BULGARIAN ACADEMY OF SCIENCES

CYBERNETICS AND INFORMATION TECHNOLOGIES • Volume **17**, No 3
Sofia • 2017 Print ISSN: 1311-9702; Online ISSN: 1314-4081
DOI: 10.1515/cait-2017-0032

Effective Gene Patterned Association Rule Hiding Algorithm for Privacy Preserving Data Mining on Transactional Database

Gayathiri P.[1], *B. Poorna*[2]

[1]*Department of Computer Science, Bharathiar University, Coimbatore 641 046, TamilNadu, India*
[2]*SSS Jain College for Women, T. Nagar, Chennai, TamilNadu, India*
E-mails: *gayathiri98@yahoo.com* *poornasundar@yahoo.com*

***Abstract*:** *Association Rule Hiding methodology is a privacy preserving data mining technique that sanitizes the original database by hide sensitive association rules generated from the transactional database. The side effect of association rules hiding technique is to hide certain rules that are not sensitive, failing to hide certain sensitive rules and generating false rules in the resulted database. This affects the privacy of the data and the utility of data mining results. In this paper, a method called Gene Patterned Association Rule Hiding* (*GPARH) is proposed for preserving privacy of the data and maintaining the data utility, based on data perturbation technique. Using gene selection operation, privacy linked hidden and exposed data items are mapped to the vector data items, thereby obtaining gene based data item. The performance of proposed GPARH is evaluated in terms of metrics such as number of sensitive rules generated, true positive privacy rate and execution time for selecting the sensitive rules by using Abalone and Taxi Service Trajectory datasets.*

***Keywords*:** *Association Rule Hiding, Data Mining, Gene Pattern, Transactional database, Multiplicative perturbation, Additive perturbation.*

1. Introduction

The privacy preserving data mining needs to ensure the sensitive information are hidden from unauthorized users. Association Rule Hiding technique in data mining is hide the sensitive association rules generated from the transactional database. The association rules hiding technique indirectly expose the other data items through false rules and hide certain data items, which are not sensitive and fail to hide certain sensitive rules, which in turn affect the privacy of data and affect the utility of the

data mining results. The aim of hiding sensitive rules should be made with minimal side effects and maximizes the data utility in the sanitized database.

A compact prelarge GA-based (cpGA2DT) algorithm to delete transactions for hiding sensitive item sets was presented in [1] using flexible fitness function with three adjustable weights aiming at minimizing the execution time and the side effects. Efficient Hiding of Sensitive Item sets based on Genetic Algorithms was used for optimizing the selected transactions to be deleted with reduced side effects. The preservation concept used with genetic algorithm, reduced rule hiding time for each transactional database. Besides, the cpGA2DT algorithm also reduced the population size at each evaluation of rule hiding with a probability distribution. Hiding sensitive item sets was performed with minimal side effects of hiding failure, missing cost, artificial cost and efficiency. However, predefined and missing item sets affect the rules being disclosed.

To Secure Association Rules, Secure Multi-party Computation (SMC) algorithm [2] was designed that hide the association rules in a horizontally distributed database. The SMC algorithm computed the union of private item subsets with which secured mining of association rules was done reducing the communication rounds, communication cost and computational cost. However, the SMC algorithm was unable to secure the transaction items and rules completely due to leakage of information from the side of the users handling the transactional database.

This paper highlights the investigation of the association rule hiding for maintain the privacy of transactional database using Gene Patterned Association Rule Hiding (GPARH) algorithm. The proposed Gene Patterned Association Rule Hiding algorithm is to evaluate the database for data items to be hidden and to be exposed from constructed vector element. In view of that data from applications, an algorithm for sensitive and non-sensitive rule identification and Gene Min-Max Mapping algorithm is proposed in the method. After that, Multiplicative and Additive Modified Association Rule generation algorithms employed to generate minimal side effects with high true positive rule privacy rate. In Multiplicative and Additive Modified Association Rule generation algorithm, two data items are selected and exchanged with each other in order to generate a new set of gene data item by using the mutation and crossover operation. Thus, GPARH model reduces the number of sensitive rules generated for hiding sensitive data with minimum size effects. The Experiments have been carried out on two data sets downloaded from UCI. The experimental results show that the proposed algorithm is effective and efficient. Experiments show that, compared with both the classical secured mining of association rules based on genetic algorithms and existing fast-distributed algorithms, the proposed algorithm can find a feasible association rule hiding method for privacy preservation of data in a much short time.

The objective of GPARH model is formulated as follows.

- To improve the privacy preservation of sensitive association rule hiding in transactional database, GPARH model is designed.
- To improve the true positive rate of privacy preservation of sensitive data, Gene Min-Max Mapping algorithm is developed in GPARH model.

- To minimize the number of sensitive rules generated for hiding to protect sensitive data with minimum size effects, Multiplicative and Additive Perturbation-based Modified Association Rules algorithm is designed in GPARH model.

The paper is structured as follows. In Section 2, the basic concepts in association rule hiding and genetic algorithm for hiding with related works are reviewed. Section 3, tells about principles and algorithms for rule hiding to preserve privacy using GPARH method are proposed. In Section 4, the experimental evaluation with two dataset is employed and the analysis of results is discussed in detail to demonstrate the effectiveness and efficiency of the algorithm and conclusion is given in Section 5.

2. Related works

Through data mining, users extract useful information that organizations do not want to disclose to the public. Therefore, several Privacy Preserving Data Mining (PPDM) techniques were employed to preserve such confidential information.

The increasing development in data mining techniques enables users to extract required knowledge from a large data collection. However, with the disclosure of sensitive data released to other users in an inappropriate manner poses severe threats. In [3], an evolutionary multi-objective optimization method was designed aiming at minimizing the side effects. In [4], sensitive rules were hidden using Evolutionary Multi-Objective (EMO) mechanisms with the objective of deleting identified transactions to hide sensitive rules. Another method based on EMO was designed in [5] to identify promising transactions to minimize side effects.

The increase in the growth of data mining techniques in recent years, meaningful information can be easily mined helping the decision-makers for efficient strategies. In [6], Hiding Missing Artificial Utility (HMAU) algorithm was employed to reduce the execution time and side effects. In [7], a GA based privacy preserving mining method was investigated to hide the sensitive high utility item sets using down closure property.

One of the most popular data mining techniques is association rule mining that discovers the interesting patterns from large transaction data. In [8], with the objective of preserving personalized privacy with high accuracy, a high-personalized data distortion model was designed. A method for hiding association rules with minimum changes in database was designed in [9] aiming at hiding sensitive rules in sparse database. Privacy problem was considered in [10] using association rule hiding algorithm.

In [11], a genetic algorithm was employed to counter the side effects for large datasets. In [12], the concept of impact factor using item lattice was employed to eradicate sensitive knowledge based on the intersection lattice. In order to ensure trust, perturbation-based privacy preserving data mining [13] model was designed to provide maximum flexibility to the data owners.

A novel technique called slicing [14] was investigated using generalization and bucketization to prevent membership disclosure. In [15], privacy preserving for

location-based query problems were investigated to introduce a security model. A privacy preserving access control mechanism [16] using heuristics for anonymization algorithm was designed to improve the precision rate.

Privacy for collaborative data publishing [17] using provider-aware anonymization algorithm was designed with the objective of ensuring high utility. Access control policies based on two layers of encryption to ensure data confidentiality and preserving the privacy of the user was provided in [18].

However, an important problem that has to be considered in public clouds is to find the measure for selectively sharing the data. This data sharing mechanism has to be designed in a way by providing fine-grained attribute based access control policies that not only assure data confidentiality but also privacy preserving of users. In [19], cryptographic techniques were investigated for addressing such problem and then present two approaches that address these drawbacks with different trade-offs. Taxonomy driven lumping for sequence mining was designed in [20] by applying Markov models for identifying trade-off between two conflicting goals, data probability maximization and complexity minimizing modeling.

A two-phase approach for retrieval of diverse and complimentary bundles were provided in [21] with the objective of solving complexity involved in mining. An Entropy based Attribute Privacy Preservation [EAPP] and Information Gain based Attribute Privacy Preservation [IGAPP] was designed in [22] for privacy preservation in multi trust level environment.

This paper develops an efficient Gene Patterned Association Rule Hiding (GPARH) algorithm based on three entropies, to explore the property of data perturbation for privacy preserving data mining in transactional database using association rule hiding. The key step of the development is that a vector element construction, we first introduce in this paper two types of rule identification based on gene property of hidden data item and expose data item. With the hidden and exposed data item identified, the gene selection operator based on mapping derives new genes being selected by making emphasis on good solutions and less emphasis on bad solutions. With these mechanisms, a Gene Min-Max Mapping algorithm is proposed for efficient mapping of corresponding vector item. After analysis of good and bad solutions, multiplicative and additive modified association rule generates modified association rule generation by exploring the mutation and crossover function. By doing so, inherent characteristics of gene comprising of hidden and exposed properties, privacy of transactional database is maintained with higher true positive rule privacy rate with minimum sensitive rules being generated in a short time.

3. Methodology

In this section, a Gene Patterned Association Rule Hiding (GPARH) algorithm for preserving association rule and therefore maintaining privacy of transactional database is presented. The rule hiding is done based on the data perturbation technique applied with the inherent characteristic of gene comprising of hidden and exposed properties. Fig. 1 shows the flow diagram of GPARH method.

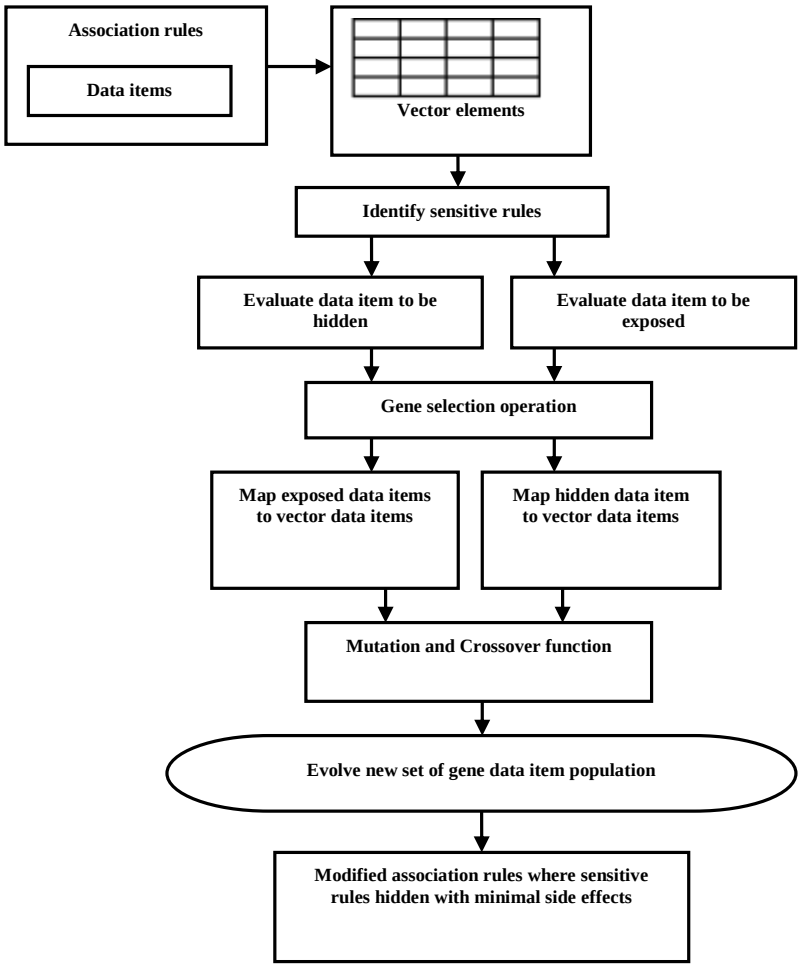

Fig. 1. Flow diagram of Gene Patterned Association Rule Hiding Construction of vector element

As shown in Fig. 1, the GPARH method is performed in three modules. The first module corresponds to the construction of vector element for generating initial sensitive rules for evaluating hidden and exposed data. The second module corresponds to the gene selection operation where these items are mapped to the corresponding data items. Finally, the third module performs Multiplicative and Additive Modified Association Rule generation with minimal side effects for hiding sensitive rules by applying mutation and crossover function. This in turn improves the performance of privacy preservation of sensitive rule hiding in transactional database.

Let us consider a Transaction Database (TD) with $I = \{I_1, I_2, \dots, I_n\}$ a set of data items purchase in a store. Then a transaction T is characterized by an ordered pair, $T = \langle \text{TID}, P \rangle$, where TID is a unique Transaction IDentifier and P represents the list of data items which the transaction comprises of. The absolute support of data item, P, is the number of transactions in TD that contains P. Fig. 2 shows the sample transactional database and its vector element representation.

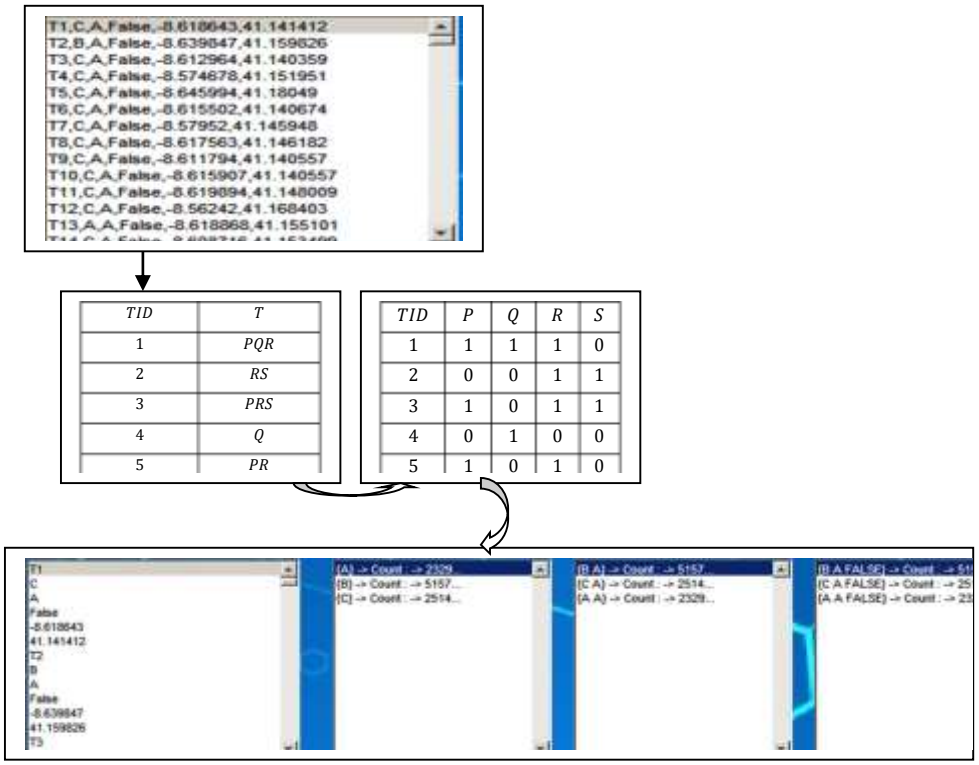

Fig. 2. Vector element representation of Transactional database

As shown in Fig. 2, a dataset is given as input and to this vector element representation is formed, "VE_{ij}, where i,j represents row and columns" is used to express a Transactional Database. Then, the relative support of P is the fraction of the transactions in a database which contain the data item P denoted as Sup (P). In the GPARH method, the data items P is said to be frequent if Sup (P) is greater than a Support Threshold (denoted as ST) by user:

(1) $$\text{Sup}(P) \geq \text{ST} \rightarrow \text{Frequent.}$$

In a similar manner, the confidence is relevant to association rules. A rule has the form of $P \rightarrow Q$, that means that antecedent P infers to the consequent Q, where both P and Q are data items with the Confidence Threshold (CT). Therefore, the measure of confidence using the GPARH method is as given below.

(2) $$\text{Conf}(P \rightarrow Q) = \frac{\text{Sup}(P \cup Q)}{\text{Sup}(P)},$$

(3) $$\text{Conf}(P \rightarrow Q) \geq \text{CT}.$$

Once the vector element representation is formed, the sensitive rules are identified with user defined criteria (ST and CT) through confidence count on the frequent item of the transactional database. Based on the sensitive rule and the privacy criteria for preserving the rule set, the data items to be exposed and hidden are evaluated:

(4) Sensitive Rule (Hidden Data Item) → More Cohesive (I),

(5) Non − sensitive Rule (Exposed Data Item) → Less Cohesive (I).

The rule comprising more cohesive data items is selected as sensitive rule. From (4), the data item obtained by applying sensitive rule is said to be hidden data item H_{ij}. On the other hand from (5), the data item obtained by applying non-sensitive rule is said to be exposed data item E_{ij} are evaluated. The threshold is set for the selection of sensitive rules with the convergence of cohesive items and divergence of the non-cohesive items. Algorithm 1 is designed for sensitive and non-sensitive rule identification.

Algorithm 1. Sensitive and Non-Sensitive Rule Identification

Input: Transaction database TD, set of data items $I = \{I_1, I_2, \dots, I_n\}$.

Output: Convergence of hidden and exposed data item.

Step 1. Begin

Step 2. For each data items

Step 3. Measure confidence value using (2)

Step 4. Measure hidden data item using (4)

Step 5. Measure exposed data item using (5)

Step 6. End for

Step 7. End

As shown in Algorithm 1, the confidence value is measured for each data item in the transaction database. Algorithm 1 finds the frequent data item present in the transaction database through confidence count. The data, which has high frequency value, is cohesive and frequency value below a threshold value is non-cohesive. For each data item, the algorithm 1 checks the sensitive and non-sensitive rule to decide the hidden and exposed data item to maintained privacy of transactional database.

3.1. Mapping-based gene selection operator

In this section, with the evaluated hidden data item H_{ij} and exposed data item E_{ij} for each transaction database TD, gene selection operator is applied for mapping it with the corresponding vector item. The gene population is filled with hidden and exposed data item characteristics. Each population comprises of several chromosomes with the best chromosome used to generate the next population. Fig. 3 shows the sample of exposed and hidden data items.

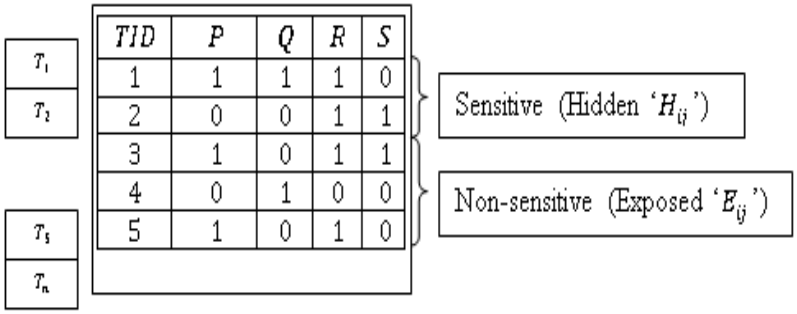

Fig. 3. Exposed and Hidden data items

As shown in Fig. 3, the columns correspond to the transactions $T_1, T_2, \ldots, T_5, \ldots, T_n$ and the rows correspond to the data items $I_1, I_2, \ldots, I_5$ respectively. As sensitive data items are limited to certain transaction, the GPARH method does not modify all of the transactions, the proposed mapping algorithm selects all the transactions that support sensitive items.

With this, the proposed mapping algorithm reach to better performance of obtaining true positive rate and less number of modification required during rule hiding process. Further, by evaluating and separating the exposed and hidden data items, the size of each chromosome decreases significantly. Followed by this, the gene selection is performed to emphasize good solutions (hiding sensitive rule) and less emphasize on bad solutions (exposing non-sensitive rules), while keeping the population size (i.e., data items) constant. This is performed using Min-Max Fitness function F as given below:

$$F = \sum_{i,j=1}^{m,n} \min\left(H_{ij} \cap E_{ij}\right) + \max\left(H_{ij} \cap E_{ij}\right). \quad (6)$$

From (6), the fitness function in the GPARH method is obtained by minimizing the hiding of sensitive rules and maximizing the exposure of non-sensitive rules. With gene selection operation, privacy linked hidden data items H_{ij} are mapped to the vector data items VE_{ij}. Then the exposed data items E_{ij} are also again mapped to the corresponding vector data item VE_{ij}.The algorithmic process of Gene Min-Max Mapping is shown in below.

Algorithm 2. Gene Min-Max Mapping

Input: Transaction Database, Transaction T, Hidden data item H_i, Exposed data item E_i.

Output: Improved true positive rule privacy rate.

Step 1. Begin
Step 2. **For** each transaction T in Transaction Database where $T \in$ TD
Step 3. Measure Min-Max Fitness function F using (6)
Step 4. Map hidden data items H_{ij} to vector data items VE_{ij}
Step 5. Map exposed data items E_{ij} to vector data items VE_{ij}
Step 6. **End for**
Step 7. End

As shown in Algorithm 2, for each transaction in transaction database, the Gene Min-Max Mapping algorithm performs mapping of vector data items to hidden and exposed data items. To do this, a Min-Max fitness function is evolved. This in turn improves the true positive rule privacy rate in an effective manner.

3.2. Multiplicative and Additive Perturbation-based Modified Association Rules

Finally, the gene based data item population is subjected to multiplicative and additive perturbation with mutation and cross over function to evolve new set of gene data item population. The new set of gene data item population is used to generate the modified association rules in which sensitive rules are hidden with minimal side effects. Fig. 4 shows the structure of Multiplicative and Additive Modified Association Rule. In the design, the GPARH method concentrate on the gene based

data item population (i.e., hidden data item) rather than exposed data items as a perturbation.

Let ΔA and ΔM denote the additive and multiplicative perturbation with mutation and crossover function. The additive and multiplicative function in the GPARH method adds and multiplies a random noise ε_{ij} to each original hidden data item H_{ij} with mutation and crossover to evolve new set of gene data item population. The crossover function applied in the GPARH method for the hidden data item $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D}, \boldsymbol{E}, F, G, H, I$ is given as below.

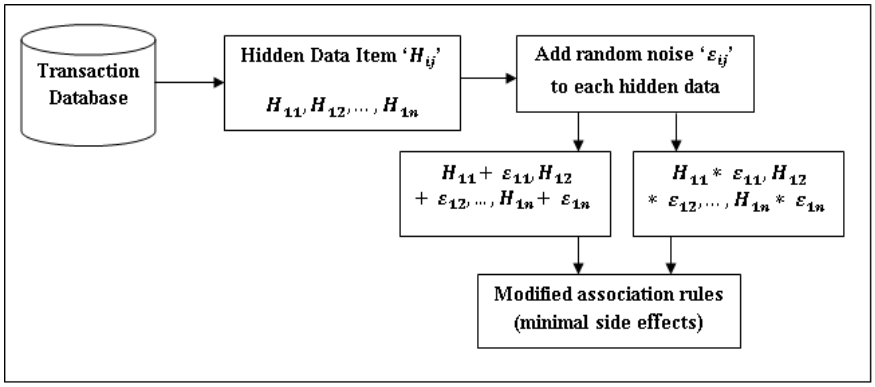

Fig. 4. Structure of Multiplicative and Additive Modified Association Rule generation

$$\Delta A = H_{ij} + \varepsilon_{ij}, \tag{7}$$

$$\Delta A = (\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D}, \boldsymbol{E}, F, G, H, I) + (\boldsymbol{D}, \boldsymbol{E}, \boldsymbol{C}, \boldsymbol{F}, \boldsymbol{H}, I, G, B, A) = \\ = (\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D}, \boldsymbol{E}, F, H, I, G). \tag{8}$$

From (7) and (8), the crossover function is designed in a way that single crossover point is selected, till this points, the data item is copied from the first parent (transaction), then the second parent (transaction) is scanned and if the data item is not in the offspring it is added.

The mutation function applied in the GPARH method is as given below:

$$\Delta M = H_{ij} * \varepsilon_{ij}, \tag{9}$$

$$\Delta M = (A, \boldsymbol{B}, C, E, F, \boldsymbol{H}, I, G) \rightarrow (A, \boldsymbol{H}, C, D, E, F, \boldsymbol{B}, I, G). \tag{10}$$

From (9) and (10), it is found that the mutation function is designed in such a manner that where two data items are selected (shown in bold face letter) and exchanged with each other to form new set of gene data item. In this manner, sensitive rules are generated with minimal side effects.

Algorithm 3. Multiplicative and Additive Perturbation based Modified Association Rule generation

Input: Transaction Database, Transaction T, Hidden data item H_i, Exposed data item E_i.

Output: Modified Association Rules (i.e., sensitive rules are generated with minimal side effects).

Step 1. Begin

Step 2. For each hidden data items H_{ij}

Step 3. Adds a random noise ε_{ij} to hidden data items through crossover functionusing (7) and (8)

Step 4. Multiplies a random noise ε_{ij} to hidden data items through mutation function using (9) and (10)

Step 5. Generate modified association rules

Step 6. End for

Step 7. End

As shown in Algorithm 3, for each hidden data items in transaction database, Multiplicative and Additive Perturbation based Modified Association Rule generation algorithm performs mutation and crossover operation. During the mutation and crossover operation, two data items are selected and exchanged with each other in order to generate a new set of gene data item. Therefore, GPARH method generates modified association rules by with aid of Multiplicative and Additive Perturbation based Modified Association Rule generation algorithm. Thus, the sensitive rules are generated with minimal side effects to hide the sensitive data from the transactional database. This in turn helps for GPARH method to improve the performance of privacy preservation for sensitive data hiding.

4. Experimental settings

This section deals with the performance of the proposed Gene Patterned Association Rule Hiding (GPARH) method in terms of the number of sensitive rules generated, true positive rule privacy rate, time for rule hiding, gene data item population and number of transaction is measured. The effectiveness of the GPARH method is studied by setting the following conditions: setting the number of rules to hide as constant, constant minimum support, and varying the number of transactions. In order to evaluate the performance of proposed, GPARH method is implemented using Java Languages.

The experiments were conducted using Abalone and Taxi Service Trajectory datasets. The Abalone dataset has seven continuous attributes among the 9 attributes and 4177 instances and Taxi Service Trajectory evaluation dataset comprising of 9 numerical attributes from UCI Machine Learning Repository. The experiment results are compared with two existing methods namely compact prelarge GA-based to Delete Transactions (cpGA2DT) [1] and Unifying lists of locally Frequent Item sets Kantarcioglu and Clifton (UNIFI-KC) [2]. The parameters and experimental results show that the GPARH algorithm is achieved better than the existing methods for association rule hiding and preserving the privacy of transactional database. The GPARH algorithm has three different experiments, such as finding number of sensitive rules, finding true positive sensitive rules and measuring execution time for selecting these rules.

4.1. Number of sensitive rules

The first experiment shows the relationship between numbers of sensitive rules arrived at applying the rule identification algorithm employing Abalone and Taxi Service Trajectory dataset with respect to different number of transactions. Different

number of transactions used in the experimentation varies in the range of 100 to 700 applied at seven different simulation runs. The numbers of sensitive rules are obtained by the difference between the total rules and the number of non-sensitive rules for each transaction,

(11) $$SR = [TR - NSR].$$

From (11), the Sensitive Rules (SR) is obtained using the Total Rules (TR), and the Non-Sensitive Rules (NSR), respectively. The efficiency of the method is measured based on minimum sensitive rules generated and is measured in terms of percentage (%). When the number of rules generated is lower, the method is said to be more efficient.

Table 1. Tabulation for number of sensitive rules

Number of transactions	Number of sensitive rules					
	Abalone dataset			Taxi Service Trajectory dataset		
	GPARH	cpGA2DT	UNIFI-KC	GPARH	cpGA2DT	UNIFI-KC
100	100	122	128	135	148	160
200	125	145	153	158	175	205
300	133	153	161	167	184	223
400	148	168	175	187	215	239
500	154	174	180	210	233	252
600	168	188	193	235	258	278
700	179	194	200	257	280	305

Table 1 shows the tabulation results for the sensitive rule generation using GPARH, cpGA2DT [1] and UNIFI-KC [2] respectively using Abalone and Taxi Service Trajectory dataset. From the table, it is observed that the number of sensitive rules generated using proposed GPARH model is lower when compared to other methods. Furthermore, the comparison results also suggest that the GPARH represents a new method for association rule hiding to preserve privacy on transactional database.

In this experiment, the number of rules generated for 100 transactions is 135 set. The result for number of sensitive rule generation using GPARH, cpGA2DT [1] and UNIFI-KC [2] is generated is depicted in Fig. 5.

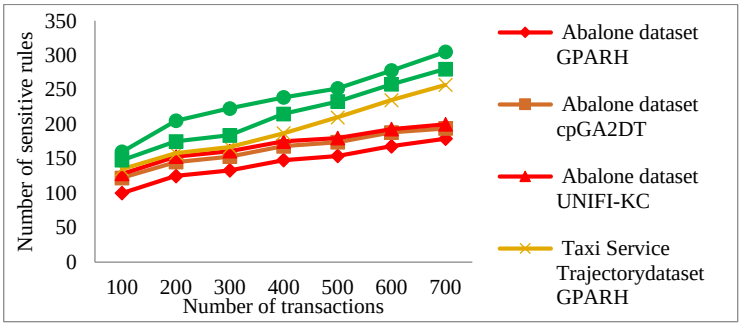

Fig. 5. Sensitive rule generation with respect to number of transactions using Abalone dataset and Taxi Service Trajectory dataset

Fig. 5 shows that number of sensitive rules generated after construction of vector element using sensitive and non-sensitive rule identification algorithm is lesser than the cpGA2DT [1] and UNIFI-KC [2] when tested with the Abalone dataset. From the Fig. 5, the red line depicts the performance of number of sensitive rules is generated while using Abalone dataset whereas green lines indicates the performance of number of sensitive rules generated while Taxi Service Trajectory dataset. Besides, while increasing the number of transactions, the number of sensitive rules is generated also is increased using all the three methods. However, comparatively the number of sensitive rules is generated using proposed GPARH model is lower. This is because the identification of sensitive rules is performed with user-defined criteria based on the value of confidence on frequent item of the transactional database. This in turn reduces the sensitive rule generation with respect to varying number of transactions using vector element representation. Moreover, with the convergence of cohesive items and divergence of the non-cohesive items, proposed GPARH model reduces the sensitive rules generated by 14.26% as compared to cpGA2DT and 19.02% as compared to UNIFI-KC respectively. In a similar manner when applied with Taxi Service Trajectory dataset, the number of sensitive rules generated using UNIFI-KC was improved by 11% when compared to cpGA2DT and 24% when compared to UNIFI-KC.

4.2. True positive rule privacy rate

In order to measure true positive rule privacy rate, sensitivity is evaluated to identify whether the hidden data items (i.e., hidden rules) are hidden and the exposed data items are exposed, measures the proportion of frequent items that are correctly identified as frequent item.

$$\text{TPR} = \frac{\text{HD correctly identified as HD}}{\text{HD correctly identified as HD+ ED incorrectly identified as HD}}. \quad (12)$$

From (12), the True Positive Rate (TPR) is obtained using the Hidden Data item HD and Exposed Data item ED. When the true positive rule privacy rate is higher, the method is said to be more efficient. The true positive rule privacy rate is measured in terms of percentage (%). To support transient performance, in Table we apply a Gene Min-Max Mapping algorithm and comparison made with two other existing methods cpGA2DT and UNIFI-KC.

Table 2. Tabulation for true positive rule privacy rate

Number of transactions	True positive rule privacy rate (%)					
	Abalone dataset			Taxi Service Trajectory dataset		
	GPARH	cpGA2DT	UNIFI-KC	GPARH	cpGA2DT	UNIFI-KC
100	94.35	71.28	66.35	90.12	68.45	61.25
200	91.14	69.23	64.14	88.65	65.64	58.16
300	88.21	67.33	62.25	85.98	64.35	56.92
400	86.32	65.14	60.14	83.16	62.15	55.16
500	89.21	70.23	65.45	84.92	68.12	53.62
600	93.14	74.19	69.31	91.16	70.91	59.85
700	95.88	78.21	75.16	93.85	77.15	64.92

The second experiment shows relation between the number of sensitive rules and the true positive rule privacy rate for varying number of transactions using Abalone and Taxi Service Trajectory dataset as shown in Table 2. The Abalone dataset contains seven continuous, one integer and one categorical attributes. Only the numerical attributes were considered for rule hiding to preserve privacy on transactional database. The number of instances is 4177. The number of sensitive rules obtained from GPARH was observed to be 100, cpGA2DT to be 122 and UNIFI-KC to be 128 respectively.

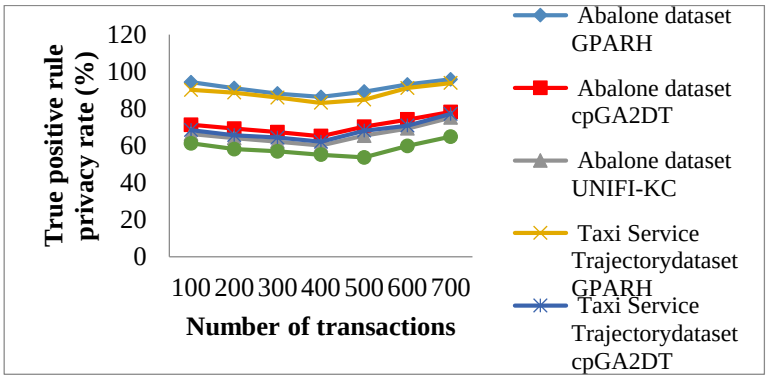

Fig. 6. True positive rule privacy rates with respect to differing number of transactions using Abalone and Taxi Service Figure Trajectory dataset

Fig. 6 depicts the relationship of the total number of transactions and the true positive rule privacy rate while keeping the population size (i.e., data items) constant. From the Fig. 6, the red line portrays the performance of true positive rule privacy rate while using Abalone dataset whereas green lines point outs the performance of true positive rule privacy rate while using Taxi Service Trajectory dataset. The true positive rule privacy rate generated by applying Gene Min-Max Mapping algorithm illustrate that if the number of transactions is increased, the number of true positive rate will not amplify. From Figure, we can see that by incorporating Mapping-based gene selection operator, GPARH method selects all the transactions that support sensitive items resulting in the improvement of true positive rule privacy rate. Moreover, GPARH method evaluates and separates the exposed and hidden data items by performing gene selection, paying emphasize to good solutions and less emphasize on bad solutions with the aid of Min-Max Fitness function. Therefore, proposed GPARH method improves the true positive rule privacy rate by 22% when compared to cpGA2DT and 27% when compared to UNIFI-KC respectively. Similarly, when using Taxi Service Trajectory dataset, proposed GPARH method improves the true positive rule privacy rate by 34% when compared to cpGA2DT and 23% when compared to UNIFI-KC respectively.

4.3. Execution time for selecting sensitive rule

The third experiment shows the time for selecting sensitive rule generated with minimum side effects for different number of modified associative rules. In Table 3 experimental results are reported with respect to modified associative rules for

abalone and Taxi Service Trajectory dataset. Fig. 7 shows the time for selecting sensitive rules when the number of modified associative rules was increased using Abalone and Taxi Service Trajectory dataset. As shown in the table, the time for selecting sensitive rules using Taxi Service Trajectory dataset was increased than using Abalone due to the increased number of attributes in Taxi Service Trajectory dataset.

While generating modified association rules with minimal side effects, the hidden rules, which are more cohesive, are said to be sensitive rule. On the other hand, the hidden rules other than cohesive are said to be non-sensitive rule. So, the execution time to generate the modified association rules in which sensitive rules are hidden with minimal side effects is mathematically formulated as given below.

(13) $$\mathrm{ET} = \text{No of modified associative rules} \times \mathrm{Time(CI)}.$$

From (13) the Execution Time (ET), for selecting sensitive rule is measured based on the number of modified associative rules generated and the time taken to extract the Cohesive Items Time(CI). When the execution time for selecting the sensitive rule is lower, the method is said to be more efficient.

Table 3. Tabulation of time for selecting sensitive rule

Number of modified associative rules	Time for selecting sensitive rule (ms)			Time for selecting sensitive rule (ms)		
	Abalone dataset			Taxi Service Trajectory dataset		
	GPARH	cpGA2DT	UNIFI-KC	GPARH	cpGA2DT	UNIFI-KC
10	0.6	0.75	0.92	1.02	1.31	1.52
20	0.91	1.12	1.26	1.53	1.92	2.35
30	1.6	1.81	1.95	2.27	2.8	2.96
40	2.22	2.71	2.98	3.15	3.4	3.65
50	3.44	3.67	3.82	3.98	4.18	4.45
60	3.91	4.08	4.25	4.55	4.88	5.15
70	4.63	4.85	5.05	5.31	5.62	6.02

Table 3 shows the measure of time for selecting sensitive rule with respect to modified associative rules generated using abalone and Taxi Service Trajectory dataset with the aid of three methods GPARH, cpGA2DT and UNIFI-KC respectively. From table value, it is clear that execution time for selecting sensitive rule using proposed GPARH model was lower as compared to other existing methods.

Fig. 7 shows the impact of execution time for selecting the sensitive rule based on different number of modified associative rule using three methods using Abalone dataset and Taxi Service Trajectory dataset. As exposed in figure, proposed GPARH model provides better execution time for selecting the sensitive rule when compared to existing methods cpGA2DT and UNIFI-KC. In addition, while increasing the number of modified association rules, the execution time for selecting the sensitive rule is also increased using all the three methods. However, comparatively the number execution time for selecting the sensitive rule using proposed GPARH model is lower. This is due to the application of Multiplicative and Additive Perturbation-based Modified Association Rules in GPARH method that provides a rational basis concentrate on the gene based data item population. This in turn reduces the time for

sensitive rule selection by 7.72% as compared to cpGA2DT when using Abalone dataset. In addition, with the application of mutation and crossover to evolve new set of gene data item population with respect to random noise, the time for sensitive rule selection is improved by 16.16% than compared to UNIFI-KC. In a similar manner, GPARH method reduces the time for sensitive rule selection by 24% as compared to cpGA2DT and 34% as compared to UNIFI-KC respectively while using the Taxi Service Trajectory dataset.

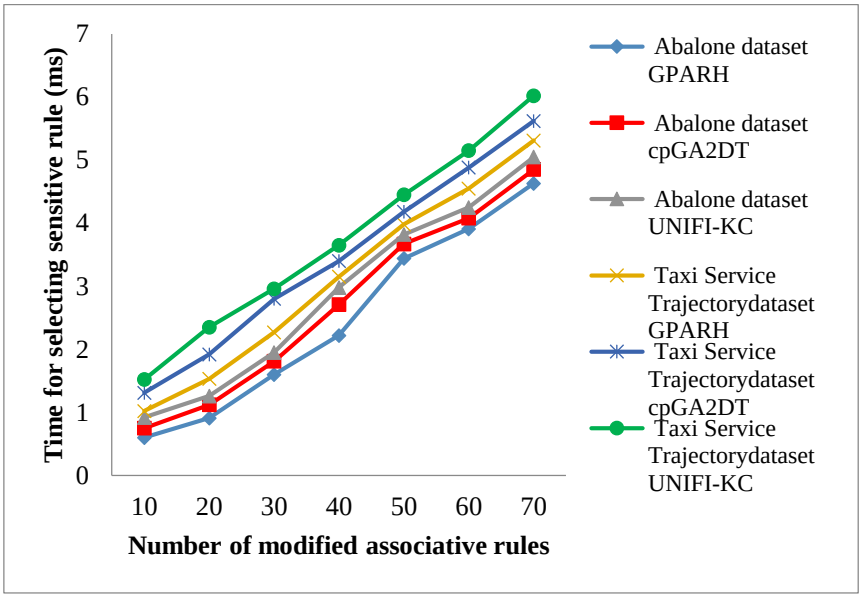

Fig. 7. Execution time for selecting sensitive rules with respect to modified associative rules using Abalone and Taxi Service Figure Trajectory dataset

5. Conclusion

To preserve the privacy of transactional database based on Gene Patterned Association Rule Hiding method with the scope of minimizing the side effects with the inherent characteristic of gene comprising of hidden and exposed proportions has been designed. The objective of providing such a design is to ensure effective preservation of association rule privacy and to decrease the processing time for sensitive rule generation. Construction of vector element is presented for the convergence of hidden and exposed data item as a measure for minimizing the number of sensitive rules using support and confidence value. Gene Min-Max Mapping algorithms also proposed to measure the true positive rule privacy rate for different instances. The proposed mapping algorithm with fitness function provides minimal side effects for different number of transactions. In addition, Multiplicative and Additive Modified Association Rule adding random noise to each data item helps in improving the true positive rule privacy rate. The experimental evaluation of GPARH method is conducted and the performance is measured in terms of time for sensitive rule selection and true positive rule privacy rate. The Performances results

reveal that the proposed GPARH method provides better performance with higher level of true positive rule privacy rate with minimal side effects; it reduces the execution time for selecting the sensitive rules when compared to state-of-the-art works.

References

1. Lin, C.-W., B. Zhang, K.-T. Yang, T.-P. Hong. Efficiently Hiding Sensitive Itemsets with Transaction Deletion Based on Genetic Algorithms. – Hindawi Publishing Corporation, The Scientific World Journal, Vol. **2014**, September 2014, Article ID 398269, pp. 1-13.
2. Tassa, T. Secure Mining of Association Rules in Horizontally Distributed Databases. – IEEE Transactions on Knowledge and Data Engineering, Vol. **26**, April 2014, Issue 4, pp. 970-983.
3. Cheng, P., J.-S. Pan. Use EMO to Protect Sensitive Knowledge in Association Rule Mining by Adding Items. – ACM, July 2014, pp. 65-66.
4. Cheng, P., J.-S. Pan. Completely Hide Sensitive Association Rules Using EMO by Deleting Transactions. – ACM, July 2014, pp. 167-168.
5. Cheng, P., C.-W. Lin, J.-S. Pan. Use HypE to Hide Association Rules by Adding Items. – PLOS ONE | DOI:10.1371/journal.pone.0127834, Vol. **10**, 12 June 2015, Issue 6, pp. 1-19.
6. Lin, C.-W., T.-P. Hong, H.-C. Hsu. Reducing Side Effects of Hiding Sensitive Itemsets in Privacy Preserving Data Mining. – Hindawi Publishing Corporation, The Scientific World Journal, Vol. **2014**, 2014, Article ID 235837, pp. 1-12.
7. Lin, C.-W., T.-P. Hong, J.-W. Wong, G.-C. Lan, W.-Y. Lin. A GA-Based Approach to Hide Sensitive High Utility Itemsets – Hindawi Publishing Corporation, The Scientific World Journal, Vol. **2014**, 2014, Article ID 804629, pp. 1-12.
8. Sun, C., Y. Fu, J. Zhou, H. Gao. Personalized Privacy-Preserving Frequent Itemset Mining Using Randomized Response. – Hindawi Publishing Corporation, The Scientific World Journal Vol. **2014**, 2014, Article ID 686151, pp. 1-10.
9. Sheykhinezhad, Z., M. Naderidehkordi, H. Rastegari. A Method for Hiding Association Rules with Minimum Changes in Database – ACSIJ Advances in Computer Science: An International Journal, Vol. **3**, September 2014, Issue 5, pp. 83-90.
10. Rao, K. S., N. M. Venkata, B. Debnath. An Association Rule Hiding Algorithm for Privacy Preserving Data Mining. – International Journal of Control and Automation, Vol. **7**, 2014, Issue 10, pp. 393-404.
11. Shah 1, R. A., S. Asghar. Privacy Preserving in Association Rules Using a Genetic Algorithm. – Turkish Journal of Electrical Engineering & Computer Sciences, Vol. **22**, February 2014, pp. 434-450.
12. Janakiramaiah, B., A. R. M. Reddy. Privacy Preserving Association Rule Mining by Concept of Impact Factor Using Item Lattice. – Wseas Transactions on Computers, Vol. **13**, June 2014, pp. 567-581.
13. Li, Y., M. Chen, Q. Li, W. Zhang. Enabling Multilevel Trust in Privacy Preserving Data Mining. – IEEE Transactions on Knowledge and Data Engineering, Vol. **24**, September 2012, Issue 9, pp. 1598-1612.
14. Li, T., N. Li, J. Zhang, I. Molloy. Slicing: A New Approach to Privacy Preserving Data Publishing. – IEEE Transactions on Knowledge and Data Engineering, Vol. **24**, March 2012, Issue 3, pp. 561-574.
15. Paulet, R., M. G. Kaosar, X. Yi, E. Bertino. Privacy-Preserving and Content-Protecting Location Based Queries. – IEEE Transactions on Knowledge and Data Engineering, Vol. **26**, May 2014, Issue 5, pp. 1200-1210.
16. Pervaiz, Z., W. G. Aref, A. Ghafoor, N. Prabhu. Accuracy-Constrained Privacy-Preserving Access Control Mechanism for Relational Data. – IEEE Transactions on Knowledge and Data Engineering, Vol. **26**, April 2014, Issue 4, pp. 795-807.
17. Goryczka, S., X. Li, B. C. M. Fung. m-Privacy for Collaborative Data Publishing. – IEEE Transactions of Knowledge and Data Engineering, Vol. **26**, October 2014, Issue 10, pp. 2520-2533.

18. N a b e e l, M., E. B e r t i n o. Privacy Preserving Delegated Access Control in Public Clouds. – IEEE Transactions on Knowledge and Data Engineering, Vol. **26**, September 2014, Issue 9, pp. 2268-2280.
19. N a b e e l, M., E. B e r t i n o. Privacy-Preserving Fine-Grained Access Control in Public Clouds. – IEEE Computer Society Technical Committee on Data Engineering, 2012, pp. 21-30.
20. B o n c h i, F., C. C a s t i l l o, D. D o n a t o, A. G i o n i s. Taxonomy-Driven Lumping for Sequence Mining. – Data Mining and Knowledge Discovery, Vol. **19**, October 2009, Issue 2, Springer, pp. 227-244.
21. A m e r-Y a h i a, S., F. B o n c h i, C. C a s t i l l o, E. F e u e r s t e i n, I. M e n d e z-D i a z, P. Z a b a l a. Composite Retrieval of Diverse and Complementary Bundles. – IEEE Transactions on Knowledge and Data Engineering, Vol. **26**, November 2014, Issue 11, pp. 2662-2675.
22. P r i y a d a r s i n i, R. P., M. L. V a l a r m a t h i, S. S i v a k u m a r i. Attribute Association Based Privacy Preservation for Multi Trust Level Environment. – Indian Academy of Science, Vol. **40**, September 2015, Issue 6, Springer, pp. 1769-1792.