Association Rule Hiding using Artificial Bee Colony Algorithm

S.Vijayarani
Assistant Professor
Department of Computer Science
School of Computer Science and Engineering
Bharathiar University
Coimbatore, Tamil Nadu, India

M.Sathiya Prabha
Research scholar
Department of Computer Science
School of Computer Science and Engineering
Bharathiar University
Coimbatore, Tamil Nadu, India

ABSTRACT

Data mining is the process of extracting the previously unknown patterns from large amount of data. Privacy preserving data mining is one of the research areas in data mining. The main objective of privacy preserving data mining is to provide the privacy for personally identifiable information in the datasets. Many privacy preserving techniques are used for protecting the confidential data items. Some of them are Privacy preserving Association Rule Mining, Privacy Preserving Clustering, Privacy Preserving Classification, Statistical disclosure control, K-anonymity etc. In this research paper, we have discussed about the association rule hiding problem. Association rule mining is one of the very important data mining techniques. The process of discovering itemsets that frequently co-occur in a transactional database so as to produce significant association rules that hold for the data is known as Association rule mining. Association rule hiding is the process of modifying the original database by hiding the sensitive data to protect the sensitive association rules. In this paper, we have proposed Artificial Bee Colony optimization algorithm for hiding the sensitive association rules. We analyze the efficiency of the Artificial Bee Colony optimization technique by using various performance factors.

General Terms

Data mining, Privacy, Security, Artificial Bee Colony algorithm.

Keywords

Privacy, Association Rule, Sensitive item, Modification, Artificial Bee Colony algorithm.

1. INTRODUCTION

Privacy Preserving Data Mining is a popular research area in data mining. The aim of the privacy preserving data mining is ensuring individual privacy while maintaining the ability of data mining techniques which develops algorithms for modifying sensitive data. In this private data and private knowledge remains private even after the mining process. Privacy preservation is primarily concerned with protecting against disclosure individual data records. In recent years, privacy preserving data mining place a vital role in data mining. The data mining applications has a great deal with health care, security, financial, behavioral, and other types of data. Different types of data mining techniques are artificial neural networks, rule induction, logistic regression, association rule, data visualization and so on. Data mining is widely used in marketing, detection of fraudulent activity, and scientific research in companies and other areas. In business, the data mining applications are customer relationship management in call centre, human-resources departments in identifying the characteristics of the employees, etc.

Privacy preserving data mining become an increasingly important issue in many data mining applications. It is a research area in both public and private sector. Currently, the research is mainly concentrated in the development of technical methods such as application of cryptography. It is also directed towards the development of specialized algorithms to meet security and privacy requirements for different data mining methods, such as classification. For example, in medical research privacy preserving data mining application is documented to protect patient’s privacy.

Association rule mining is one of the very interesting and important techniques of data mining. Association rule mining is the process of discovering itemsets that frequently co-occur in a transactional database so as to produce significant association rules that hold for the data [1]. An association rule consists of two parts, they are antecedent and consequent. The antecedent is also known as the Left Hand Side (LHS) is the part on the left of the arrow of the rule and the Right Hand Side (RHS) is the part on the right of the arrow of the rule. Two metrics are of the association rule mining are support and confidence. In privacy preserving data mining, Association rule hiding is a challenging research problem. There are several algorithms that are used for generating association rules. Some of them are apriori algorithm, partition algorithm, pincer-search algorithm, dynamic item set counting algorithm, fp-tree growth algorithm, Eclat algorithm and Dclat algorithm etc.

To preserve privacy in data mining a number of miscellaneous methods such as association rule mining and query processing is found. This problem is related to the disclosure control in statistical databases. The advances in data mining methods provide increasingly sophisticated methods for adversaries to make inferences about the behavior of the underlying data. The association rules may represent sensitive information for target marketing purposes in case of sharing the commercial data which needs to protection from inference. It refers to the impact of changing the database in the same method particular sensitive association rules terminates with out critically impacting the data and non-sensitive rules [2]. The three types of association rule hiding algorithms are 1) Heuristic approaches, 2) Border-based approaches and 3) Exact approaches. The Heuristic approaches are used to modify the selected transactions from the database for hiding the sensitive data. The Border-based approaches is the sensitive rule hiding can be done through the modification of the original borders in the lattice of the frequent and the infrequent patterns in the data set. The Exact approaches are non-heuristic algorithms which envisage the hiding process as a constraint satisfaction problem that may be solved using integer programming or linear programming.

This paper is organized as follows; Association rule hiding and the related works are discussed in Section 2. The general problem formulation and the basic definitions of association rule mining are discussed in Section 3. In Section 4, the proposed Artificial Bee Colony optimization technique for sensitive item modification is given. Section 5 gives the experimental results of the proposed technique and the efficiency of Artificial Bee Colony algorithm. Finally, Section 6 provides the Conclusion and Future work.

2. RELATED WORKS

In the data mining application, the privacy preservation have important role. It provides security for sensitive data. It is used to develop algorithms and techniques for extracting knowledge from large amounts of data while protecting sensitive information. The analyzed data is called knowledge. One type of data mining method is used discovering these data is known as knowledge hiding. 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 we differentiate hiding algorithms into single rule and multiple rule schemes. 4) The nature of the hiding algorithm, which can be either heuristic or exact [2].

In the paper "An Efficient Association Rule Hiding Algorithm for Privacy Preserving Data Mining [8]"proposes the association rule hiding algorithms to hide some generated association rules, by increase or decrease the support or the confidence of the rules[3].We observed from this paper Increase Support of Left Hand Side (ISL) algorithm is working only for modification of LHS and Decrease Support of Right Hand Side (DSR) algorithm works for the modification of RHS, but they hide the rules in less number of modification by increasing and decreasing the support of the LHS and RHS item of the rule. The hiding of association rule, $X \rightarrow Y$ was done by either decrease its support or its confidence by smaller than user-specified minimum support transaction (MST) and minimum confidence transaction (MCT). By decreasing the confidence of a rule either (1) by increasing the support of X in the left hand side of the rule, or (2) by decreasing the support of the item set $X \cup Y$.

The paper "Security Information Hiding in Data Mining on the bases of Privacy Preserving Technique [11]"proposes an modified support for Privacy Preserving Data Mining that ensures that the mining process will not break privacy up to a certain degree of security. They consider the data interference control problem and give the security information hiding and privacy by using Unified Modelling Language (UML) model. The Snooping results are covering insecure information that is sensitive knowledge and individuals' privacy.

The target of the reconstruction based approach is reconstructing the original data base by using only supporting transaction of non-sensitive rules. The paper, "Reconstruction-Based Association Rule Hiding[9]" is provide an easily controllable and robust association rule hiding secure mechanism in privacy preserving data sharing context. The FP-tree based method for inverse frequent set mining is used for reconstruction-based framework.

In the paper, "Privacy Preserving Fuzzy Association Rules Hiding in Quantitative Data [7]"proposes a fuzzy association rules hiding algorithm for hiding rules discovered from a quantitative database. They use Apriori mining algorithm to find association rules and hiding these rules using privacy preserving techniques. Decreasing the support value of item in either Left Hand Side (L.H.S.) or Right Hand Side (R.H.S) of the rule is used for hiding the rules.

The paper "A Novel Method for Privacy Preserving in Association Rule Mining Based on Genetic Algorithms [10]" introduces new multi-objective method for hiding sensitive association rules based on the genetic algorithm concept. It provides security of database and keeping the utility and certainty of mined rules at highest level.

The paper "Tabu Search based Association Rule Hiding [12]" provides association rule hiding techniques based on tabu search optimization techniques. They uses heuristic approach particularly distortion process for hiding sensitive into non-sensitive items. Finally, all sensitive rules are hidden, non-sensitive rules are protected, no fake rules are produced and efficiency of the algorithm is high.

3. PROBLEM FORMULATION

Consider a database D, consisting of N transactions, m items and thresholds minfreq and minconf set by the owner of the data. After performing association rule mining in During thresholds minfreq and minconf, we get a set of association rules, denoted as A_R, among which a subset S_R of A_R contains rules which are considered to be sensitive from the owner's perspective. Given the set of sensitive association rules S_R, 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 A_R from disclosure, while minimally affecting the non-sensitive rules existing in A_R.In this work, we 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 rules are created by applying Eclat algorithm to D with minimum support and minimum confidence value. The items presented in the sensitive rules are considered as sensitive items and number of modification needed to hide these items by using Artificial Bee colony algorithm to the original transactional database (D) and we get modified database (D'). Finally the modified database (D') is subjected to Eclat algorithm and we have found that there is no occurrence of sensitive association rules and fake rules. The non sensitive rules are protected and the hiding failure is null.

4. PROPOSED SOLUTION

The following steps are required for the proposed solution.
Step 1: Consider a transactional database with set of items and transactions.
Step 2: Eclat algorithm is used to find the frequent itemsets based on the minimum support threshold.
Step 3: The set of association rules can be generated based on the minimum support and minimum confidence thresholds from the frequent itemsets.
Step 4: Select the sensitive items from the set of association rules.
Step 5: Artificial Bee Colony algorithm is used for modifying the sensitive items.
Step 6: Repeat the steps 2 in the modified data set
Step7: Verify (i) the hiding failure, (ii) misses cost, (iii) dissimilarity (iv) efficiency

4.1. Eclat Algorithm

The algorithm called Eclat (Equivalence Class Transformation) was introduced by Zaki (2000). ECLAT algorithm is used for generating association rules. It employs prefix based classes to reduce the search space [5]. Eclat was using the power of a vertical database layout and intersection based counting mechanism along with a depth-first search strategy and a bottom–up search for finding frequent patterns from a database[1].The Eclat algorithm is used to perform itemset mining. The basic idea for the Eclat algorithm is use tidset intersections to compute the support of a candidate itemset avoiding the generation of subsets that does not exist in the prefix tree[6]. The advantage of Eclat algorithm is much quicker than the Apriori algorithm. It is a depth-first algorithm. It has lower memory consumption than the Apriori algorithm [5]. Recursive process is eliminated because of the single scan of database. The data structures are processor-cache friendly that means it is just an array of numbers.

Items in each itemset are kept sorted in their lexicographic order. The Eclat algorithm recursively merges discovered frequent itemsets and uses "tidlists" to evaluate support. When two frequent (k-1)-itemsets are merged to form a candidate k-itemset, their "tidlists" are intersected to form the "tidlist" of the new candidate. Association rule or sensitive rules are generated from the frequent item sets if the confidence threshold value is greater than the minimum threshold value. The frequent itemset generation phase of the Eclat algorithm is given below. In this work distortion technique is used for hiding the sensitive rules. This is replacing the values from 1 to 0 and vice versa.

Table 1: Eclat Algorithm

```
Eclat ($S_{k-1}$):
for all itemsets $I_a, I_b \in S_{k-1}$, a < b do begin
    C=$I_a$.tidlist ∩ $I_b$.tidlist;
    if (|C| ≥ minsup)
        add C to Lk
end
Partition $L_k$ into prefix-based (k-1)-length prefix classes
for each class $S_k$ in $L_k$
    Eclat ($S_k$);
end
Answer = $\cup_k L_k$;
```

4.2 Optimization techniques

Optimization techniques place a vital role in the field of privacy preserving data mining. This technique determines how privacy is preserved in various tasks. An efficient and accurate solution depends not only on the size of the database but also on characteristics of the sensitive rules and items. There are various optimization algorithms available to provide optimal solution. Some of these are Bees algorithm, genetic algorithm, particle swam optimization, etc. In this research work, we consider the Artificial Bee Colony optimization algorithm.

4.3. Artificial bee colony algorithm

Artificial Bee Colony Algorithm (ABC) is an optimization algorithm based on the intelligent foraging behavior of honey bee swarm, proposed by Karaboga in 2005. ABC tries to model natural behavior of real honey bees in food foraging. Honey bees use several mechanisms like waggle dance to optimally locate food sources and to search new ones. This makes them a good candidate for developing new intelligent search algorithms. In the ABC algorithm, the colony of artificial bees contains three groups of bees: employed bees, onlookers and scouts. Each cycle of the search consists of three steps:

- Placing the employed bees onto the food sources and then calculating their nectar amounts.
- Selecting the food sources by the onlookers after sharing the information of employed bees and determining the nectar amount of the foods.
- Determining the scout bees and placing them onto the randomly determined food sources. In the ABC, a food source position represents a possible solution to the problem to be optimized and the nectar amount of a food source corresponds to the quality (fitness) of the associated solution.

Let us consider original data set (D) that contains connect items. In the first step, all the transactions and items are initialized. In the second step, we initialized the sensitive items and number of modification required for the sensitive items. In the third step, the nectar amount, fitness function and probability are calculated for finding best food source.

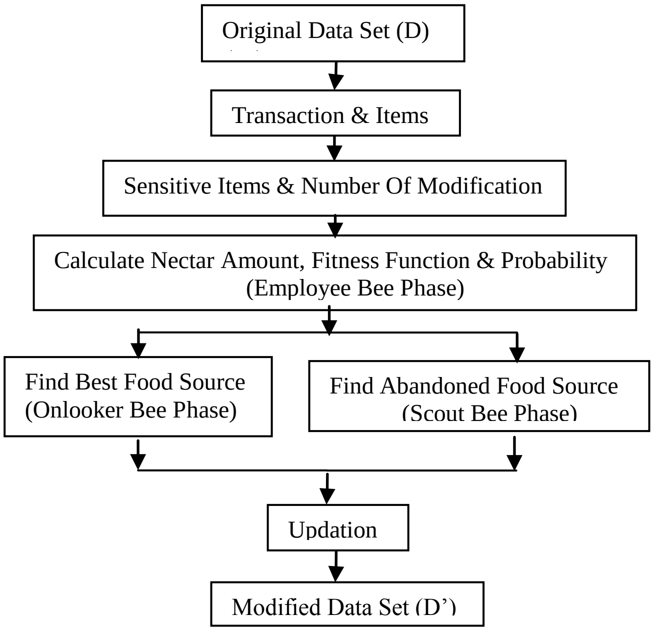
Original Data Set (D)
Transaction & Items
Sensitive Items & Number Of Modification
Calculate Nectar Amount, Fitness Function & Probability (Employee Bee Phase)
Find Best Food Source (Onlooker Bee Phase)
Find Abandoned Food Source (Scout Bee Phase)
Updation
Modified Data Set (D')

Fig 1: Artificial Bee Colony Optimization

Table 2: Artificial Bee Colony Algorithm for Sensitive Item Modification

Step1: Initialize the transaction, items, sensitive items and number of modification required for each sensitive item.
Step2: Cycle=1
Step3: Repeat
Step4: Consider the optimization problem for finding cost for the employed bees using the formula,
Cost(f_i)=(No., of item present)^2+(No., of sensitive item)^2 and evaluate them.
Step5: Calculate the probability values Pi for the solutions xi by means of their fitness
values using the equation (1)
$P_i = fit_i / \sum fit_i$ ------------ ----→(1)
In order to calculate the fitness values of solutions we employed the following
equation (eq. 2):
$Fit_i = 1/(1+f_i)$ if $f_i>=0$ ----------------→(2)
$1+abs(f_i)$ if $f_i<0$
Normalize Pi values into [0,1]
Step6: Apply the greedy selection process for the onlookers using P_i,Select highest P_i and corresponding transaction then applying distortion. Then, apply greedy process until there is no one presented.
Step7: Repeat the step8 until we find the item value as 1.
Step8: Determine abandoned solution, if exists. Then select the maximum P_i and replace1 as 0 if presented.
Step9: The original transaction present in the dataset is replaced by the modified transaction.
Step10: Cycle=Cycle+1
Step11: Until Cycle= Maximum Cycle Number (MCN) or modification=0.

The Artificial Bee Colony optimization algorithm for sensitive item modification is given in the above steps. Each food source is found depending on the value of probability and distance between the neighbour food sources. By applying greedy selection process, the best food is found depending on the probability value. After that the best food source ie., transaction which contains highest nectar amount is selected for modifying the sensitive item. The modification is done by replacing 1 as 0. This means changing sensitive item into non sensitive item. If abandoned food source is exists, select that transaction and replace 1 as 0 in sensitive item. Repeat these steps until modification becomes zero. In the terminating step contains no sensitive items and no modification required for that items. Finally, we got the modified data set which contains only non-sensitive items. After performing Artificial Bee Colony optimization algorithm, the Eclat algorithm is applied to the modified database for finding the frequent item sets and generates the sensitive rules from the database. In the modified data set, we ensure hiding failure was zero, all the sensitive rules are hidden, no false rules are generated and non-sensitive rules are not affected, efficiency of the algorithm is high.

5. EXPERIMENTAL EVALUATION

5.1. Dataset

Dataset contains the multiple transaction. Each transaction contains the set of items. Dataset is collected from the website fimi.ua.ac.be/data/connect. Various types of datasets are available in this website such as Mushroom, Chess, Connect, etc. In this research work, connect dataset is used; it contains 127 items and 67557 transactions. Items are considered as binary values such as 0's and 1's. If particular item is presented in a transaction that item have a value 1, otherwise 0.We have taken various transactions and items such that 1000 transaction and 23 items, 2000 transaction and 30 items, 3000 transaction and 40 items, 5000 transaction and 50 items.

5.2. Creation of Sensitive Rules

In the first phase work, we have taken various different data sets for generating the frequent itemsets using Eclat algorithm based on the minimum support value. After generating the frequent itemsets, the number of sensitive rules are generated by association rule algorithm (Eclat) with the various threshold values namely support10% and confidence 20%, support20% and confidence 40% and support30% and confidence 50% for each data sets. The items presented in the sensitive rules are considered as the sensitive items. To provide association rule hiding, these items are converted into non-sensitive items by it is subjected to number of modification which is done by applying artificial bee colony algorithm.

5.3. Modification of Sensitive Rules

In general, ABC algorithm consists of three steps for each cycle of the search. The first step is sending the employed bees onto their food sources and evaluating their nectar amounts. After sharing the nectar information of food sources, the selection of food source regions by the onlookers and evaluating the nectar amount of the food source is the second step. The final step is determining the scout bees and then sending them randomly onto possible new food sources.

In this work, the step of the search is initialize the transaction, items which contains both sensitive and non-sensitive items and number of modification required to convert sensitive items into non-sensitive items. The second step consists of calculating cost(nectar amount), fitness function and probability to find the best food source region for employed and onlooker bees. The final step is determining the abandoned food source for scout bees and randomly selecting the possible transaction depending on probability value and replace 1 as 0 in sensitive items.

Frequent items in the sensitive transaction are modified as infrequent. For every modification the number of modification of a sensitive item is reduced. Item having highest number of modification is selected first for modification than the other items. This process is continued until the number of modification should become 0. Finally applying the Eclat algorithm in the modified database, no frequent items are retrieved and also the sensitive rules are also hided from the data base. Considering the original data sets required modification only done through the algorithm. No extra modification is done and data is not lost. It ensures no false rules are generated.

5.4. Analysis of Results

In this section, the result of artificial bee colony optimization algorithm is analysed. The experimental results are analysed based on the following performance factors.

1. Hiding failure
2. Misses cost
3. Dissimilarity
4. Efficiency

5.4.1. Hiding Failure

The hiding failure is measured by the ratio between the number of sensitive rules in modified dataset and the number of sensitive rules in original data base. We have analysed that the hiding failure of this algorithm with various transactions and items. It is found that the hiding failure with ABC algorithm for the transactions is zero. The hiding failure is calculated by the following equation.

$$HF = \frac{R_P(D')}{R_P(D)} \text{-------------}\rightarrow (3)$$

Where $R_P(D')$ corresponds to the sensitive rules discovered in the modified dataset D', $R_P(D)$ to the sensitive rules appearing in the original dataset D. Various hiding failure results for different datasets is listed in above table

Table 3: Hiding Failure for Different Thresholds

Thresholds	Hiding Failure			
	1000T 23I	2000T 30I	3000T 40I	5000T 50I
Support 10% &confidence 30%	0	0	0	0
Support 20% &confidence 40%	0	0	0	0
Support 20% &confidence 50%	0	0	0	0

.

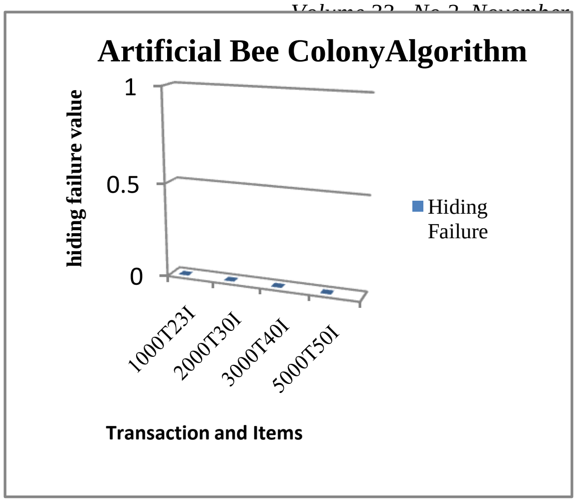

Fig 3: Hiding Failure

5.4.2 Misses Cost

This performance factor is used to measurethe percentage of the nonrestrictive patterns that are hidden as a side-effect of the modification process. It is computed as follows:

$$MC = \frac{|R_P(D)| - |R_P(D')}{|R_P(D)|} \quad \text{………………}> (4)$$

where $RP(\overline{D})$ is the set of all non-sensitive rules in the original database D and RP (D') is the set of all non-sensitive rules in the modified database D.

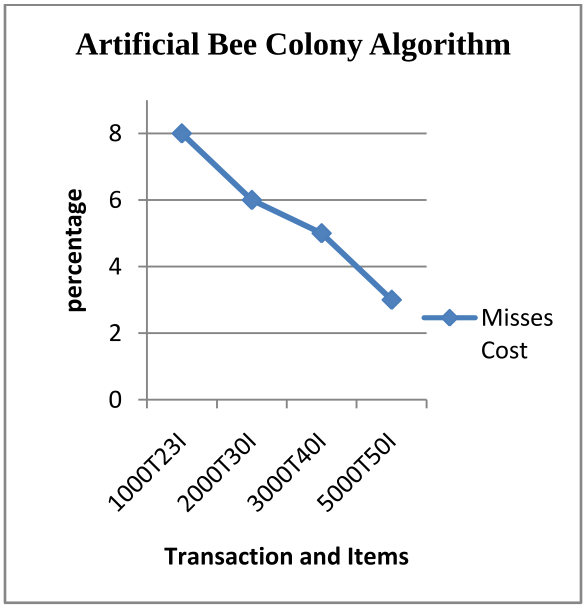

Fig: 4 Misses cost for Artificial Bee Colony algorithm

This analysis consists of data sets with four different sizes namely 23 items 1000 transactions, 30 items 2000 transactions, 40 items 3000 transactions and 50 items 5000transactions. The misses cost percentage for different data sets are null for all thresholds.

5.4.3. Dissimilarity

This performance factor qualifies the difference between original and modified datasets by drawing graphs. Here the horizontal axis contains items and the vertical axis contains their frequencies (number of transaction contains that particular item). The following chart shows the dissimilarity measures for the data set contains 1000 transaction and 23 items at the threshold value of support 30% and confidence 50%. In this data set, items 12, 14 and 21 are sensitive items.

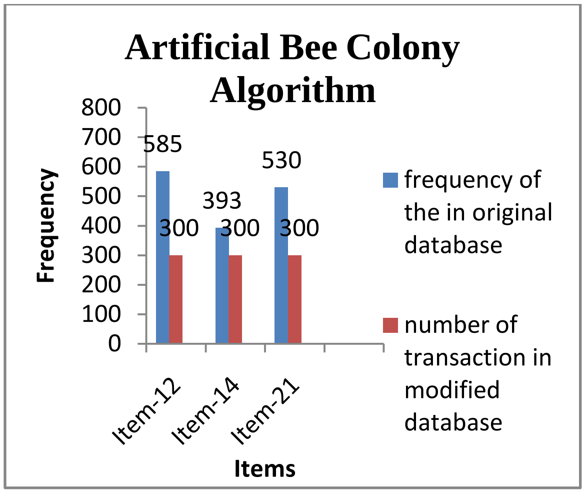

Fig 5: Dissimilarity between Items

5.4.4 Efficiency

In this work, the efficiency of the algorithm is calculated by using the CPU time. Here we are taking the data set which contains 1K, 2K, 3K, 5K transaction and 23,30,40,50 items respectively. The efficiency of the algorithm for various threshold values namely support 10% and confidence 30%, support 20% and confidence 40%, support 30% and confidence 50% is shown below.

Table 4: CPU Time at Different Thresholds

Thresholds	CPU time(In Seconds)			
	1000T 23I	2000T 30I	3000T 40I	5000T 50I
Support 10% and confidence 30%	36	86	164	978
Support 20% and confidence 40%	31	64	136	792
Support 30% and confidence 50%	5	23	48	360

The efficiency of the algorithm is shown in following figure (4).

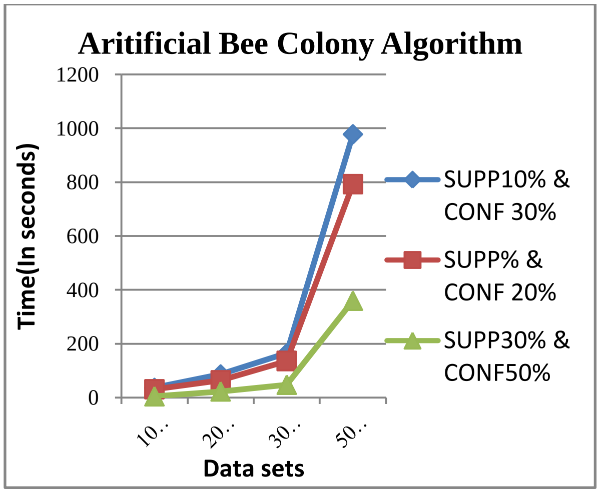

Fig 6: Efficiency of algorithm

6. CONCLUSIONS

Association rule hiding technique is the most important application in privacy preserving data mining. In this work, we have taken heuristic approach for hiding sensitive association rules by using artificial bee colony optimization algorithm. In this research work, we analyzed the four different performance factors with various data sets and threshold values. The experimental results show the hiding failure is zero, misses cost is null, dissimilarity measures and high efficiency. In future, our goal is to implement a new optimization technique to reduce the iterations.

7. REFERENCES

[1] Arun K Pujari, “Data mining techniques”, second edition, ISBN: 978 81 7371 672 0

[2] Agarwal CC. and Yu PS., “Privacy-preserving data mining: Model and Algorithms”,(editors) CharuC.Aggarwal and Philip S. Yu, ISBN: 0-387-70991-8, 2008.

[3] Assaf Schuster, Ran Wolff, BobiGilburd, “Privacy Preserving data mining on data Grids in the presence of Malicious Participants”, IEEE International Symposium on High Performance Distributed Computing - HPDC 2004.

[4] Christian Borgelt, “Efficient Implementations of Apriori and Eclat”, Workshop of Frequent Item Set Mining Implementations (FIMI 2003, Melbourne, FL, USA).

[5] Hahsler, M.; Buchta, C.; Gruen, B. &Hornik, K. “Arules: Mining Association Rules and Frequent Itemsets”,2009.

[6] R. Kessl SUI, “Frequent Substructure Mining- An Introduction”, 7. May 2009

[7] Manoj Gupta and R. C. Joshi, “Privacy Preserving Fuzzy Association Rules Hiding in Quantitative Data”, International Journal of Computer Theory and Engineering, Vol. 1, No. 4, October, 2009, 1793-8201

[8] Yogendra Kumar Jain, “An Efficient Association Rule HidingAlgorithm for Privacy Preserving Data Mining”, International Journal on Computer Science and Engineering (IJCSE)

[9] YuhongGuo,” Reconstruction-Based Association Rule Hiding”, Proceedings of SIGMOD2007 Ph.D. Workshop on Innovative Database Research 2007(IDAR2007), June 10, 2007, Beijing, China.

[10] Mohammad NaderiDehkordi , KambizBadie, Ahmad KhademZadeh , “A Novel Method for Privacy Preserving inAssociation Rule Mining Based on Genetic Algorithms”, journal of software, vol. 4, no. 6, august 2009

[11] VarunYadav, RichaJindal, “Security Information Hiding in Data Mining on the bases of Privacy Preserving Technique”, 2010 International Journal of Computer Application(0975-8887)Volume 1- No.15.

[12] S.Vijayarani,Dr. A.Tamilarasi, R.SeethaLakshmi, “Tabu Search based Association Rule hiding”, International Journal of Computer Application (0975-8887) Volume 19-No.1,April 2011.