International Journal on Emerging Technologies **10**(4): 472-477(2019)

**ISSN No. (Print): 0975-8364**
**ISSN No. (Online): 2249-3255**

# Quality Preserving ICOA for ARH

***G. Bhavani[1] and S. Sivakumari[2]***
*[1]Research Scholar, Department of Computer Science and Engineering, Avinashilingam Institute for Home Science and Higher Education for Women, School of Engineering, Coimbatore (Tamil Nadu), India.*
*[2]Professor and Head, Department of Computer Science and Engineering, Avinashilingam Institute for Home Science and Higher Education for Women, School of Engineering, Coimbatore (Tamil Nadu), India.*

*(Corresponding author: G. Bhavani)*
*(Received 18 September 2019, Revised 11 November 2019, Accepted 19 November 2019)*
*(Published by Research Trend, Website: www.researchtrend.net)*

**ABSTRACT: Data mining performs knowledge extraction from large databases. That knowledge has to be hidden in today's world to preserve privacy. One of the classic knowledge hiding approaches is Association Rule Hiding (ARH).ARH hides the sensitive association rules by modifying the real database without influencing the insensitive rules and real data. ARH can be done through Heuristic, Border based and Exact approaches. Optimization algorithms have been developed under the Meta-heuristic approach which lies under heuristic approach to perform maximum hiding of association rules. Among many optimization algorithms, Cuckoo Optimization Algorithm (COA) was an association rule hiding technique in which each cuckoo has been optimally distorted to clean the real database. Still, a fixed number of transaction modifications are carried out in COA which is not suitable for a variety of datasets. So an Improved COA (ICOA) was proposed to decide a minimum number of transactions for modifications. Additionally, ICOA-Crowding Distance (ICOA-CD) was introduced to minimize the conflicts between the multiple fitness functions in ICOA. In this paper, Quality Preserving ICOA (QPICOA) is proposed to hide the sensitive rules with multiple LHS and RHS. In QPICOA, to reduce the hiding failure and lost rules on insensitive rules, an item is chosen based on the correlation between sensitive and insensitive rules. The selected item is removed and reinserted in the transaction of each cuckoo based on objective parameters and sensitivity of transactions. This process is continued until a non-dominated best solution is obtained. Thus, each cuckoo hides the sensitive association rules by removing and reinserting items in the transactions. The experiments are conducted in adult, bank marketing and hardware store sales datasets to prove the effectiveness of QPICOA.**

**Keywords:** Association rule hiding, Crowding distance, Cuckoo Optimization Algorithm, Quality Preserving.

## I. INTRODUCTION

As a result of the rapid development of electronic data in different organizations, privacy-preserving data mining (PPDM) has become an important concern [1]. Such types of data contain sensitive information that could be misused when disclosed. Due to fast improvement in data mining technology, sensitive information of a user could be done easily. This makes the privacy of data a very important factor. Under PPDM, Association rule hiding is a subfield that analyzes the hiding failure and lost rules.

In a large number of application scenarios, data is collected or the extracted knowledge patterns have to be shared with other entities for specific purposes that affect privacy. The ARH process is to disinfect the data and make the ARM algorithms being applied to this data can mine all non-sensitive rules and unable to mine the sensitive rules [20].

A Cuckoo Optimization Algorithm (COA) was proposed to hide the sensitive association rule [2]. COA was hiding the sensitive rule by data distortion technique. Each cuckoo has been optimally distorted to clean the real database. However, a fixed number of transaction modifications are carried out in COA which is not suitable for a variety of datasets. So, Improved COA (ICOA) was proposed where a minimum number of transactions were selected for modification based on MST and MCT- threshold value for minimum support and confidence [3]. To improve ARH a new fitness function was introduced. The multi-objective problem was solved using Crowding Distance (CD) to improve the convergence of the Pareto-optimal solution. However, it is unable to hide the rules with multiple LHS and RHS items in the rules.

In this paper, Quality Preserving ICOA for ARH (QPICOA for ARH) is proposed to reduce hiding failure and lost rules on non-sensitive rules by hiding multiple LHS and RHS items in the rules. Initially, association rules are generated from the real transaction database by cuckoo search optimization. After pre-processing the real database, a minimum number of transactions are selected. In each iteration, the item correlation between insensitive and sensitive rules are calculated to reduce the hiding failure and lost rules on non-sensitive rules. A victim item has minimum influence in the insensitive rule. The influence of an item in the insensitive rule is measured based on objective parameters and sensitivity of transactions. The selected item is removed and reinserted in the transactions to sanitize a database. By removing and reinserting items a further refined sanitized database is obtained.

This advantage of the proposed approach is that, it produces results with minimal hiding failure and lost rules on insensitive rules.

## II. MATERIALS AND METHODS

The materials which were used for the literature survey are as follows:

Cuckoo Optimization Algorithm was proposed to hide sensitive association rules [2]. Distortion technique is used to prevent the rising sanitization time. Pre-processing was done in two phases - the selection of critical transactions, the selection of sensitive items with a critical role in sanitization. Three fitness functions were defined which were used to achieve the best solution and hide the sensitive association rules.

A novel approach for ARH was based on the data distortion technique in which the sensitive item's position was changed and the support of the item was unchanged [4]. Based on the idea of representative rules, the rules were pruned and then hide the sensitive rules based on a threshold value. But the efficiency of this approach is low.

An ARH algorithm (AARHIL) was proposed to hide sensitive association rules [5]. Two heuristics were devised to solve the issue of association rule hiding. To restrict the lost rules a victim item was determined in the first heuristic and focused on preserving item sets in the creating set. In the second heuristic, a weight value was assigned to hide sensitive rules. AARHIL needs improvement in terms of accuracy.

A Border based approach with BRDA for sensitive association rule hiding was done by removing certain items in a database [6]. In BRDA, positive and negative border rule concepts were used to find the rules which are affected by the modifications in the database. The supporting transactions were evaluated based on the relation with those rules and then weakly relevant ones were selected for modification. However, it consumes more time to hide sensitive rules.

Efficient algorithms for association rule hiding is based on the full item and consequent item sensitivities [7]. These item sensitivities are calculated using respective algorithms for all consequent sensitive rules of an item. Consequent sensitivity and full sensitivity rules were computed for each sensitive rule. The sensitive rules were arranged in decreasing order of their RCS/RFS and stored in Rule Consequent Sensitive Set/Rule Full Sensitivity Set. These were enabled to hide multiple rules. However, a threshold value greatly influences the efficiency of consequent and full item sensitive algorithms.

A hybrid approach called MDSLE was proposed for hiding association rules [8]. This approach was a combination of heuristic approach and ECLAT algorithm. The ECLAT was used to find sensitive items and frequent item sets in the transactions. The sensitive association rules were hidden by the heuristic approach. In MDSLE, hiding is done many times with minimizing modifications done to the database.

Another Optimization Algorithm (EFO) was proposed for sensitive association rule hiding [9]. In the EFO algorithm, fitness functions were utilized to gain a solution with minimal hiding failure and lost rules. Initially, using the fitness function electromagnetic particles were created and arranged. Then, electromagnetic particles were split into three groups as higher fitness magnitude, lower fitness magnitude, and neutral fitness magnitude.

HGA and DIC techniques were proposed for association rule hiding [10]. The individual transaction cost was computed and then the sensitive items in the transactions were chosen one by one for modifications. All transactions were ordered based on their cost. The sensitive item in each transaction was modified and then the new transaction cost was calculated. By repeating this, a modified database was obtained from all sensitive items. DIC was employed to hide the sensitive rules and it also created dummy items for the modified sensitive items. However, this technique has a high artifactual error rate.

ARH with different heuristic approaches discussed the advantages of this approach and its types [12]. ARM without pre-assign weights was proposed for discovering frequent itemset, but this transaction ranking method didn't fit for large database [13].

An optimization for ARM using an improved GA [14] predicted rules which contain negative attributes but the toolkit needs modifications on database for accurate results. Pandect on ARH discussed different hiding techniques, applications, importance and approaches for study [15]. The various ways in ARH expressed the types and goals of genetic, fuzzy and rough sets [16]. Data extraction in data mining from a large set of database reduces the size of data storage with increased accessing capacity [17]. Overview on ARM algorithms discussed on Apriori- basic, tid, hybrid and FP-Growth algorithms [18, 19]. MDSRRC algorithm along with Matrix Apriori algorithm were developed to improve the efficiency of ARH [21].

From the materials used for literature survey Cuckoo Optimization Algorithm was chosen to be effective compared to the other optimization techniques.

The methods implemented are as follows:

The QPICOA method for sensitive association rule hiding is described as follows. Initially, the cuckoo search optimization algorithm is applied in the real database to create association rules $R_s$ and then sensitive rules $R_{sens}$ are chosen from $R_s$ [11]. Transactions which supports one or more sensitive rules called critical transactions are selected in the real database to reduce the sanitization time. After this preprocessing, the important sensitive items in sanitization are selected for change. After addressing the sensitive items in the transactions, a minimum number of transactions are selected for sanitization based on two properties [3].

The minimized transaction database is processed by QPICOA where each cuckoo removes or reinserts a victim item. These items are chosen based on objective parameters are γ, δ, and sensitivity of transactions. The selected victim items are removed or reinserted based on fitness functions [3]. The multi-objective optimization problem is solved by a Pareto-optimal solution where CD is used to find a better solution for association rule hiding.

### *A. Quality Preserving ICOA for ARH*

The database containing the minimum number of transactions is further refined by minimizing the hiding failure and lost rules.

Each cuckoo in ICOA 4 ARH removes or inserts items in the transaction database for sensitive ARH. Sensitive rule hiding in the form of s → t can be achieved by decreasing either confidence or support of the rule to below the MCT and MST respectively. By decreasing the frequency of item set st, the support of the rule s → t can be decreased. By decreasing the support of consequence or increasing the support of the antecedent, the confidence of the rule is reduced.

The main intention of QPICOA is to remove or reinsert items in the transaction which consists of multiple LHS and RHS. In QPICOA, a Boolean variable S is used to describe the states of all sensitive rules. Initially, set the state of all sensitive rules as false. An association rule which has multiple LHS and RHS is in the form of Os→ Pt, where O, P ϵ I, O, P ⊂ I, and I is the item sets. Here, O and P are single items chosen by the QPICOA to be inserted into or removed from LHS or RHS of the rule, respectively. The support of the rule Os→ Pt can be decreased.

The confidence of a rule is decreased by inserting a selected item in the appropriate transaction. To reduce the support and confidence of a rule, the QPICOA finds a set of items. The selected item is removed or reinserted in the transaction from the selected minimum number of transactions. For this purpose, two objective parameters called γ and δ are calculated. The objective parameters are described as follows:

Parameter γ: The frequency of LHS items of the sensitive rules. It builds a list Rγ where LHS items are arranged in the increasing order of γ.

Parameter δ: The number of occurrences of an item in the sensitive set of rules for calculating the transaction sensitivity.

Transaction Sensitivity: It is calculated as the summation of δ values of all sensitive items included in that transaction.

The parameters γ, δ and sensitivity of transactions are calculated for the selected minimum number of transactions. After the calculation of γ, δ and sensitivity of transaction, transactions are sorted based on their sensitivity and length. Then, Rγ is calculated and the process of removing and reinserting of victim items is started from the first sensitive rules. In the first sensitive rule, from the number of LHS items, any item with a low value of γ is removed. Then the selected item is inserted in the transactions which have the large item sets that partially support LHS and don't or partially support RHS. The support and confidence values of the sensitive rules in the minimized database are updated after the removal and reinsertion process. If the support and confidence value reached below MCT and MST, then the false state is changed to true.

When a sensitive rule becomes disclosed because of inserting an item, the state is changed from true to false. At that time, the insertion will not be carried out in the minimized transaction database. But the removal of suitable items of the left side will be carried out to refine the transaction database with minimal hiding failure and lost rules on non-sensitive rules. This process is continued until all sensitive rules become true. This process is carried out in all cuckoos where the better solution for hiding association rule is achieved through the fitness functions are hiding failure, lost rule, the distance for hiding rule, the distance for the lost rule, number of ghost rules and transactions that are sanitized. The fitness function conflict is solved by using crowding distance to find a better ARH solution. Thus, the proposed QPICOA preserves the privacy of sensitive data and maintains the quality of the data. The overall process of QPICOA is given in the following algorithm.

*B. QPICOA Algorithm*

*Input*: Original Dataset D, MCT, MST, N trans
*Output*: Sanitized dataset D′

1. Use cuckoo search optimization, extract association rules $R_s$ from D
2. Choose sensitive rule set $R_{sens}$
3. State of all sensitive rules = false
4. Arrange confidence value $R_s$ in descending order.
5. Pre-process the D
6. Choose the least count of transactions $N_{trans}$
7. Create the first population
8. Every cuckoo arbitrarily count the sensitive items
9. Compute γ, δ and sensitivity of transactions
10. Create Rγ by arranging LHS in increasing order of their γ.
11. Arrange transactions based on sensitivity and length in descending order.
12. While (state ($R_{sens}$) ≠ true) do
13. Find the first rule $R_k$ from $R_{sens}$ such that state ($R_k$) is false
14. Choose item I from LHS of rule $R_k$ based on Rγ
15. For (m=1; m<$N_{trans}$; m++)
16. If ($T_m$ supports both parts of rule $R_k$)
17. Remove selected item I from transaction $T_m$
18. If ($R_k$ .Disclosed is false)
19. For (n=m; n< $N_{trans}$; n++)
20. If ($T_n$ does not include item I and partially support rule $R_k$)
21. Insert selected LHS item I in transaction $T_n$
22. End for
23. End if
24. Compute the fitness count of every cuckoo
25. Discover the top solution from the fitness count
26. Move all results to the top solution
27. Compute the fitness value for the latest top solution
28. Discover the top solutions from the capable function
29. Till end criterion is contented
30. Find the non-dominated solution (rule) using Pareto-optimal solution
31. For every rule R in $R_s$
32. Recalculate R. support and R. confidence for R
33. If (R. Support < MST|| R. Confidence < MCT)
34. Set State(R) to true
35. Else
36. If (State(R) == true)
37. Set R. Disclosed== true
38. Set State (R) == false
39. End if
40. End for
41. End while

By using the above QPICOA algorithm, the transaction database is sanitized for association rule hiding with minimum hiding failure and lost rules on non-sensitive rules

## III. RESULTS AND DISCUSSION

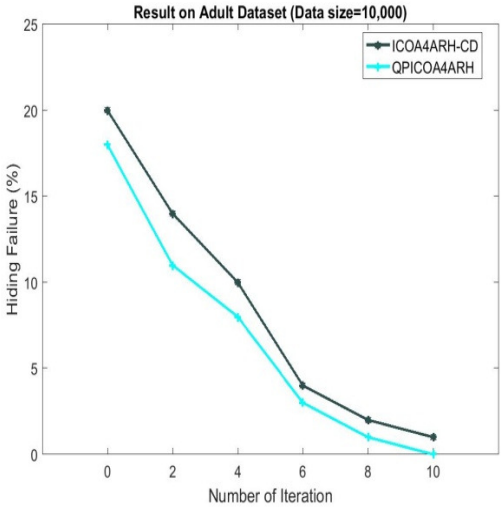
The effectiveness of the existing ICOA-CD and proposed QPICOA is evaluated in terms of lost rule and hiding failure. Adult, bank marketing and hardware store sales dataset are used for the experimental purpose. The adult dataset consists of 32,561 transactions, 14 items with average transaction length is 15. The bank marketing database consists of 4522 transactions and 17 items with average transaction length as 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, in which 1000 items with average transaction length of 100. The ICOA-CD and QPICOA have implemented in Java JDK 1.6 language and runs on a Microsoft Windows 7 with Intel processor running at 2.70 GHz and 4GB memory.

*A. Hiding Failure*

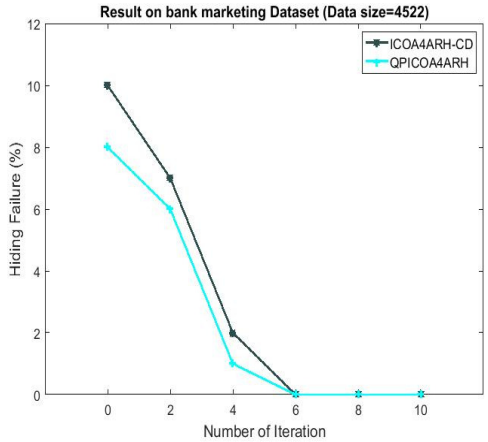
The HF value represents the number of delicate rules which were still unhidden in the sanitized data. HF is computed as,

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

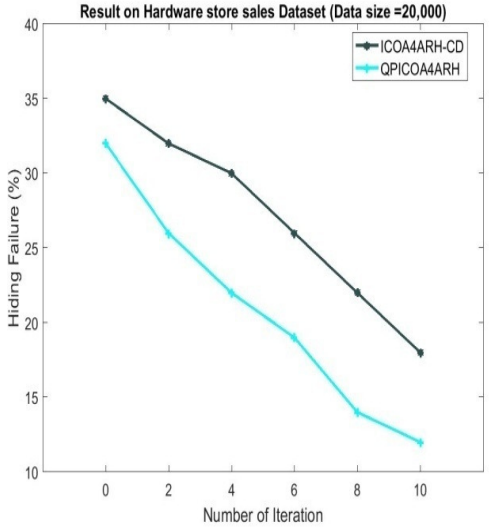
here, $R_s(D')$ – the number of delicate rules in D′
$R_s(D)$ - the number of delicate rules in D.




(a) Adult Dataset.




(b) Bank Marketing Dataset.




(c) Hardware store sales Dataset.

**Fig. 1.** Relationship between Hiding Failures.

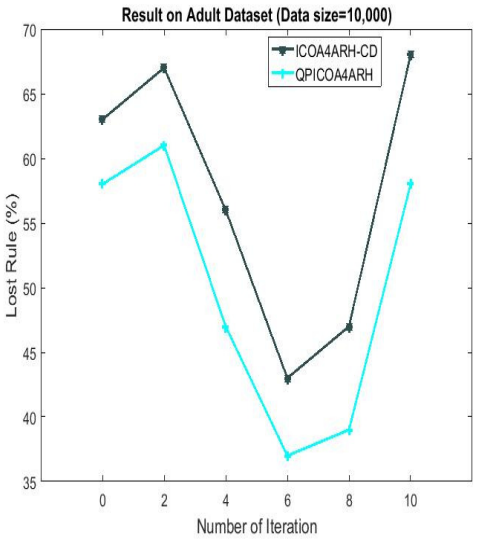
Fig. 1 shows the comparison between ICOA-CD and QPICOA algorithms in terms of hiding failure on adult, bank marketing and hardware store sales datasets. If the quantity of repetition is 4, then the hiding failure in the adult dataset is 20% reduced, in the bank marketing dataset it is reduced by 50% whereas in the hardware store sales dataset it is reduced by 26.7%. From this comparison, it is shown that the proposed QPICOA has better hiding failure than ICOA-CD for adult, bank marketing and hardware store sales datasets.

*B. Lost Rule*

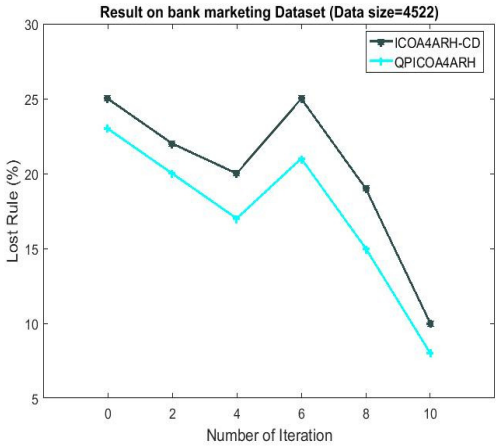
The LR denotes the amount of non-delicate rules that are lost. The non-delicate rule is not mined from D′. LR is computed as,

$$LR = \frac{|\sim R_S(D)| - |\sim R_S(D')|}{|\sim R_S(D)|}$$

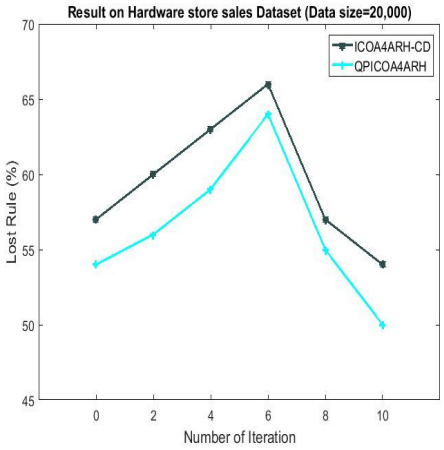
Here, $|\sim R_s(D)|$- the amount of non-delicate rules in D.
$|\sim R_s(D')|$- the amount of non-delicate rules in D′.




(a) Adult Dataset




(b) Bank Marketing Dataset.




(c) Hardware store sales Dataset.

**Fig. 2.** Relationship between Lost Rules.

Fig. 2 shows the relationship between ICOA-CD and QPICOA in terms of the lost rule on adult, bank marketing and hardware store sales datasets. If the amount of repetition is 6, then the lost rule 14% reduced in the adult dataset, 3.03% reduced in hardware store sales dataset, whereas if the amount of repetition is 4, then it is 16% in the bank marketing dataset. From this relationship comparison, it is shown that the proposed QPICOA has better-lost rule than ICOA-CD for adult, bank marketing and hardware store sales datasets.

## IV. CONCLUSION

In this paper, QPICOA is proposed to reduce the hiding failure and lost rules on non-delicate rules. A cuckoo search optimization is applied in the real transaction database to generate the association rules. Then, the least amount of transactions for modifications are selected. The minimized transactions are processed by a cuckoo optimization algorithm where an item is removed and reinserted in the transactions based on the objective parameters of the transactions. Thus the delicate rules in the transaction database are unseen with minimal hiding failure and lost rules on non-sensitive rules. The investigational results prove that the recommended QPICOA has improved hiding failure and lost rules on adult, bank marketing and hardware store sales datasets.

## V. FUTURE SCOPE

In future the variable limits are adjusted dynamically based on the fitness value of each cuckoo in the population. This would increase the convergence rate of the sanitized database and increase the efficiency of optimization further.

## REFERENCES

[1]. Refaat, M., Aboelseoud, H., Shafee, K., & Badr, M. (2016). Privacy preserving association rule hiding techniques: Current research challenges. *International Journal of Computer Applications*, *136*(6), 11-17.

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

[3]. Patel, B. P., Gupta, N., Karn, R. K., & Rana, Y. K. (2011). Optimization of Association Rules Mining Apriori Algorithm Based on ACO. *International Journal on Emerging Technologies, 2*(1), 87-92.

[4]. Gulwani, P. (2012). A novel approach for association rule hiding. *International Journal of Advance Innovations, Thoughts & Ideas*, *1*(3), 1-9.

[5]. Quoc Le, H., Arch-Int, S., & Arch-Int, N. (2013). Association rule hiding based on intersection lattice. *Mathematical Problems in Engineering*, *2013,* pp. *1-11*.

[6]. 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.

[7]. Shahsavari, A., & Hosseinzadeh, S. (2014). CISA and FISA: efficient algorithms for hiding association rules based on consequent and full item sensitivities. *In 7th International Symposium on Telecommunications (IST'2014) IEEE. 978*(1), 977-982.

[8]. Femandes, M., & Gomes, J. (2017). Heuristic approach for association rule hiding using ECLAT. In *2017 2nd International conference on communication systems, computing and IT applications (CSCITA)* (pp. 218-223).

[9]. Talebi, B., & Dehkordi, M. N. (2018). Sensitive association rules hiding using electromagnetic field optimization algorithm. *Expert Systems with Applications*, *114*, 155-172.

[10]. 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.

[11]. Mohammed, R. A., & Duaimi, M. G. (2018). Association rules mining using cuckoo search algorithm. *International Journal of Data Mining, Modelling and Management, 10*(1), 73-88.

[12]. Chhatrapati, M., & Serasiya, S. (2015).A Research on Privacy Preserving Data Mining Using Heuristic Approach. *International Journal of Computer Science and Mobile Computing (IJCSMC), 4*(5), 349-357.

[13]. Singhal, M. N., & Richariya, V. (2011). An Efficient Association Rule Mining without Pre-assign Weight. *International Journal on Emerging Technologies*, *2*(2), 18-20.

[14]. Jain, N., & Sharma, V. (2012). Distance Weight Optimization of Association Rule Mining with Improved Genetic Algorithm. *International Journal of Electrical, Electronics and Computer Engineering, 1*(2), 25-27.

[15]. Prabha, K., & Suganya, T. (2016). A Pandect on Association Rule Hiding Techniques. *International Journal on Emerging Technologies(Special Issue on ICRIET), 7*(2), 65-69.
[16]. Parmar, U. S, Motwani, A., & Shrivastava, A. (2015). Association Rule Mining- Various ways: A Comprehensive Study. *International Journal of Electrical, Electronics and Computer Engineering, 4*(2), 134-138.
[17]. Shrivastava, J., & Shrivastava, N. (2014). A Review of Data Reduction/Extraction in Data Mining from the Large set of Database. *International Journal of Electrical, Electronics and Computer Engineering, 3*(2), 149-153.
[18]. Kumbhare, T. A., & Chobe, V. (2014). An Overview of Association Rule Mining Algorithms. *International Journal of Computer Science and Information Technologies, 5*(1), 927-930.
[19]. Gayathri, P., & Poorna, B. (2017). Association Rule Hiding for Privacy Preserving Data: A Survey on Algorithmic Classifications. *International Journal of Applied Engineering Research, 12*(23), 13917-13926.
[20]. Verykios, V. S. (2013). Association rule hiding methods. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 3*(1), 28-36.
[21]. 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.

**How to cite this article:** Bhavani, G. and Sivakumari, S. (2019). Quality Preserving ICOA for ARH. *International Journal on Emerging Technologies*, *10*(4): 472–477.