. Remove Rule r from R h
24. }
25. }
26. Take next transaction.
27. }
28. End while
29. }
30. End for
31. Update the modified transactions in D.
32. End.

3.3 Randomization

Architecture

Transactions



IV. EXPERIMENTAL RESULTS

In the previous section, we have theoretically proved that increasing the randomization factors for non-sensitive items can lead to the reduction of variance for support estimates. Intuitively, the smaller the variance is, the better the mining result will be. In this section, we conduct experiments on traffic accident dataset using our Dual Clustering algorithm to verify this idea. This data set of traffic accidents is obtained from the National Institute of Statistics (NIS) for the region of Flanders (Belgium) for the period 1991-2000. More specifically, the data are obtained from the Belgian "Analysis Form for Traffic Accidents" that should be filled out by a police officer for each traffic accident that occurs with injured or deadly wounded casualties on a public road in Belgium. In total, 340.184 traffic accident records are included in the data set. The traffic accident data contain a rich source of information on the different circumstances in which the accidents have occurred: course of the accident (type of collision, road users, injuries …), traffic conditions (maximum speed, priority regulation …), environmental conditions (weather, light conditions, time of the accident …), road conditions (road surface, obstacles …), human conditions (fatigue, alcohol …) and geographical conditions (location, physical characteristics …). In total, 572 different attribute values are represented in the data set. On average, 45 attributes are filled out for each accident in the data set.

Performance Metrices

Performance of any privacy preserving association rule mining is estimated using the Hiding Failure (HF), Misses Cost and Artifactual patterns.

Hiding Failure (HF)

It is the measure of restrictive association rules that appear in the sanitized database. It is the percentage of data that remain exposed in the sanitized dataset. It is calculated by using the formula below:

$$H_F = \frac{|S_R(D')|}{|S_R(D)|}$$

Where D is the original data set, D' is the sanitized data set, SR is the number of sensitive association rules.

Misses Cost

It is the measure of amount of legitimate association rules that are hidden by accident after sanitization. It is the percentage of non-sensitive data hidden during sanitization process. It is calculated as follows:

$$M_c = \frac{|S_R(D')| - |S'_R(D')|}{|S'_R(D)|}$$

where $|S'_R(D)|$ is the size of set of all non-sensitive rules.

Artifactual Patterns

It is the measure of artificial association rules created by adding the noise in the data. It is the measure of discovered artifacts. It is calculated by:

$$DIFF(D, D') = \frac{1}{\sum_{i=0}^{n} F_D(i)} \times \sum_{i=0}^{n} \left| F_D(i) - F_{D'}(i) \right|$$

where P is the set of discovered association rules in the original database D and P' is the set of association rules in the sanitized database D'

Frequent item sets and association rule mining

Frequent item set generation algorithm digs out frequently occurring item sets, subsequences, or substructures from large data sets. A common example of frequent item set applications is market basket analysis. Market basket analysis is a process that helps retailers to develop their marketing strategies by finding out associations between different items that customers place in their shopping baskets. Besides market basket data, frequent item sets mining has been applied in, for example, bio informatics and web mining.

Accident	Gender	Age	Alcohol	Speed Limit	Fatals
1	M	Young	Yes	>= 100	Yes
2	M	Young	Yes	70-90	Yes
3	M	Middle	No	70-90	Yes
4	F	Young	No	<=60	Yes
5	M	Old	No	70-90	Yes

Table 1: Sample Dataset for Accident data

Mining of frequent item sets using an artificial accident data given in Table 1 will be illustrated next. The data set contains five fictitious accidents that are described with four explanatory variables and one consequential variable. The frequent item sets are generated according to apriori algorithm.

Rule	Support	Confidence
1=>4 6	0.5	0.625
6=>1 4	0.5	0.83333
4 6=>1	0.5	0.83333
1 6=>4	0.5	1
1 4=>6	0.5	0.625
2=>4 5	0.4	0.63158
2 5=>4	0.4	0.8
2 4=>5	0.4	0.63158
1=>4 7	0.53333	0.66667
7=>1 4	0.53333	0.84211
4 7=>1	0.53333	0.84211
1 7=>4	0.53333	1
1 4=>7	0.53333	0.66667
5=>4 6	0.5	0.6
6=>4 5	0.5	0.83333
5 6=>4	0.5	1
4 6=>5	0.5	0.83333
4 5=>6	0.5	0.6

Table 2: Predicted Values using Apriori Algorithm

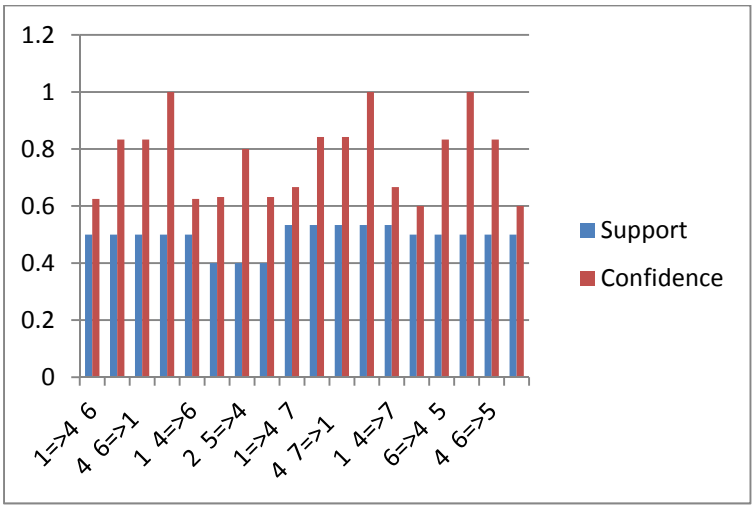


Figure 1: Graphical representation of Predicted Values

Rules	Hiding Failures
1=>4 5	0.4042
1=>4 6	0.40625
1=>4 7	0.4058
2=>4 5	0.4087
5=>4 6	0.4165

Table 2: Hiding Failures for Corresponding Rules

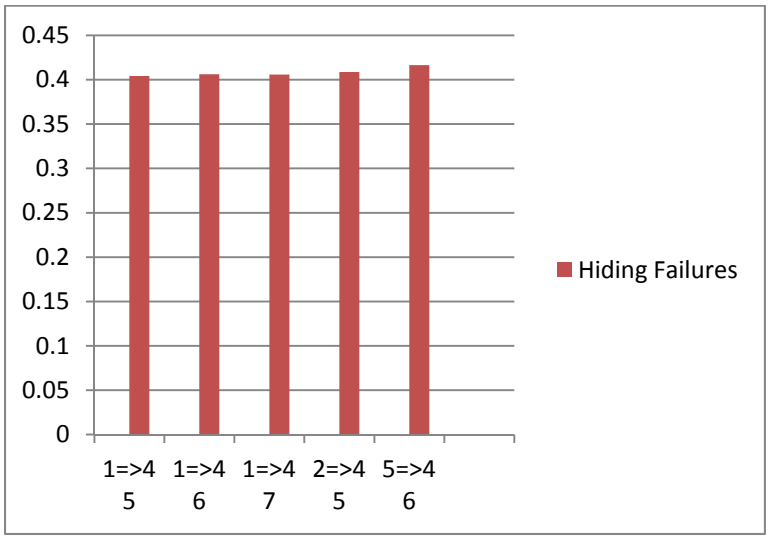


Figure 2: Privacy Preserving predicted values

The example has been executed using accident data in california dataset. The experimental results were examined with various support, confidence and hiding failures.The results shows that our proposed model exhibts only less than one percentage of hiding failure. So, after the analysis of various previous methods, the proposed method of dual clustering is more suitable for privacy preserving of data. It can be concluded that,the DCR algorithm in accident databases has no hiding failures.

V. Conclusion and Future Work

In this paper, we proposed a heuristic algorithm named DCR which hides many sensitive association rules. Several existing approaches regarding sensitive rule hiding problem were also discussed. An example was also demonstrated to show the effectiveness of the proposed DCR algorithm. This algorithm aimed at reducing the hiding process side effects, especially hiding failure and misses cost. In future, proposed algorithm can be modified to hide sensitive rules which contain different number of R.H.S items.

References

1. "Privacy Preserving Association Rule Mining in Large-Scaled Distributed Systems" by Assaf Schuster, Ran Wolff, Bobi Gilburd.
2. "Privacy Preserving Mining of Association Rules" by Alexandare Evfimievski Ramakrishnan Srikant Rakesh Agrawal Johannes Gehrke.
3. "A Survey on Privacy Preserving Association Rule Mining " by K. Sathiyapriya and Dr. G. Sudhasadasivam.
4. "Maintaining Privacy and Data Quality in Privacy Preserving Association Rule mining" by Chirag N.Modi,Dhiren R.Patel.
5. "Privacy-Preserving Distributed Mining of Association Rules on Horizontally Partitioned Data" by Murat Kantarcioglu and Chris Clifton,Senior Member,IEEE.
6. "A Study on Association Rule Hiding Approaches" by K.Shah,A.Thakkar, and A.Ganatra.
7. "Association Rule Hiding by Heuristic Approach to Reduce Side Effects and Hide Mutiple R.H.S Items" by K.Shah,A.Thakkar,and A.Ganatra.
8. "Security Information Hiding in Data Mining on the bases of Privacy Preserving Technique" by V.Yadav and R.Jindal.
9. "A Novel Algorithm for Completely Hiding Sensitive Association Rules" by C.C Weng, S.T.Chen, and H.C Lo.
10. "Hiding Collaborative recommendation association rules" by S.Wang, D.Patel,A.Jafari, and T-P.Hong.

