Privacy Preserving Data Mining based on Ant Colony Optimization

S.Narmadha
Assistant Professor
Department of Computer science
Vivekanandha College of Arts and Sciences for Women
Tiruchengode,
Tamilnadu, India

ABSTRACT

The security of the large database that contains certain decisive information, it will become a serious issue when sharing database to the network against unauthorized access. Specifically consider a scenario in which two parties owning confidential databases wish to run a data mining algorithm on the merger of their databases, without revealing any unnecessary information. Our work is motivated by the need to both protect privileged information and enable its use for business or other purposes. Privacy preserving data mining technique is a new research area in data mining and statistical databases where mining algorithms are analyzed for the side effect they acquire in data privacy. Association rule hiding is a worthwhile technique for protecting confidential data and crucial information in a database. Data modification techniques, query auditing methods, statistical techniques are developed and used for protecting the database. Many optimization techniques also used with the data mining concept for protecting the data base. In this paper using ant colony optimization technique with association rule mining for hiding sensitive items in large data base.

General Terms

Optimization technique, privacy of database

Keywords

Data Mining, Privacy, Rule hiding, Ant colony algorithm.

1. INTRODUCTION

Data mining techniques have been developed successfully to extracts knowledge in order to support a variety of fields like marketing, weather forecasting, medical diagnosis and national security. Privacy provides autonomy from unauthorized access. Providing security to sensitive data in opposition to unauthorized access has been a long term goal for the database security, research community and for the government statistical agencies. In recent years, with the swift development in Internet, data storage and data processing technologies, privacy preserving data mining has been drawn increasing consideration. Privacy problems associated to the application of data mining techniques are divided into two broad kinds, data hiding and knowledge hiding [1]. Data hiding is the removal of confidential or sensitive information from the data before it is disclosed to others. Knowledge hiding is the results of data mining techniques, after having analyzed the data, these may find out the hidden knowledge. Such knowledge should be protected from others.

A primary necessity of privacy-preserving data mining is to defend the input data, yet still consent to data miners to extort useful knowledge models. Generally modification or sanitization techniques can be categorized into two groups: data blocking and data distortion approaches. The major concept of blocking approaches, are replacing the actual values of the items with "unknown" symbols in the proper transactions [12]. The main reason of using blocking techniques is that algorithms do not add simulated information in the database. This is so important when the source database contains crucial information that extracting wrong known will consequences dangerous effects. One way of using either 1’s or 0’s in order to achieve the best possible results.

The method developed in this paper uses twofold transactional dataset as an input and modifies the original dataset based on the concept of ant colony optimization algorithms in such a way that all of sensitive rules become hiding without loss of data. The most possible style for transaction modification is distortion of original database (i.e., by replacing 1’s by 0’s and vice versa).The rest of the paper is organized as follows. In section 2 presents related work of association rule hiding. In section 3 presents an overview of association rule hiding technique. In section 3 have proposed work. In section 4 presents experimental results. In section 4 has conclude the result with performance.

2. RELATED WORK

Privacy preserving data mining technique is a new research area in data mining and statistical databases where mining algorithms are analyzed for the side effect they attain in data privacy. The objective of privacy preserving data mining is modifying the original data in some way, so that the private data and private knowledge stay private even after the mining process. Knowledge hiding is related with the sanitization of secret knowledge from the data. The knowledge hiding is also called association rule hiding. The objective of association rule hiding is to protect sensitive knowledge. The hiding scenario is the sanitization process can accomplished in the original dataset that affects minimum and preserves the general forms that achieves to hide the sensitive knowledge. The Association Rule Hiding Techniques are having set of orthogonal dimensions. Some of these are 1) the hiding algorithm uses the support or the confidence of the rule to drive the hiding process. 2) The modification in the raw data that is caused by the hiding algorithm. The two types of the modification comprise the distortion and the blocking of the original values.3) a single rule or a set of rules can be hidden during an iteration of the hiding algorithms. Based on this criterion differentiate hiding algorithms into single rule and multiple rule schemes. 4) The nature of the hiding algorithm, which can be either heuristic or exact.

Based on the above dimensions so many techniques and methods are utilized in the following papers. “Association rule hiding” has been mentioned for the first time in 1999 in a workshop paper by Atallah et al. some authors M. Atallah, E. Bertino, A. Elmagarmid, M. Ibrahim, and V. S. Verykios tried to apply general ideas regarding the implications of data mining in security and privacy of information to the association rule mining framework. They proposed a number of solutions like fuzzification of the source database, limiting access to the source database, as well as releasing of samples instead of the entire database.

Mohammad Naderi Dehkordi et al., gives Association rule Mining based on genetic Algorithms. In that research they introduce new multi-objective method for hiding sensitive association rules based on the concept of genetic algorithms. The main purpose of this method is fully supporting security of database and keeping the utility and certainty of mined rules at highest level.

A lot of research has done in this area but most of them focused on perturbation of original database heuristically. Therefore the final accuracy of released database falls down intensely and generated some fake rules. In addition to accuracy of database the main aspect of security in this area is privacy of database that is not warranted in most heuristic approaches, perfectly due to the modification and generating additional rules. Due to their efficiency and scalability, heuristic approaches have been focus in the data mining area.

3. PROBLEM FORMULATION

Consider a database D, consisting of N transactions, m items and thresholds minsub and minconf set by the proprietor of the data. After performing association rule mining in during thresholds minsub and minconf, get a set of association rules, denoted as AR, among which a subset SR of AR contains rules which are considered to be sensitive from the proprietor's perspective. Given the set of sensitive association rules SR, the goal of association rule hiding methodology is to construct a new modified database D' from D, which achieves to protect the sensitive association rules AR from disclosure, while minimally affecting the non-sensitive rules existing in AR. In this work, select various datasets which contains sensitive items and non-sensitive items. The main objective of this work is hiding sensitive rules by converting sensitive items into non-sensitive items. The sensitive items are received by apriori algorithm in association rule mining. From the set of association rules, approachable association rules can be extracted from a specific data collection and are considered to be sensitive, the task of association rule hiding algorithms is to properly transform the original data so that the association rule mining algorithms that may be applied to this modified data (i) will be incompetent to determine the sensitive rules (ii) will be capable to mine all the non sensitive rules that become visible in the original dataset and (iii) will be incapable to discover false rules.While generating the association rules after applying the ant colony algorithm all sensitive association rules are hided.

4. PROPOSED SOLUTION

Optimization Technique

Optimization is a mathematical regulation that concerns the discovery of minima and maxima of functions, subject to so-called constraints. Optimization techniques used for solving problems in which one seeks to minimize or maximize a real function by analytically choosing the values of real or integer variables from within an allowed set. Many types of optimization techniques and optimization algorithms are using in various types of approaches for solving problems. In this paper use the Ant colony optimization algorithm for modifying the database in a optimal way.

Ant colony Optimization

Ant Colony Optimization (ACO) is a powerful technique for designing metaheuristic algorithms for combinatorial optimization problems. The first algorithm which can be classified within this framework was presented in 1991[12].The essential feature of ACO algorithms is the combination of a priori information about the structure of a promising solution with a posteriori information about the structure of previously obtained good solutions.

In the real world, ants (initially) wander randomly, and upon finding food return to their colony while laying down pheromone trails. If other ants find such a path, they are likely not to keep traveling at random, but instead follow the trail laid by earlier ants, returning and reinforcing it if they eventually find food. Over time, however, the pheromone trail starts to evaporate, thus reducing its attractive strength. The more time it takes for an ant to travel down the path and back again, the more time the pheromones have to evaporate. At the same concept of database if items are occurred in the frequent transactions, it will be taken as a sensitive item. If change the item at randomly using functions, can hide the frequent items.

Ant colony optimization algorithm

Ant colony optimization(MetaHeuristic)

```
1.        {Initialization 1}
$\lambda_{1,}\lambda_{2}............\lambda n$
2.        {initialization 2}
$n_1(\theta)\ \varepsilon\ \lambda_{1....}\ n_n(\theta)\ \varepsilon\ \lambda_n$
3.        {construction}
For each t ε D
Do
    $Cost(t) = w1 * \sum_{i=1}^{n(t)} + w2 * n(\lambda(t))$
4.        Select t = maxcost(t) ¥ t and
Select i = max (n (θ)) ¥ i
5.        Update i to 0 at above i in t.
$n(\theta) = n(\theta) - 1$
    repeat step 2
    until $n(\theta) = 0$.
6.        {terminating condition}
If $n(\theta) = 0$
Exit;
```

Process of Ant colony optimization Algorithm

First sensitive item are initialized from the association rule mining. Based on the construction of transactions, update the number of modifications for each sensitive item. Initialization is common for all iterations and also no of modifications are updated for each transaction based on the modification of the item. If a particular item is modified that item count is reduced by 1.After performing the ant colony optimization algorithm apply the Apriori algorithm for find frequent itemsets and generate the sensitive rules from the database. So uses of ant colony algorithm all sensitive rules become hide without fake rules.

Experimental Results
Sample database

Table 1 sample database

A	B	C	d	e	F
1	0	1	1	0	0
1	1	1	0	1	0
0	1	1	0	1	0
0	1	0	1	0	0
1	1	1	0	0	1

The above table represents the some of the transactions and items. It contains 5 transactions and 6 itemsets. The presence of the item is represented is represented as 1, and absence of the item is 0.Threshold value such as support is 28% and confidence is 58%.From the above table items **a,b,c** is taken as a frequent itemset using the Apriori algorithm.

Rules from the frequent itemsets are as shown below,

Min sup=25% Min conf=58%

Table 2 association rules

Association rules	Confidence
a→b	66.67%
a→c	100%
a→bc	66.67%
ab→c	100%
ac→b	66.67%
b→c	75%
bc→a	66.67%
bc→e	66.67%
be→c	100%
c→a	75%
c→b	75%
ce→b	100%
e→c	100%
e→bc	100%
e→b	100%

After applying the ant colony optimization result will be shown below,

Table 3 result table

A	B	C	D	E	F
1	0	0	1	0	0
1	1	0	0	1	0
0	0	1	0	1	0
0	1	0	1	0	0
0	0	1	0	0	1

After applying the Apriori algorithm in the modified database sensitive rules are hided with the same support and confidence.

The following performance factors are used for evaluating the ant colony optimization algorithm

1. Hiding Failure
2. Sensitive Rule Protection
3. False rule generation
4. Misses cost or protection of Non sensitive rules
5. No. of iterations
6. Time complexity

Hiding Failure (HF)
This measure quantifies the percentage of the sensitive patterns that remain exposed in the sanitized dataset. It is defined as the fraction of the restrictive association rules that appear in the sanitized database divided by the ones that appeared in the original dataset. Formally,

$$HF = \boxed{|ARsen(D)| \;/\; |ARsen(D')|}$$

Where $|AR_{sen}(D')|$ corresponds to the sensitive rules discovered in the modified dataset D', $|AR_{sen}(D)|$ corresponds to the sensitive rules appearing in the original dataset D and |X| is the size of set X. Ideally, the hiding failure should be 0%.

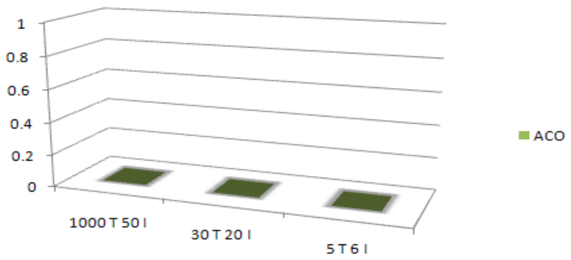
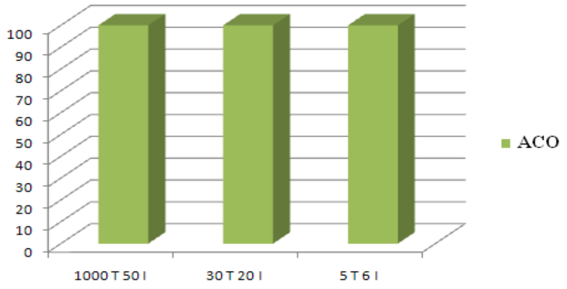


Figure 1: Hiding Failure

Sensitive Rule Protection

Figure 1 shows that the hiding failure is 0% which means all the sensitive rules are protected from the disclosure. The accuracy of sensitive rule protection is 100%. It is represented in the figure 2.



Ant colony optimization algorithm has protected all the sensitive rules. It gives 100% privacy protection for any different size of data sets.

False Rule Generation

This measure computes the percentage of the false rules can be generated which is to be considered as a side-effect of the modification process. It is computed as follows:

FR= [ARnonsen(D)+ARsen(D)] -

[ARnonsen(D')+ARnonsen(D')]

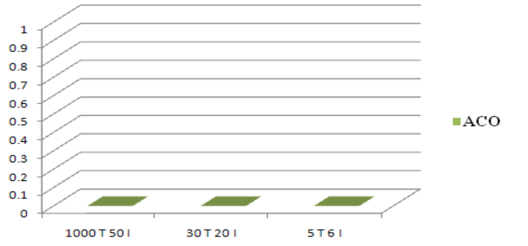


Figure 3: False Rule Generation

No false rules are generated after modifying the sensitive items using ant colony optimization. It is represented in Fig 3 which shows 0% of false rule generation

Cost or Protection of Non-Sensitive Rules

This measure computes the percentage of the non-sensitive rules that are hidden as a side-effect of the modification process. It is computed as follows:

MC= |ARnonsen(D) – ARnonsen(D')| / |ARnonsen(D)

where $AR_{NON-SEN}(D)$ is the set of all non-sensitive rules in the original database *D* and $AR_{NON-SEN}(D')$ is the set of all non-sensitive rules in the sanitized database *D'*. Ant colony optimization gives the 100% privacy protection for all the non sensitive rules.

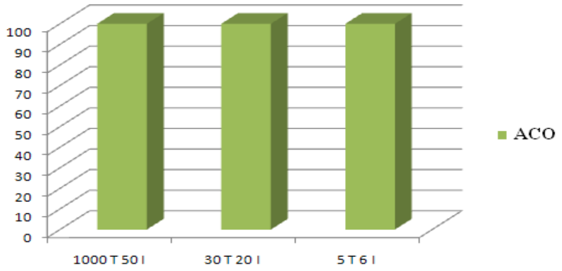


Figure 4: Non sensitive Rule Protection

Number of Iterations Required

The chart shows the iterations required by the ant colony optimization for performing the modification process. Various threshold values are applied to modify the sensitive items and protect the sensitive rules.

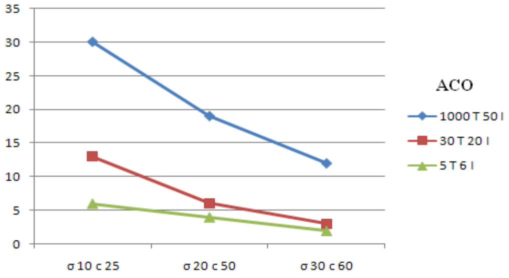


Figure 5: No. of Iterations

Time Complexity

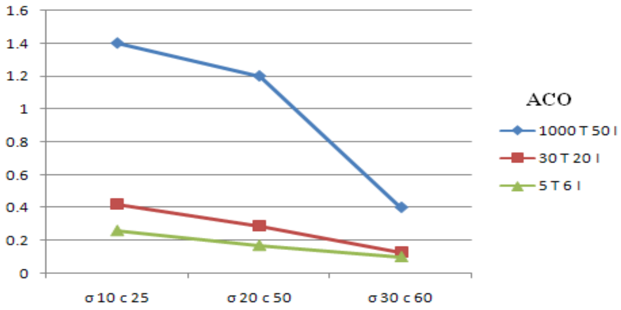


Figure 6: Time Complexity

5. CONCLUSION

In this paper use the association rule with ant colony optimization algorithms for hiding the frequent item sets. In this paper developed new algorithm for choose the transaction for modifying the process. Compared with the previous approaches this approach modified the data without loss. Performance and time complexity is better than the other optimization techniques. So keep the privacy and also accuracy of the database without loss of data and minimum number of iterations.

6. REFERENCES

[1] A.Gkoulalas-Divanis and V.S.Verykios, Association Rule Hiding for Data Mining, *Advances in Database Systems 41, DOI 10, 1007/978-1-4419-6569-1_2, Springer Science + Business Media*, LLC 2010

[2] Bikramjit Saikia, Debkumar Bhowmik "Study of Association Rule Mining And different hiding Techniques" Department of computer Science Engineering, National Institute of Technology,Rourkela.

[3] Charu C.Aggarwal IBM T.J. Watson Research Center, USA and Philip S. "Privacy Preserving Data Mining: Models and algorithms" Yu University of AIllinois at Chicago, USA.

[4] Dasseni,V.S. Verykios, A. Elmagarmid, and E. Bertino.Hiding association rules by using confidence and support. In: Proc. of the 4th *Int'l Information Hiding Workshop (IHW'01*). Springer-Verlag, 2001. 369-383.

[5] E.D. Pontikakis, A. Tsitsonis, and V.S. Verykios. An experimental study of distortion-based techniques for association rule hiding. In Proc. of the 18th Annual IFIP WG 11.3 *Working Conf. on Data and Applications Security. 2004.*

[6] M. Atallah, E. Bertino, A. Elmagarmid,M. Ibrahim and V. Verykios. Disclosure limitation of sensitive rules. Proc. Of *IEEE Knowledge and Data Engineering Exchange Workshop (KDEX),* November 1999.

[7] Mohammad Naderi Dehkordi, Kambiz Badie, Ahmad Khadem Zadeh "A Novel Method for Privacy Preserving in Association Rule Mining Based on Genetic Algorithms" , *Journal of Software*, Vol.4, No.6 August 2009

[8] Nada M. A. Al Salami "Ant Colony Optimization Algorithm" Rakesh Agrawal,Tomasz lmielinski,Arun Swami "Mining Association Rules between sets of items in Large Databases".IBM Almaden Research Center,San Jose,CA 95120.

[9] T. Johnsten, V. Raghavan, K. Hill: The security assessment of association mining algorithms. In: Proceedings of the 16th Annual IFIP WG 11.3 *Working Conference on Database Applications Security*, pp. 163–174 ,2002.

[10] Vaidya, j.Clifton, .W; Zhu, Y.M 2006, X, 121 p, 20 illus, Hardcover"Privacy Preserving Data Mining" ISBN: 979-0-387-25886-7.

[11] Vittorio Maniezzo, Luca Maria Gambardella, Fabio de Luigi "Ant colony optimization".

[12] V. Verykios, A. Elmagarmid, E. Bertino, Y. Saygin and E. Dasseni. Association Rule Hiding. *IEEE Trans. On Knowledge and Data Engineering, 16(4), 2004.*