International Journal of Applied Engineering Research ISSN 0973-4562 Volume 15, Number 4 (2020) pp. 425-430

BSE- QP– ICOA for Association Rule Hiding

G.Bhavani[1], Dr.S.Sivakumari[2]

[1]Research Scholar, [2]Professor and Head,
Department of Computer Science and Engineering,
Avinashilingam Institute for Home Science and Higher Education for Women,
School of Engineering, Coimbatore, Tamilnadu, India.

ABSTRACT

Association rule hiding is a well-organized key that helps data users evade the risk initiated by sensitive information leak when allocating data in their collaborations. Quality Preserving-Improved Cuckoo Optimization Algorithm for Association Rule Hiding (QP-ICOA for ARH) is an efficient association rule hiding algorithm. QP-ICOA for ARH reduces the side effects on insensitive rules while hiding the sensitive items in the transaction database. In this paper, the QP-ICOA4ARH is improved by proposing Balancing Stochastic Exploration within QP-ICOA for the ARH method (BSE-QP-ICOA) for ARH. In BSE-QP-ICOA, the variable limits are adjusted dynamically according to the fitness value of each cuckoo in the population. It also helps to sanitize the database with a high convergence rate. The efficiency of QP-ICOA for ARH is improved by dynamically changing the value of the switching parameter and the total number of iteration. A balancing between exploration and exploitation in QP-ICOA for ARH achieved by using crossover and mutation operators in the cuckoo search algorithm. The population size is modified, at each iteration, by the linear population reduction method that helps to increase the convergence rate and to escape from local optima problem. An experimental analysis is carried out in adult, bank marketing, and hardware store sales datasets in terms of hiding failure and lost rules to prove the effectiveness of BSE-QP-ICOA.

Keywords: Association rule hiding, cuckoo search algorithm, switching parameter, Quality Preserving-Improved Cuckoo Optimization Algorithm for the Association Rule Hiding, Balancing Stochastic Exploration.

1. INTRODUCTION

In the current information age, ubiquitous and pervasive computing is continually creating a vast volume of information. The investigation of this data has shown to be beneficial to numerous services such as cybersecurity, health care, banking transportation, commerce, and many others. However, most of the collected information may be sensitive to private data, which raises privacy concerns. Privacy-Preserving Data Mining (PPDM) [9] protects the data privacy of an individual without dropping its value. Being aware of the intrusions that could happen on their data, people hesitate to share confidential information.

In sharing private data, the data holder must have a solution to preserve privacy while confirming exact mining results [10]. Association rule mining is one of the main productive approaches for privacy-preserving data mining that acts as a more authoritative tool to discover relationships hidden in large datasets. Such links represent the association rules, which identify all the possible ways in which public data can reveal private information. A Cuckoo Optimization Algorithm for ARH (COA4ARH) [8] proposes the technique to hide the sensitive association rule. COA4ARH was a type of data distortion technique where each cuckoo in the population optimally distorted delicate objects to sanitize the unique dataset. But, a fixed number of transaction modifications in COA4ARH which was not suitable for a variety of datasets.

So, Improved COA (ICOA) [1] proposes a minimum number of transactions for modifications that were decided on the minimal level of support and confidence - MST and MCT respectively. Moreover, a new fitness function was introduced which enhance the association rule hiding process. In addition to this, a multi-objective problem in COA4ARH was solved by the Pareto-optimal solution. Still, it is unable to hide the rules with multiple LHS and RHS items in the rules. So, Quality Preserving ICOA for ARH (QP-ICOA for ARH) [2] was proposed to reduce the hiding failure and lost rules on insensitive rules by hiding multiple LHS and RHS items in the rules. Based on the influence of an item, QP-ICOA for ARH decided whether the items were removed or reinserted in the transactions to sanitize the database.

In this paper, Balancing Stochastic Exploration within QP-ICOA for ARH method (BSE-QP-ICOA) is proposed for effective association rule hiding. In BSE-QP-ICOA, the searching space of cuckoos is adjusted in each iteration by modifying the low limit and high limit of each variable based on minimizing the fitness value [1]. In QP-ICOA for ARH, switching parameter and number of iteration is kept constant. By dynamically changing the switching parameter and number of iterations, the efficiency of QP-ICOA for ARH is improved. In addition to this, stochastic exploration and population reduction features are introduced to enhance the association rule hiding process.

2. LITERATURE SURVEY

Ponde & Jagade [12] developed a heuristic-centered algorithm to hide delicate association rules. Based on the investigation

of Frequent Pattern-Growth (FP-Growth) and Apriori algorithm, a matrix Apriori algorithm was proposed which generated association rules. The transactions were arranged in descending order of their sensitivity. Then, a binary matrix was formed by assigning 0 to the deleted items in the ordered transactions. The confidence and support of sensitive rules which contained deleted items were updated. Finally, the delicate rules are deleted depending on the support and confidence threshold. This process gets repeated to completely hide all the delicate rules. However, side effects due to reducing the database need to be reduced.

Cheng et al. [11] developed a border rule-depending distortion algorithm for ARH by deleting some items in the transaction database with decreased confidence and support levels of delicate rules below precise thresholds. In addition to this, positive and negative border rules were described to identify the delicate rules. The supporting relations were estimated based on their relation with positive and negative border rules. Generally, feeble relevant items were chosen for modifications. However, it requires high CPU time for association rule hiding.

Afzali & Mohammadi [7] proposed a fuzzy logic approach for ARH in big data. In this approach, anonymization methods were used and it avoided undesired effects of removing frequent itemsets on new entrance data. An appropriate membership function was used to find the sensitive degree of each association rule and based on that membership function anonymization process was carried out. Based on the confidence value of each transaction and confidence threshold, the fuzzy logic approach hides the sensitive rules. However, if there are any changes in membership function, it causes some effects on the height of appropriate generalization.

Menaga & Revathi [6] introduced whale optimization and Least Lion Optimization Algorithms (LLOA) for ARH. From the input database, the whale optimization mined the association rules and those were validated by a newly formulated fitness function. A modified version of the Lion Optimization Algorithm (LOA) was LLOA where Least Mean Square (LMS) was included along with LOA processes that created an undisclosed key to preserve privacy. The sanitized database was created by LLOA using an optimal secret key which hides delicate evidence using the privacy and utility factor. Yet, the convergence speed of LLOA rest on on the ending standard.

Murthy et al. [5] developed a MAXARH algorithm that identifies the delicate rules and preserves privacy. It splits the association rule process into conversion and mining phase, identification and hiding of rules. The input transaction database was converted into binary values then applied Apriori-based ARM to mine the association rules. The best rules were identified based on maximum confidence and maximum support and the best rules were hidden by replacing the binary values. But still, MAXARH has side effects due to hiding association rules.

Mohan & Angamuthu [4] developed a genetic algorithm based on hiding technique and a technique for creation of dummy item for ARH. Initially, the cost of each transaction in the database was calculated and then chosen the sensitive items one by one for modifications. Based on the descending order of transaction costs, all transactions in the database were arranged. The sensitive item in each transaction was modified from 1 to 0 and then the new cost of transactions was calculated. A modified database was obtained by continuing the same process for all sensitive items. However, the artifactual fault level is high.

3. PROPOSED METHODOLOGY

In this section, the proposed BSE-QP-ICOA for association rule hiding is described in detail. The BSE-QP-ICOA process is started by applying the cuckoo search optimization algorithm in the original database which creates association rules R_s. After that, the sensitive rules R_{sens} are selected from R_s based on the minimal support and confidence level. Then the database is pre-processed by selecting the transaction which supported one or more sensitive rules and these are called critical transactions. The delicate items with a serious role in sanitization are addressed for alteration. As per the properties in [1], a least amount of transactions are selected. After the selection of a minimum number of transactions, QPICOA for ARH is processed where each cuckoo removes or reinserts a victim item to hide the sensitive association rules. The fitness value of each cuckoo in the population is calculated and selects the best solution.

For producing new solutions, each parent solution is merely permitted to alter a restricted number of victim items in the minimized transactions that are called as modification radius Mod_Rad. It is calculated as,

$$Mod\_Rad = \left[\alpha \times \frac{Current\ solution's\ K}{Total\ of\ all\ solution's\ K}\right] \times \left(V_{high} - V_{low}\right) \quad (3.1)$$

In Eq. (3.1), K is the quantity of a new solution that should be produced by each parent solution, α represents the quantity of entire repetitions of RHS items of delicate rules in serious relations, V_{high} is the high limits of each variable in optimization problem and V_{low} is the low limits of each variable in the optimization problem. V_{high} and V_{low} are ranging from 0 to 1.

The modification radius could also be used to set the searching space of cuckoos. In BSE-QP-ICOA, the range of V_{low} and V_{high} can be altered depending on the following fitness function,

$$\text{Minimize } \vec{f} = [f_1, f_2, f_3, f_4, f_5] \quad (3.2)$$

In Eq. (3.2), f_1 denotes the hiding failure, f_2 denotes the lost rule,

$$f_3 = Rule\ Hiding\ Distance + Rule\ Lost\ Distances,$$

$$f_4 = \frac{No.of\ Ghost\ Rules}{Total\ No.of\ Rules},$$

$$f_5 = \frac{No.of\ transactions\ that\ are\ sanitized}{Size\ of\ the\ transaction\ database}$$

The range of V_{high} and V_{low} is increased or decreased based on the following condition,

1. If $\vec{f}$ is maximum, then increase V_{high} and decrease V_{low}

2. If $\vec{f}$ is minimum, then decrease V_{high} and increase V_{low}

By adjusting the variable limit, the searching space of the cuckoos can be adjusted. Hence, it sanitizes the transaction database with a high convergence rate. To produce a new solution x_i^{t+1} for the ith cuckoo at $t+1$ iteration, Levy flight is executed which is referred as global random walk as follows:

$$x_i^{t+1} = x_i^t + \alpha \times Mod\_Rad \otimes \text{Levy } (\lambda)(\mathrm{x}_{\mathrm{best}} - x_i^t) \quad (3.4)$$

The local random walk is given as follows:

$$x_i^{t+1} = x_i^t + \alpha \times Mod\_Rad \otimes H(P_a - \varepsilon)\left(x_j^t - x_k^t\right) \quad (3.5)$$

In Eq. (3.4) and Eq. (3.5), x_i^t is the parent solution, $\alpha > 0$ is the step size related to problem scales, H is the heavy-side function, P_a is the switching parameter, ε is the random number, MR is the modification radius, $\otimes$ is entry wise multiplication, $\mathrm{x}_{\mathrm{best}}$ is the existing best solution, x_j^t, and x_k^t are randomly selected solutions. In BSE-QP-ICOA, the fixed values of P_a and α are dynamically changed with the number of generations and have been expressed as,

$$P_a(g_n) = P_{a\_max} - \frac{g_n}{NI}\left(P_{a\_max} - P_{a\_min}\right) \quad (3.6)$$

$$\alpha(g_n) = \alpha_{max} \exp(c.g_n) \quad (3.7)$$

$$c = \frac{1}{NI} L_n\left(\frac{\alpha_{min}}{\alpha_{max}}\right) \quad (3.8)$$

Where NI is the number of total iterations and g_n is the current iteration. By using Eq. (3.6-3.8), the P_a and α are changed dynamically which are used in the local random walk and global random walk to generate a new solution from parent solution.

A balancing between exploration and exploitation [6] is achieved by using crossover and mutation operators in cuckoo search. During the mutation process, all the items in the transaction database are changed. The mutation process is typically a unary operation. To preserve nonsensitive items, the crossover process is introduced. The crossover operation demands an interaction between two or more cuckoos and enables a flow of information inside the population. This flow is controlled by MR. The mutation strategy is introduced in the global random walk and local random walk of the BSE-QP-ICOA which are expressed in Eq. (3.8) and Eq. (3.9).

$$u_{ig}^t = x_i^t + \alpha \times Mod_{Rad} \otimes \text{Levy } (\lambda) \times F_i(x_{best}^t - x_i^t) + \left(x_j^t - x_k^t\right) \quad (3.9)$$

$$u_{il}^t = x_i^t + \alpha \times Mod\_Rad \otimes H(P_a - \varepsilon) \times F_i(x_{best}^t - x_i^t) + \left(x_j^t - x_k^t\right) \quad (3.10)$$

Where F_i denotes a scaling factor regulating the magnitude of the change and x_{best}^t is the current best solution. A crossover operator is introduced in the mutation process which has a crucial impact on the association rule hiding performances. The crossover operator is given as follows:

$$w_i^t = \begin{cases} u_i^t, & rand_j(0,1) \le Mod\_Rad \vee j = j_{rand} \\ x_i^t, & otherwise \end{cases} \quad (3.11)$$

In Eq. (3.11), the condition $j = j_{rand}$ ensures the new solution differs from the original solution x_i^t in at least one element. Finally, the new solution is obtained from the following Eq. (3.10).

$$x_i^{t+1} = \begin{cases} w_i^t, & if \vec{f}(w_i^t) \le \vec{f}(x_k^t) \wedge k \ne i \\ x_i^t, & otherwise \end{cases} \quad (3.12)$$

In Eq. (3.12), $k = rand\ (0, Pop\_Size)$ is a randomly selected integer drawn from uniform distribution in interval $[0, Pop\_Size)$. The population size Pop_Size of the nest makes a great impact on the association rule hiding performance. The small-sized population tends towards faster convergence but also increases the risk of getting trapped in a local optimum. On the other hand, the larger-sized population convergence slower but provides an exploration of the search space. In BSE-QP-ICOA, the population size is modified by the linear population reduction which is expressed as follows:

$$Pop\_Size^{t+1} = round\left[\left(\frac{Pop\_Size_{max} - Pop\_Size_{min}}{5}\right) \times \vec{f}_{Num}\ Pop\_Size_{min}\right] \quad (3.13)$$

In Eq. (3.13), Pop_Size_{max} is the starting population size, Pop_Size_{min} is the user-specified minimal population size and $\vec{f}_{Num}$ is the existing quantity of the fitness function. By using Eq. (3.13), the population size of the cuckoos is modified in iteration and it increases the convergence speed and provides a better exploration of the search space. Based on the new population size, the next iteration of BSE-QP-ICOA is processed. The total number of existing solutions is limited to a threshold value N_{max}.

In this manner, the worst solutions need to be removed, so that the quantity of remaining solution remains equal to N_{max}. To improve the remaining solution, the other solutions are migrated towards the best solution based on immigration function [8]. This process is repeated till a extreme quantity of repetition is achieved.

BSE-QP-ICOA Algorithm

Input: Original dataset D, α, V_{high}, MST, MCT, V_{low}, Pop_Size, Itr_{max}, N_{max}

Output: Sanitized Database D'

1. Begin
2. Create association rules R_s by applying cuckoo search algorithm in D
3. Choose sensitive rules R_{sens} from R_s based on MST and MCT
4. Preprocess D by choosing critical transactions

5. Select the minimum number of transactions for association rule hiding
6. Generate an initial population and compute the fitness value $\vec{f}$ of each cuckoo using Eq. (3.2).
7. for $itr = 1: Itr_{max}$ do
8. Calculate K value and change a limited number of victim items based on Mod_Rad value using Eq. (3.1).
9. Based on the condition in Eq. (3.3), modify V_{high} and V_{low} values
10. for K times do
11. for Mod_Rad times do
12. Generate a new solution using Eq. (3.12).
13. Set $Rand[0,1]$ to the selected victim item.
14. Modify the population size using Eq. (3.13).
15. end for
16. end for
17. Calculate $\vec{f}$ of the new solution.
18. Limit the quantity of solutions to N_{max}.
19. for each solution in a population do
20. Transfer all solutions towards the finest solution
21. end for
22. Calculate fitness value $\vec{f}$
23. Choose the best solution
24. end for
25. End

4. RESULTS AND DISCUSSION

In this section, the performance of BSE-QP-ICOA is analyzed and compared with QP-ICOA4ARH in terms of hiding failure and lost rule. In this investigational resolution, three different datasets called adult, bank marketing and hardware store sales dataset are used.

The adult dataset consists of 32,561 transactions, 14 items with average transaction length is 15. The bank marketing database consists of 4522 transactions and 17 items with average transaction length are 17. Hardware store sales dataset is a real-time dataset collected from MVS traders about sales details from January-1, 2017 to January-1, 2018. It consists of 1, 00,000 transactions, 1000 items with an average transaction length of 100.

4.1 Performance Metrics

Hiding Failure

Hiding Failure (HF) denotes the count of delicate rules which QP-ICOA4ARH and BSE-QP-ICOA algorithms could not hide and are still mined from the sanitized data. It is calculated as,

$$HF = \frac{|R_s(D')|}{|R_s(D)|} \tag{3.14}$$

In Eq. (3.14), $R_s(D')$ is the number of sensitive rules discovered in D' and $R_s(D)$ is the number of sensitive items discovered in D.

Lost Rule

Lost Rule (LR) measures the number of non-sensitive rules that are lost because of the act of QP-ICOA4ARH and BSE-QP-ICOA algorithms. The non-sensitive rule will not mine from the D'. It is calculated as,

$$LR = \frac{|\sim R_s(D)| - |\sim R_s(D')|}{|\sim R_s(D)|} \tag{3.15}$$

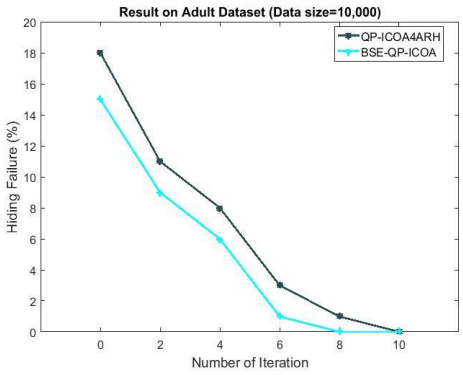
In Eq. (3.14), $|\sim R_s(D)|$ is the number of non-sensitive rules explored in D and $|\sim R_s(D')|$ is the number of non-sensitive rules explored in D'.

5. PERFORMANCE ANALYSIS

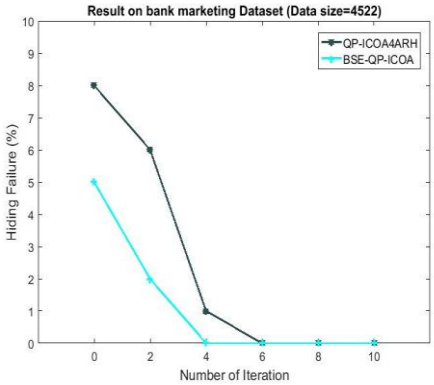
The performance of QP-ICOA4ARH and BSE-QP-ICOA are tested in terms of hiding failure and lost rules under three different datasets

Fig.1. Hiding Failure on

a) Adult Dataset



b) Bank Marketing Dataset



c) Hardware Stores Dataset

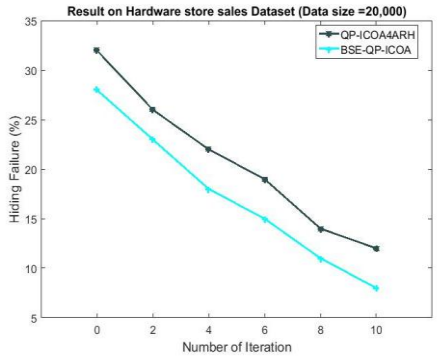
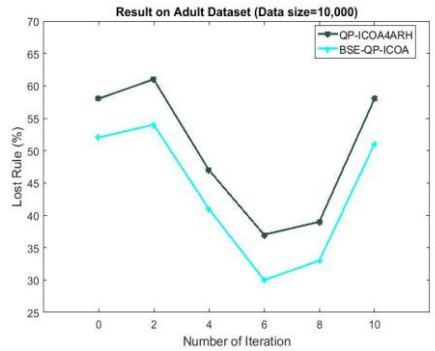
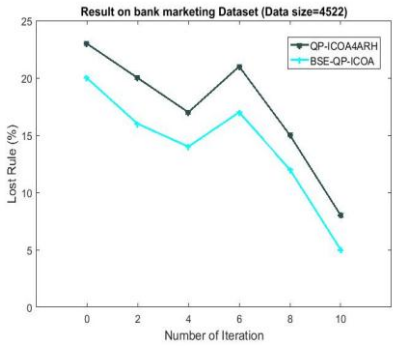


Fig.2. Lost Rule on

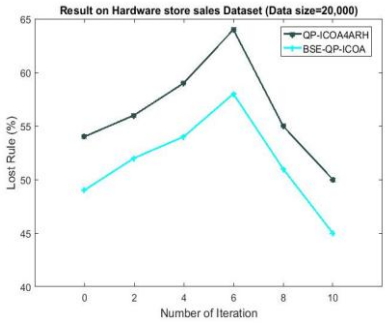
a) Adult Dataset



b) Bank Marketing Dataset



c) Hardware Stores Dataset



From the above fig.1 and 2, it gets proved that the proposed BSE-QP-ICOA has less hiding failure and lost rules than QP-ICOA for ARH for adult, bank marketing, and hardware store sales datasets.

5. CONCLUSION

In this paper, BSE-QP-ICOA is proposed for efficient association rule hiding. Initialize the population size and then cuckoo sanitize the database removing or inserting a victim in the selected number of transactions. Then, the fitness of each cuckoo calculated, and a new best solution is selected based on the fitness value. While generating a new solution, the variable limits are adjusted dynamically, which adjusted the modification radius of cuckoos in each of the iterations. The switching parameter and number of the iteration are modified to increase the efficiency of the association rule hiding process. Crossover and mutation operators are introduced in QP-ICOA for ARH to achieve the balance between exploration and exploitation. A linear population reduction is used to modify the population size that increases the convergence rate and helps to escape from local optima. The experimental results prove that the proposed BSE-QP-ICOA has better hiding failure and lost rules than QP-ICOA for ARH for adult, bank marketing, and hardware store sales dataset.

REFERENCES

[1] Bhavani, G., & Sivakumari, S. (2019). Improved cuckoo optimization algorithm for association rule hiding with minimal side effects. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, *8*(10), 337-342.

[2] Bhavani, G., & Sivakumari, S. (2019). Quality Preserving Improved Cuckoo Optimization Algorithm for the sensitive Association Rule Hiding. *International Journal on Emerging Technologies (IJET),10(4), 472-477*.

[3] Menaga, D., & Revathi, S. (2018). Least lion optimization algorithm (LLOA) based secret key generation for privacy-preserving association rule hiding. *IET Information Security*, *12*(4), 332-340.

[4] Murthy, T. S., Gopalan, N. P., & Venkateswarlu, Y. (2018). An Efficient Method for Hiding Association Rules with Additional Parameter Metrics. *International Journal of Pure and Applied Mathematics*, *118*(7), 285-290.

[5] Mohan, S. V., & Angamuthu, T. (2018). Association Rule Hiding in Privacy-Preserving Data Mining. *International Journal of Information Security and Privacy (IJISP)*, *12*(3), 141-163.

[6] Salgotra, R., Singh, U., & Saha, S. (2018). New cuckoo search algorithms with enhanced exploration and exploitation properties. *Expert Systems with Applications*, *95*, 384-420.

[7] Afzali, G. A., & Mohammadi, S. (2017). Privacy-preserving big data mining: association rule hiding using fuzzy logic approach. *IET Information Security*, *12*(1), 15-24.

[8] Afshari, M. H., Dehkordi, M. N., & Akbari, M. (2016). Association rule hiding using cuckoo optimization algorithm. *Expert Systems with Applications*, *64*, 340-351.

[9] Shah, A., & Gulati, R. (2016). Privacy-preserving data mining: Techniques classification and implications—A survey. *Int. J. Comput. Appl.*, *137*(12), 40-46.

[10] Mogtaba, S., & Kambal, E. (2016, July). Association Rule Hiding for Privacy-Preserving Data Mining. In *Industrial Conference on Data Mining* (pp. 320-333). Springer, Cham.

[11] Cheng, P., Lee, I., Pan, J. S., Lin, C. W., & Roddick, J. F. (2015). Hide association rules with fewer side effects. *IEICE TRANSACTIONS on Information and Systems*, *98*(10), 1788-1798.

[12] Ponde, P. R., & Jagade, S. M. (2014). Privacy-Preserving by Hiding Association Rule Mining from Transaction Database. *IOSR Journal of Computer Engineering (IOSR-JCE)*, *16*(5), 25-31.