ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

GRAVITATIONAL SEARCH ALGORITHM FOR EFFECTIVE SELECTION OF SENSITIVE ASSOCIATION RULES

GAYATHIRI P[1] , Dr. B POORNA[2]
Research Scholar,Department of Computer Science,Bharathiar University, Coimbatore 641 046, TamilNadu, India
[2]Principal, SSS Jain College for Women,T.Nagar, Chennai,TamilNadu, India
e-mail: gayathiri98@yahoo.com; poornasundar@yahoo.com

ABSTRACT

Association rule mining is the process of identifying the frequent items and associative rules in a market basket data analysis for large set of transactional databases. Association rules are employed in different data mining applications including web mining, intrusion detection and bioinformatics. In recent years it has been seen tremendous advances in the ability to perform effective association rule mining. It causes the need of sensitive rule selection to enhance the privacy preservation of data transactions. A new technique called Gravitational Search Based Sensitive Rule Selection (GS-SRS) is proposed to select the sensitive rules is to be hided for improving the privacy preservation of transactional database. The GS-SRS technique is introduced to select the sensitive rules from the derived association rule through conditional probability. The sensitive rules contain the sensitive information of transactional database. Sensitive rules are identified for many applications. One of the applications of sensitive rule identification is to preserve the privacy of an organization or an individual by hiding these rules. The GS-SRS technique initially generates association rules through identifying the frequent items by using conditional probability-based association rules and support count and confidence value. Next, GS-SRS technique used gravitational search algorithm that lists the cohesive and non-cohesive items for the given transactional database in shorter time than the conventional means of feature selection using Rough Set technique. The association rule containing more cohesive items is selected as sensitive rule. A threshold value is used to select sensitive rules with the convergence of cohesive items and divergence of the non-cohesive items. Finally, the sensitive rules selected are hided for preserving the privacy of transactional database. The experiments have been carried out on transaction database using four data sets and compared with state of art existing techniques. The experiment results show that the proposed GS-SRS technique is able to improve the accuracy of privacy preservation with minimum execution time when compared to state-of-the-art works.

Keywords: *Association Rule Mining, Gravitational Search, Sensitive Rule Selection, Conditional Probability, Cohesive Items, Non Cohesive Items.*

1. INTRODUCTION

In data mining, association rule mining technique identifies the frequent items and associative rules for a large size of transactional databases. The conditional probability of the transactional data items are evaluated to present the associative rule, which indicates the scenario of the buying habits and products in demand. Privacy preservation of transactional data has been considered as an important research problem in recent applications. However, disclosing the sensitive associative rule of the transactional data item may expose the confidentiality and privacy of the organizations and individuals. Identifying and selecting sensitive association rule of a transactional database is a crucial process for preserving the privacy of sensitive information.

Association Rule hiding technique hides the sensitive association rules generated from the transactional data items. The privacy preservation with data mining needs to ensure hiding sensitive information. Hiding sensitive rules should be made with minimal side effects. Sensitive association rule selection play major role in effectiveness of association rule hiding technique. But, the association rules hiding technique indirectly generates some data items which are not exist in original database and hide certain data items which are not sensitive, which in turn affects the privacy of rules and affects the utility of the data mining results. This problem is addressed by developing the new technique called Gravitational Search Based Sensitive Rule Selection (GS-SRS) to improve the selection of sensitive rules to enhance the Association Rule Hiding (ARH) technique using Gravitational Search Algorithm (GSA). This research work is an optimization process for selection of sensitive association rules from the possible set of association rule of a transactional database. The gravitational algorithm determines how appropriate association rules for hiding have been selected.

In this study, a Gravitational Search Based Sensitive Rule Selection (GS-SRS) technique is designed to select the sensitive rules is to be hided and thereby preserving the privacy of sensitive information's in business transactions. The identification of frequent items based on support count which results in the generation of association rules. The GS-SRS technique is a sensitive rule selection-based technique that efficiently reduces the execution time for sensitive rule selection, the rule are governs the organization's and individual's privacy. GS-SRS also incorporates gravitation search algorithm to select the sensitive rules from derived association rule, enabling reduced memory consumption for measuring cohesive and non-cohesive items. After that, the sensitive rules selected are hided with aim of preserving the privacy of transactional database. Experimental results shows that the gravitation search based technique can improve accuracy of privacy preservation over existing selection methods. The propose algorithm select optimum number of association rules, which has threat to privacy of data owners.

The remaining of this paper is organized as follows: Section 2 reviews related work in the fields of association rule hiding for privacy preserving data mining. Section 3 provides the necessary background along with the proposed sensitive rule selection technique with the aid of gravitational search algorithm. Section 4 contains the experimental evaluation and the detailed discussion provided in Section 5. Section 6 gives the conclusion.

2. RELATED WORKS

In recent trends, numerous researches have been done in association rule hiding for privacy preserving data mining. Group Incremental Approach using Rough Set (GIA-RS) [1] is used as a group incremental feature selection algorithm to identify the new feature subset in a short time, when multiple objectives were added to a decision table. Incremental feature selection algorithm was based on the information entropy and it dealt with effective and efficient mechanism. The feature selection algorithm generated more feasible subset in a short period of time with increasing number of data. However, the rules extracted from dynamic dataset need to be updated in time to improve the selection of sensitive rules.

Locality Sensitive Hiding for Privacy Preserving (LSH-PP) [2] with a Homomorphic Matching Technique identifies the candidate record pairs. The matching of pairs was implemented by a basic protocol which performs simple distance computations. Matching Technique is mostly used for Privacy-Preserving Record Linkage. The performances of the distance-preserving properties were highly correlated due to anonymization format. The parameters of blocking system in LSH-PP were selected in such a way to attain highest possible accuracy there by significantly reducing the possible running time. However, it failed to produce accurate results because of the anonymization format, failing to preserve the privacy of the individual.

Association rule hiding methodologies aim at sanitizing the original database [22]. This methodology makes all the sensitive rules disappear from the sanitized database, when the database is mined under the same or higher levels of support and confidence as the original database. Then all the non-sensitive rules that were mined from the original database should also be mined from its sanitized counterpart at the same or higher levels of confidence and support. At last no false rules also known as ghost rules should be produced when the sanitized database is mined at the same or higher levels of confidence and support. A false rule is an association rule that was not among the rules mined from the original database.

Sensitive association rule selection for association rule hiding has received considerable attention in recent years, especially in the context of transaction databases. For example, the work employed in [3] Locality Sensitive Hashing (LSH) was applied to reduce the data utility and execution time by proposing two novel anonymization methods for sparse high dimensional data. However, the time complexity involved in anonymization method remains unaddressed. In [4], it is shown that by applying Feature Relation Networks, using rule-based multivariate text feature selection, features were selected in a more computationally efficient manner. Despite, efficient in terms of computational effort, redundancy and relevancy remained unsolved. To address this issue, in [5], fuzzy approximation was used during rule extraction that concentrated on non-redundant and relevant features. Similarly, another method addressed in [6] that used accuracy constrained privacy preservation mechanism aiming at improving the imprecise bounds.

Association Rule Hiding of a transaction database has been identified as a significant problem and probabilistic data management received a lot of attentions to deal with uncertain data. An optimized Monte Carlo Algorithm [7] drastically minimized the number of iterations in probabilistic databases, however, little concentration were made towards attribute redundancy. To minimize the attribute redundancy, Fuzzy Rough Set [8] designed a rule-based classifier model that improved the sensitivity of rules being generated. However, focused remained unsolved, in differentiating the data between positive and negative samples. In [9], Probabilistic Latent Semantic Analysis (PLSA) was designed using minimum information from the user and made efficient use of the positive and unlabeled data by applying iterative algorithm. Since a better performance was achieved through PLSA and side effects was not discussed. Hiding Missing Artificial Utility (HMAU) [10] algorithm was applied to reduce the number of side effects and number of deleted transactions. Another method using randomized response technique [11] was applied to preserve the personalized privacy in frequent itemset mining.

Data mining techniques are used to mine useful information and knowledge from several databases. In [12], GA-based privacy preserving utility mining method was introduced to hide sensitive items. Yet, another GA-based approach was introduced in [13] for hiding sensitive information. However, with the introduction of smart phones and web services, privacy of metadata remained a major concern. To address this issue, in [14], a personal metadata management framework was introduced with the objective of reducing the execution time and dynamically protecting the personal data. Another Multi Objective Optimization (MOO) was designed in [15] that discovered useful relationships from shared data. This in turn minimized the confidence of sensitive rules.

Data publishing is extensively applied in the field of information sharing and scientific research, and providing security for user's privacy. Privacy protection method for multiple sensitive attributes was introduced in [16] to improve the sensitive rules. Though improvement was observed in sensitive rule generation, privacy preservation in distributed environment remained a major breakthrough. A novel algorithm for privacy preserving in data mining in distributed environment was applied in [17] using Elliptic Curve Cryptography(ECC) and Diffie Hellman key exchange. The application of ECC ensured privacy preservation and also minimized the performance time. A comprehensive review on privacy preserving data mining was studied in [18]. Though several methods were presented to hide sensitive association rules, the modification in database made certain amount of side effects. Genetic algorithm to preserve privacy was applied in [19] to preserve confidential information and counter side effects of lost rules. In [20], two fundamental approaches were designed with the aim of protecting sensitive rules from disclosure. Besides, it presented three strategies and five algorithms for hiding a group of association rules which was characterized as sensitive.

A novel technique was designed in [23] for privacy preserving mining of association rules from outsourced transaction a database. But, privacy vulnerabilities were remained unaddressed. A hybrid partial hiding algorithm (HPH) was presented in [24] for improving the privacy preservation of association rule mining. Privacy preserving association rule mining was developed in [25] for distributed databases with the application of genetic algorithm. A novel method was intended in [26] for secure mining of association rules in horizontally distributed databases. However, computational cost was higher. An accuracy-constrained privacy-preserving access control framework was designed in [27] to achieve privacy requirements in relational data. The association rule hiding technique was intended in [28] for hiding the sensitive data or information in transaction database.

Based on the aforementioned techniques and methods, in this work a gravitational search based sensitive rule selection technique is introduced with the objective of improving sensitive rules selection to enhance the business transactions.

3. DESIGN OF GRAVITATIONAL SEARCH BASED SENSITIVE RULE SELECTION

The Gravitational Search Based Sensitive Rule Selection (GS-SRS) technique is designed to select the sensitive rules is to be hidden and to enhance the privacy preservation of sensitive information in transactional database. This section presents two models to select sensitive rules to enhance association rule hiding and to enhance the privacy preservation of business transactions. The first model generates the association rules with support count and confidence value using Conditional Probability-based Association Rule generation. On the other hand, the second model selects sensitive rules from generated rules using Gravitational search based sensitive rule selection. The elaborate description of these two models is presented in the forthcoming sections.

3.1 Conditional Probability-based Association Rule Generation

Association rule mining [3] discovers items frequently occurring in a transactional database with the objective of producing significant association rules that hold for the data. In GS-SRS, a transactional database using association rule mining identifies frequent items with support count and generates association rules with confidence value based on conditional probability. To preserve the privacy of an individual or an organization, sensitive rules are identified in the transactional database. In this work, Conditional Probability-based Association Rule Generation is used to identify the frequent items in transactional database for efficient association rule generation. The following Figure 1 shows the process involved in Conditional Probability-based Association Rule generation.

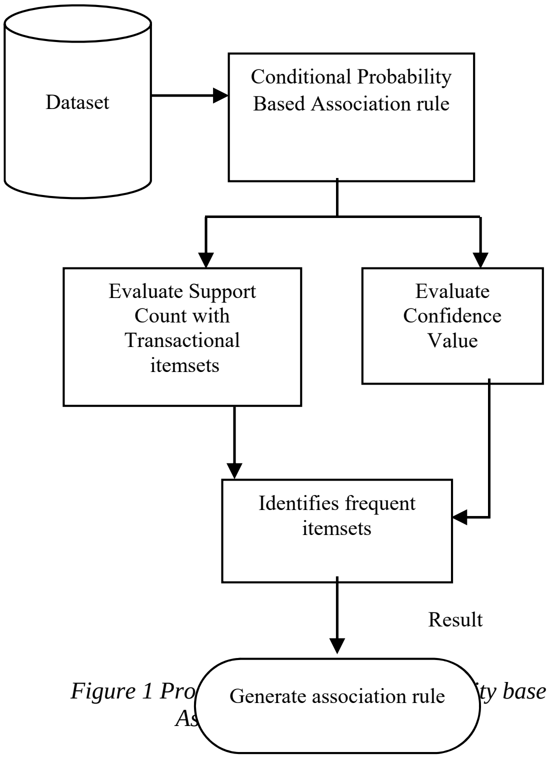
Dataset
Conditional Probability Based Association rule
Evaluate Support Count with Transactional itemsets
Evaluate Confidence Value
Identifies frequent itemsets
Result
Generate association rule

Figure 1 Pro ty based As

As shown in the Figure 1, Conditional Probability-based Association Rule Generation process initially takes Taxi Service Trajectory dataset as input and then evaluates the support count and confidence value with help of transactional item sets in the given data set. The frequent item sets are identified using evaluated support count values with the objective of generating the association rules. The GS-SRS technique uses conditional probability measure to estimate the support and confidence value and therefore identifies the frequent items. The conditional probability is a measure of probability of support and confidence value (for an itemset) given that another support and confidence (for another itemset) has occurred. Finally, it generates the association rule based on the identified frequent item sets with the aid of evaluated confidence value.

Let us consider a dataset, DS, a set of rules R over DS, as well as a sensitive rule SR. The objective of Conditional Probability-based Association Rule Generation is to identify a dataset DS such that when mining DS for generation of rules using same parameters as those used in the mining of DS, only the non-sensitive rules in $R - SR$ are derived. The selection of sensitive rule influences the analytical strategies which governs the effectiveness of association rule hiding technique.

Let $I = i_1, i_2, \dots . i_n$ denotes the set of items, for transaction $T = t_1, t_2, \dots . t_n$ where $T \in I$ with transaction id represented by TID, then the transaction T contains P, a set of items in P, if the following condition is said to be satisfied.

$$P \leq T \qquad (1)$$

The association rule is an implication and is formulated as given below

$$P \rightarrow Q, where\ P, Q < I\ and\ P \cap Q = \emptyset \qquad (2)$$

From the equation (2), P and Q denotes the items in the set of items I. The strength of the rule is measured using confidence and support value based on conditional probability. The support measures the frequency of the rule whereas the confidence measures the strength of the relation between item sets. The support of an association rule or probability of joint P and $Q$are mathematically formulated as given below.

$$Sup(P \rightarrow Q) = \left(\frac{Support\ count\ of\ P \cup Q}{Total\ number\ of\ records\ in\ DS}\right) \qquad (3)$$

From the equation (3), the support Sup is measured using the fraction of records that contain $P \cup Q$ to the total number of records in the database DS is said to be conditional probability (that involves both P and Q). On the other hand, the confidence value is measured in such a way that the transactions that contain P also contain Q and are mathematically formulated as given below.

$$Conf(P \rightarrow Q) = \left(\frac{Support\ count\ of\ P \cup Q}{Support\ count\ of\ P}\right) \qquad (4)$$

By using the equation (3) and (4), all the rules that satisfy the user-specified minimum support '$minsup$' and minimum confidence $minconf$ are retrieved. This in turn helps for efficient generation of association rules in a significant manner. The algorithmic description of Conditional Probability based Association Rule Generation is shown in below.

```
// Conditional Probability based Association Rule Generation Algorithm
Input: dataset '$DS$' Items '$I = i_1, i_2, \dots . i_n$' , Transaction '$T = t_1, t_2, \dots . t_n$', minimum support '$minsup = 1$', minimum confidence '$minconf = 10$'
Output: Efficient generation of frequent items (i.e.association rule generation)
Step 1: Begin
Step 2:    For each dataset '$DS$'
Step 3:       For each Transaction '$T$'
Step 4:          Sort all rules according to support value using (3)
Step 5:          If ( $Sup > minsup$) then
Step 6:             return $Sup$
Step 7:          End if
Step 8:          Sort all rules according to confidence value using (4)
Step 9:          If ($Conf > minconf$) then
Step 10:            return $Conf$
Step 11:         End if
Step 12:      End for
Step 13:   End for
Step 14: End
```

Algorithm 1 Conditional Probability Based Association Rule Generation

As shown in Algorithm 1, the Conditional Probability-based Association Rule generation algorithm consisting of three steps. For each dataset and transaction, all rules are first sorted according to the support value. Secondly, all rules are then sorted according to the confidence value in order to obtain the derived association rule through conditional probability. Finally, comparison of support and confidence value is made with the user-defined minimum support and minimum confidence to measure the strength of the rule. This in turn assists for efficient generation of association rule which resulting in minimized number of associative rules being generated.

3.2 Gravitational Search Algorithm for sensitive association rule selection

In GS-SRS technique, Gravitational Search Algorithm (GSA) is introduced to select the sensitive rules is to be hided from the derived association rule for enhancing the privacy preservation of transactional database. The idea of GSA came from the Newtonian laws of gravitation and motion [21], which says that all objects move due to the attraction with each other by gravitational forces. Therefore, objects with heavier mass have stronger attraction and move slower than the objects with relatively smaller mass.

Based on this fact, this work extended it by designing the gravitational search where data items with more cohesiveness has greater impact on the transactions rules being generated (i.e. mass interactions) than non- cohesive data items (i.e.

distance). Due to its fast convergence rate, the gravitational search is conducted to list the cohesive and non-cohesive items in each of the association rule generated for the given transactional database (i.e. search space). The rule comprising more cohesive items in turn is selected as sensitive rule is to be hided for improving privacy preservation of transactional database. The threshold is set for the selection of sensitive rules with the convergence of cohesive items and divergence of the non-cohesive items. The following Figure 3 shows the process of Gravitation Search model for sensitive rules selection.

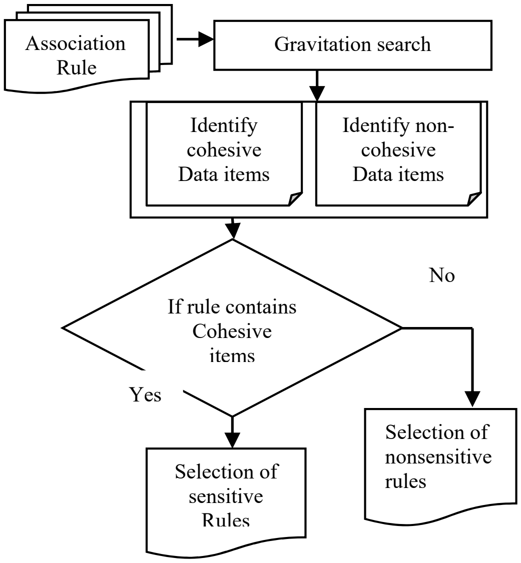
Association Rule
Gravitation search
Identify cohesive Data items
Identify non-cohesive Data items
If rule contains Cohesive items
No
Yes
Selection of sensitive Rules
Selection of nonsensitive rules

Figure 3 Process Of Gravitation Search Model For Selection Of Sensitive Rules

shown in the Figure3, Gravitation Search algorithm initially takes association rule as input. Then, this algorithm identifies cohesive data items and non cohesive data items (i.e. association rule). If association rule contains more cohesive items, then it selected as the sensitive rules otherwise it selects as non-sensitive rules. This efficient separation of cohesive and non cohesive data items helps to reduce the execution time for selecting sensitive rules. The Gravitational Search algorithm is based on the law of gravity is formulated as given below.

$$F = G * \left(\frac{M1*M2}{R2}\right) \qquad (5)$$

From the equation (5), F denotes the gravitational force, G is gravitational constant, M_1 and M_2 are the mass of the first and second data items and R is the distance between two data items. In a Transactional database with N items, the position of the i^{th}item is defined as:

$$X_i = \left(X_i^1 \ldots.., X_i^d \ldots.., X_i^n\right) for\ i = 1,2, \ldots \ldots, n \qquad (6)$$

Where present the position of i^{th}item in the d^{th}dimension and N is dimension of the search space. At time t, a force acts on mass i from mass, j and this force F_{ij}^d'is defined as:

$$F_{ij}^d = G(t)\frac{m_{pi}(t) * m_{pj}(t)}{R_{ij}+\in}\left(X_j^d(t) - X_i^d(t)\right) \qquad (7)$$

Where is M_{pi} active gravitational mass of item i, M_{pj} is passive gravitational mass of item j, $G(t)$is gravitational constant at time t, $\in$ is a small constant and $R_{ij}(t)$ is Euclidian distance between two items i and j respectively. Then, the total force acting on mass, i in the dth dimension at time t is definedas:

$$F_i^d(t) = \sum_{j \in KBest, j \neq i}^{N} frand_j\, F_{ij}^d(t) \qquad (8)$$

Where $frand_j$ is a random number in the interval [0.0, 1.0], Kbest is the set of first K items withthe best fitness value. The acceleration related to mass i in time t in the d^{th}dimension is given as follows:

$$a_i^d = \frac{F_i^d(t)}{M_{ii}(t)} \qquad (9)$$

From the equation (9), M_{ii} is inertial mass of i^{th}item. The velocity of an item could be calculated as a fraction of its current velocity added to its acceleration. Position and velocity of agent is calculated as follows:

$$V_i^d(t+1) = frand_i V_i^d(t) + a_i^d(t) \qquad (10)$$

$$X_i^d(t+1) = X_i^d(t) + V_i^d(t+1) \qquad (11)$$

Where $frand$ is a uniform random variable in the interval [0.0, 1.0]. Let n number of rules be generated through conditional probability which is denoted as,

$$R_i = r_1, r_2, \ldots, r_n \qquad (12)$$

Based on the rules being generated, the gravitation search model is introduced to select the sensitive rules. In order to evaluate the sensitive rules, best fitness and worst fitness are evaluated using maximization problem which is mathematically formulated as,

$$Best\ \left(R_j\right) = \max fit(R_j) \qquad (13)$$

$$Worst\ (R_j) = \min fit(R_j) \qquad (14)$$

From the equation (13) and (14), $Best\ (R_j)$denotes the best fitness at iteration j and $Worst\ (R_j)$ denotes the worst fitness at iteration j respectively. Followed by this, the gravitational search is performed to identify the cohesive and non-cohesive items in each of the association rule generated for given transactional database.

The fitness function provided in this algorithm based on high support value of the rule serves as a basic criterion in extracting association rules. On the other hand, support value of the rule is considered to be non sensitive and not interesting for the user. The threshold T is set for the selecting the sensitive rules with the convergence of cohesive items and divergence of the non-cohesive items using gravitation constant. The gravitation constant is measured for each iteration j is formulated as,

$$GR(j) = Te^{\left(\frac{-j}{n}\right)} \qquad (15)$$

From the equation (15), the value of threshold T is initialized at the start of selection of sensitive rules with n representing the total number of iterations. Finally, using the gravitation constant, the rule comprising more cohesive items influences the privacy of an individual are selected as sensitive rule. It is formulated as given below.

$$CI = \left(\frac{fit(R_j) - Worst(R_j)}{Best(R_j) - Worst(R_j)}\right) \qquad (16)$$

$$NCI = \left(\frac{fit(R_j) - Best(R_j)}{Best(R_j) - Worst(R_j)}\right) \qquad (17)$$

From the equation (16) and (17), the cohesive and non-cohesive items are differentiated with the rule comprising more cohesive items are selected as the sensitive rule. With the separation of cohesive and non-cohesive items, execution time for selecting sensitive rule is minimized in a significant manner. The algorithmic process of Gravitation Search based Sensitive Rule Selection is shown in below.

// Gravitation Search based Sensitive Rule Selection Algorithm

Input: Frequent Item (FI), Set ($Item\ 1, Item\ 2, \ldots Item\ n$),Association Rules ($R_i$)

Output:Sensitive Rule (Best (R_j)), Non Sensitive Rule (Worst (R_j)), Occurrence Frequency (OF), Relative Frequency (RF), Threshold (TH), Cohesive Item (CI), Non Cohesive Item (NCI)

```
Step 1: Begin
Step 2:     Initialize $TH$ for $RF$ in the range from 0.0 to 1.0
Step 3:     For each $FI$
Step 4:         IF (($TH$ >=0.0) and ($TH$ <=1.0))
Step 5:             Select $FI$ as $CI$ for corresponding $R_i$ and Sort Descending $CI$ by Fitness
Step 6:             Select all $R_i$ with $CI$ as Best ($R_j$)
Step 7:         else
Step 8:             Select $FI$ as $NCI$ for corresponding $R_i$
Step 9:             Select all $R_i$with $NCI$ as Worst ($R_j$ )
Step 10:        End if
Step 11:        If $TH$ <=0.0 and no rule is sensitive
Step 12:            Re-initialize $TH$ with range value decremented to0.01
Step 13:            Process step 1 to step 7
Step 14:        Else
Step 15:            List all the Best ($R_j$ ) and Worst ($R_j$ )
Step 16:        End if
Step 17:        For ∀ FI ∈ Item do
Step 18:            Calculate Acceleration_OF(Item[t+1],GI)
Step 19:            Calculate Velocity_RF (Item[t+1],GI)
Step 20:            If Frand() < S, ($V_i^d(t+1)$) then
Step 21:                Exchange ($X_i^d[t+1]$), t++
Step 22:                Sensitive Rule -> Best ( $R_j$) DiscoveredBestRules
Step 23:            End if
Step 24:        End for
Step 25:    End for
Step 26: End
```

Algorithm 2 Gravitational Search Based Sensitive Rule Selection

As shown in algorithm 2, the threshold variance is initially set in the range of 0 to 1. The value of threshold variance changes during frequent item pruning, while evaluating the relative frequency to identify the sensitive and non-sensitive items. When the threshold value lies between 0 and 1, frequent items are selected as cohesive items to the corresponding association rules and it is also chosen as the sensitive rule. When the condition is not satisfied, frequent items are selected as non-cohesive items to the corresponding associative rules, referred to as the non-sensitive rules. When the threshold value is less than or equal to zero, there is no occurrence of sensitive rules. So, the threshold value gets reinitialized and the process gets repeated. Finally, association rules that are selected as sensitive rule is hidden in order to improve the privacy preservation of sensitive information's in transactional database.

4. EXPERIMENTAL SETUP

Motivated by the work in Group Incremental Approach using Rough Set (GIA-RS) [1] and Locality Sensitive Hiding for Privacy Preserving (LSH-PP) [2] that deals with dynamically increasing dataset and has undergone an anonymization transformation, a Gravitational Search Based Sensitive Rule Selection (GS-SRS) technique is introduced to improve the selection of sensitive rules to enhance the business transactions using MATLAB tool. The objective of the following experiments is to shown the effectiveness and efficiency of the proposed GS-SRS technique. Three datasets namely Taxi service trajectory, Tic-tac-toe and shuttle datasets are used in measuring the efficiency of GS-SRS technique. Their description is provided in the following sections.

4.1 Dataset Descriptions

4.1.1 Taxi service trajectory dataset

The performance evaluation of GS-SRS technique is performed using Taxi Service Trajectory dataset extracted from UCI repository. The dataset includes entire Taxi Service Trajectory evaluation dataset comprising of 9 attributes (Trip_ID, Call_Type, Origin_Call, Origin_Stand, Taxi_ID, Timestamp, Day_Type, Missing_Data and Polyline) with the aid of Matlab. This dataset has been chosen because it gives a clear picture that helps in evaluating the trajectories performed by all the 442 taxis running in the city of Porto, in Portugal from the view of the company, where sensitive rules are hidden and the attributes (i.e. characteristics) are displayed to the customer.

4.1.2 Tic-tac-toe dataset

The Tic-tac-toe database encodes the complete set of possible board configurations at the end of tic-tac-toe games, where "x" is assumed to have played first. The target concept is "winning for x" (i.e., true when "x" has one of 8 possible ways to create a "three in-a-row").

4.1.3. Shuttle dataset

The shuttle dataset comprises of 9 attributes. All the nine attributes are numerical. The examples in the original dataset were in time order, and this time order is more suitable for classification. However, this was not deemed relevant for Stat Log purposes, so the order of the examples in the original dataset was randomised, and a portion of the original dataset is removed for validation purposes. All the experiments have been carried out on a personal computer with Windows 7, Inter(R) Core (TM) i7-2600 CPU (2.66 GHz) and 4.00 GB memory. Experiment is conducted on the factors such as number of associative rules, selected sensitive rules, data utility rate, execution time for selecting sensitive rule, and size of the transaction database. The results of the metrics of GS-SRS technique is compared against with the existing methods such as Group Incremental Approach using Rough Set (GIA-RS) [1] and Locality Sensitive Hiding for Privacy Preserving (LSH-PP) [2].

4.1.4 Online Retail Data Set

The online retail data set is a transnational data set that includes the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail. This data set consists of 8 attributes namely InvoiceNo, StockCode, Description, Quantity, InvoiceDate, UnitPrice, CustomerID and Country.

No. of association rules	Accuracy Of Privacy Preservation (%)								
	Taxi service trajectory dataset			Shuttle dataset			Online Retail Data Set		
	GS-SRS	GIA-RS	LSH-PP	GS-SRS	GIA-RS	LSH-PP	GS-SRS	GIA-RS	LSH-PP
20	81.53	76.95	71.90	78.22	70.12	66.85	85.62	79.92	73.68
40	83.65	78.16	72.86	79.56	73.25	69.52	87.25	82.36	75.91
60	84.95	81.25	74.62	81.26	74.97	71.62	88.31	84.25	77.63
80	86.29	82.93	77.26	84.62	77.54	72.96	89.96	85.96	78.92
100	87.62	85.36	78.92	85.92	78.92	74.62	91.23	86.65	81.06
120	89.63	86.97	80.26	87.32	80.20	77.25	93.65	89.36	84.26
140	90.05	88.11	82.37	88.13	81.86	80.23	96.15	90.85	86.92

5. DISCUSSION

The Gravitational Search based Sensitive Rule Selection (GS-SRS) technique is compared against the existing Group Incremental Approach using Rough Set (GIA-RS) [1] and Locality Sensitive Hiding for Privacy Preserving (LSH-PP) [2] using MATLAB tool.

5.1 Case study

The results in Tables 1 and 2 shows that the impact of accuracy obtained and execution time for selecting sensitive rules have a profound influence on the final rule hiding. This part employs Frequency Item (FI) and Cohesive Items (CI) with Gravitational Search to conduct a brief case study on the data set taxi service trajectory, shuttle and Online Retail dataset respectively.

5.2 Case scenario 1: Impact of accuracy for privacy preservation

The accuracy of privacy preservation measures the ratio of number of sensitive rules that are correctly hidden and number of non sensitive rules that are correctly exposed to the total number of association rules generated. The accuracy of privacy preservation (A) measured in terms of percentages (%) and mathematically formulated as follows,

$$A = \left(\frac{\substack{number\ sensitive\ rules\ that\ are\\ correctly\ hidden+\\ number\ non\ sensitive\ rules\ that\ are\\ correctly\ exposed}}{number\ of\ assocation\ rules\ generated} \right) * 100 \quad (18)$$

From the equation (18), the privacy preservation accuracy is measured. While the accuracy of privacy preservation is higher, the method is said to be more efficient.

Table 1 Tabulation for accuracy of privacy preservation using Taxi service, Shuttle and Online Retail dataset

Table 1 show the accuracy of privacy preservation results is obtained with respect to different number of association rules in the range 20-140 using three datasets namely taxi service, shuttle and online retail data set. From the table value, it is illustrative that the accuracy of privacy preservation using proposed GS-SRS technique is higher as compared to other existing works using all three datasets.

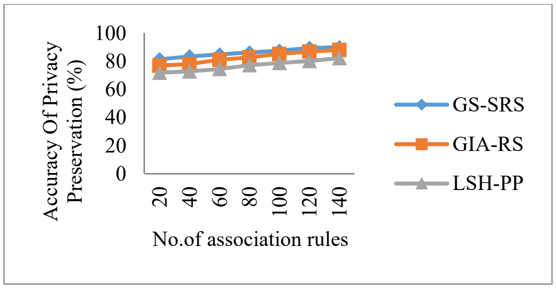

Figure 5 a) Measure of accuracy of privacy preservation using Taxi Service Trajectory dataset

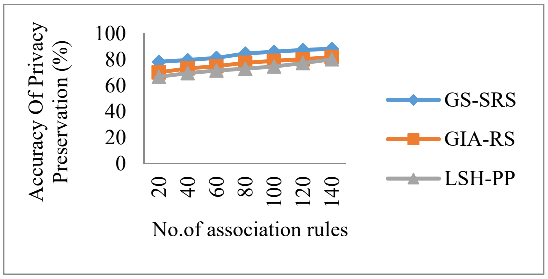

Figure 5 b) Measure of accuracy of privacy preservation using Shuttle dataset

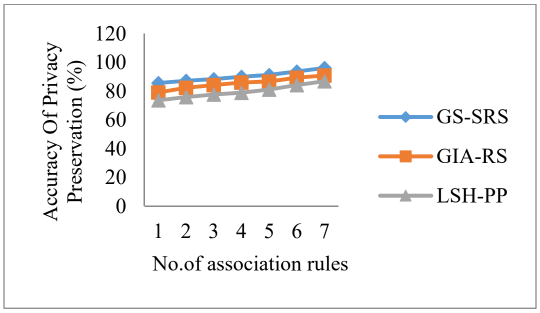

Figure 5 c) Measure of accuracy of privacy preservation using Online Retail dataset

Figure 5 shows the impact of privacy preservation accuracy results is obtained versus different number of association rules using three datasets namely taxi service, shuttle and online retail data set. As exposed in figure 5 a), b) c), proposed GS-SRS technique is provides better accuracy for privacy preservation while compared to existing GIA-RS [1] LSH-PP [2]. This is due to application of Gravitational Search Based Sensitive Rule Selection algorithm. By using this algorithmic process, proposed GS-SRS technique selects the sensitive rules to be hided and thereby efficiently hides the sensitive information about the transactional databases in order to preserve the privacy. This in turn helps for improving the accuracy of privacy preservation in an effective manner. Therefore, proposed GS-SRS technique increases the accuracy of privacy preservation by 4% as compared to GIA-RS [1] and 11% as compared to LSH-PP [2] when using Taxi Service Trajectory dataset. Further, proposed GS-SRS technique improves the accuracy of privacy preservation by 8% as compared to GIA-RS [1] and 12% as compared to LSH-PP [2] when using shuttle dataset. For applying the online retail dataset, proposed GS-SRS technique improves the accuracy of privacy preservation by 5% as compared to GIA-RS [1] and 12% as compared to LSH-PP [2] respectively.

5.3 Case scenario 2: Impact of execution time for selecting sensitive rule

The second case scenario considered for rule hiding is the execution time for selecting sensitive rule. To measure the execution time, two factors are considered, namely, number of associative rules and the time required to extract the cohesive item. To this, 20 to 140 association rules were considered at different intervals and seven iterations performed to measure the execution time. The comparison of execution time for selecting sensitive rule is presented in table 2 with respect to the number of associative rules generated for the given transaction database (Taxi Service Trajectory and Shuttle) at varied time intervals. As shown in the table, that it is feasible to perform rule hiding based on the experimental settings of associative rules selected to determine the execution time.

The execution time determines the time required for association rule generation and sensitive rules selection. The execution time is measured in terms of milliseconds (ms) and mathematically formulated as,

$$ET = n * time(association\ rule\ generation + sensitive\ rule\ selection) \qquad (19)$$

From the equation (19), the execution time ET for association rule generation and sensitive rules selection is obtained. To investigate upon the factors impacts the results of execution time, the technique apply the Gravitation Search algorithm to further check the results for the given transactional database. In table 2 it employ the Gravitation Search algorithm to arrive at the execution time. With increase in the number of associative rules, the execution time for selecting sensitive rule is also increased though not observed to be linear. This is because of the different types and nature of the transaction, collected at different time intervals based on the GPS data stream that invariably varies according to the timestamp.

Table 2 Tabulation for execution time using Taxi Service Trajectory and Shuttle and Online Retail dataset

No. of Association Rules		GS-SRS	GIA-RS	LSH-PP
20	Taxi Service Trajectory	12.6	15.1	18.8
40		17.1	18.5	23.6
60		20.5	24.9	27.2
80		25.9	30.7	33.8
100		29.3	33.2	39.6
120		32.4	35.8	42.7
140		36.7	40.4	48.3
20	Shuttle dataset	10.2	13.5	16.7
40		12.6	16.9	20.3
60		18.6	21.3	26.4
80		21.5	28.5	31.7
100		23.6	31.7	36.2
120		27.1	33.4	40.1
140		32.8	38.9	44.8
20	Online Retail	8.2	11.6	14.9
40		10.5	13.9	17.2
60		16.4	18.3	22.8
80		19.6	22.5	26.3
100		22.2	27.6	30.4
120		25.9	30.8	33.7
140		30.5	32.7	35.8

To ascertain the performance of the execution time for selecting sensitive rule, comparison is made between two other existing methods Group Incremental Approach using Rough Set (GIA-RS) [1] and Locality Sensitive Hiding for Privacy Preserving (LSH-PP) [2] applying the taxi service trajectory and shuttle and online retail dataset.

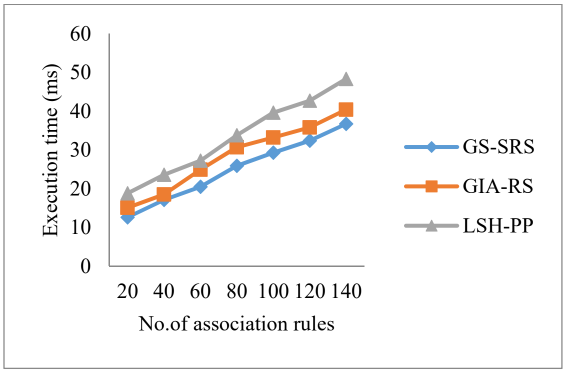

Figure 6 a) measure of execution time using taxi service trajectory dataset

No. of iterations	Memory Consumption (MB)								
	Taxi Service Trajectory dataset			Shuttle dataset dataset			Online Retail dataset		
	GS-SRS	GIA-RS	LSH-PP	GS-SRS	GIA-RS	LSH-PP	GS-SRS	GIA-RS	LSH-PP
2	3.4	4.8	5.5	2.5	3.6	4.2	1.9	2.4	3.5
4	4.1	5.5	6.2	3.1	4.2	4.8	2.2	3.2	4
6	4.8	5.8	6.5	3.3	4.4	5	2.7	3.9	4.9
8	4.2	5.6	6.3	3.6	4.7	5.3	3.1	4.3	5.2
10	5.3	6.7	7.4	3.9	5	5.6	3.9	4.7	5.4
12	5.5	6.9	7.6	4.3	5.4	6	4.3	5.1	5.7
14	5	6	6.7	4.5	5.6	6.2	4.5	5.3	5.9

Table 3 Tabulation for memory consumption using Taxi Service Trajectory and Tic-tac-toe dataset and Online Retail dataset

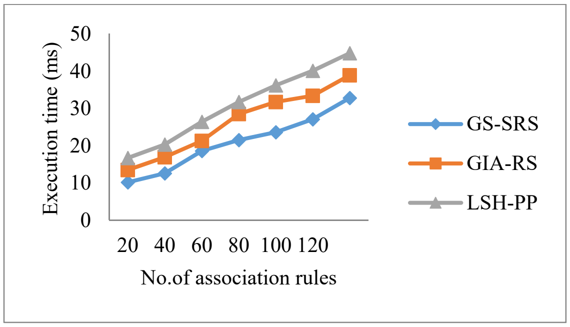

Figure 6 b) Measure of execution time using Shuttle dataset

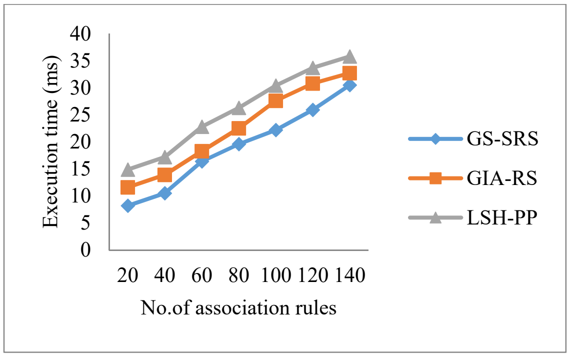

Figure 6 c) Measure of execution time using Online Retail dataset

Figure 6 a), b), c) shows a measure of execution time for sensitive rule selection using Taxi service trajectory, shuttle dataset, and online retail dataset. From figure 6 a), b), c) , it is found that the execution time for selecting sensitive rule is less using the proposed GS-SRS technique when compared to the two other existing methods. Also by applying shuttle dataset, it showed better performance when compared to the taxi service trajectory. With the application of Gravitation Search algorithm, the cohesive and non-cohesive data items are measured for extracting the sensitive rules. This separation of cohesive and non-cohesive items minimizes the execution time while selecting sensitive rules using GS-SRS by 15% as compared to GIA-RS [1] and 36% as compared to LSH-PP [2] while using Taxi Service Trajectory dataset. Besides, by applying the gravitational constant, best fitness and worst fitness are evaluated. Followed by this, the gravitational search performed to identify the cohesive and non-cohesive items for the given transactional database, which further reduces the execution time for selecting sensitive rule using GS-SRS technique by 27% as compared to GIA-RS [1] and 50% as compared to LSH-PP [2] while Shuttle dataset. Besides for applying the online retail dataset, proposed GS-SRS technique reduces the execution time by 22% as compared to GIA-RS [1] and 43% as compared to LSH-PP [2] respectively.

5.4 Case scenario 3: Impact of memory consumption

Finally, the third case considered is the memory consumption. To measure the memory consumption, three variables namely, the memory consumption for deriving sensitive rules, the memory for cohesive and memory for non-cohesive are used. Convergence characteristics for the measure of memory consumption for 14 iterations are taken into consideration and compared with two other methods.

With two iterations considered, the memory consumption for cohesive items was found to be 1MB (2MB using GIA-RS and 2MB using LSH-PP) and with that for obtaining non-cohesive items to be 1MB (2MB using GIA-RS and 2MB using LSH-PP), whereas the memory required for obtaining sensitive rules is found to be 1MB using GS-SRS technique, GIA-RS and LSH-PP respectively. Memory consumption for cohesive and non-cohesive items is mathematically formulated as given below.

$$M = M(SR) - [M(CI) + M(NCI)] \qquad (20)$$

The memory consumption is measured in terms of megabytes (MB). While the memory consumption for achieving privacy preservation is lower, the method is said to be more efficient.

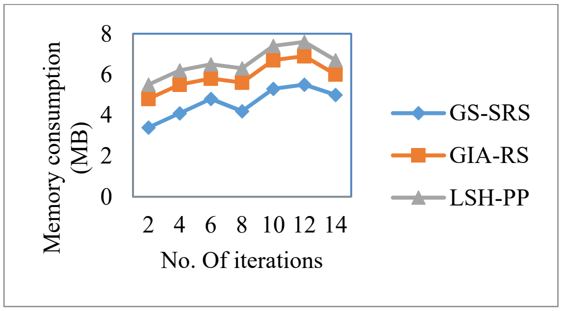

Figure 7 a) Measure of memory consumption using Taxi Service Trajectory dataset

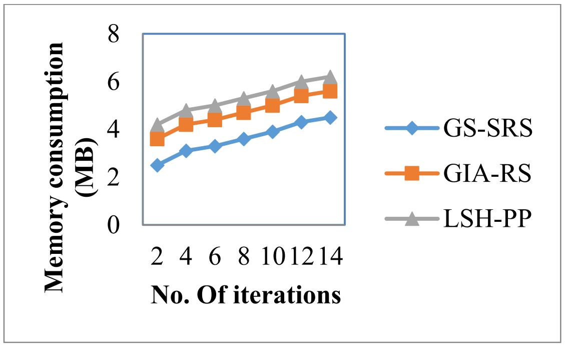

Figure 7 b) Measure of memory consumption using Shuttle dataset dataset

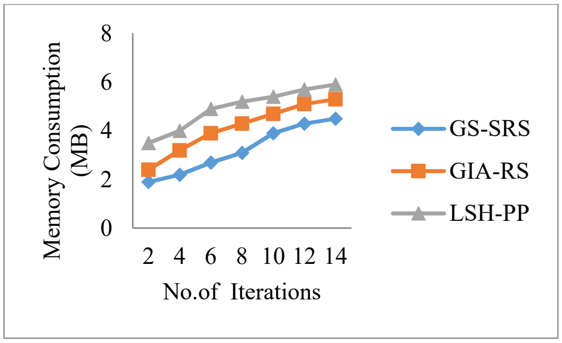

Figure 7 c) Measure of memory consumption using Online Retail dataset

The targeting results of memory consumption for cohesive and non-cohesive item generation using GS-SRS technique is compared with two state-of-the-art methods [1], [2] in Figure 7 a), b), c) is presented for visual comparison based on the number of iterations. This technique differs from the GIA-RS [1] and LSH-PP [2] in that, the technique have incorporated gravitational search in that cohesive and non-cohesive items are listed for the given transaction database in a much shorter period of time. The threshold is set for the selecting the sensitive rules with the convergence of cohesive items and divergence of the non-cohesive items using gravitation constant. As a result, the memory consumption for cohesive and non-cohesive items generations using GS-SRS technique is decreased by 28% as compared to GIA-RS [1] and 44% as compared to LSH-PP [2] while using Taxi Service Trajectory dataset. Though nine attributes were used in Tic-tac-toe and Taxi Service Trajectory dataset, with samples greater than Taxi Service, the memory consumed is lesser because of a single factor considered (win for x). Therefore, the memory consumed using GS-SRS when applied with Tic-tac-toe dataset is 33% reduced when compared to GIA-RS and 51% reduced when compared to LSH-PP respectively. Besides for applying the online retail dataset, proposed GS-SRS technique reduces the memory consumption by 30% as compared to GIA-RS [1] and 60% as compared to LSH-PP [2] respectively.

6 CONCLUSIONS

The Gravitational Search Based Sensitive Rule Selection Algorithm (GS-SRS) presented in this work is concluded as an effective sensitive rule selection technique for preserving the privacy of transactional database. The GS-SRS technique improves the selected sensitive rules that enhance the performance of decision making in business and/or association rule hiding technique for privacy preservation. The goal of Gravitational Search Based Sensitive Rule Selection is to select the sensitive association rules generated from frequent item sets, extracted from the transactional database to enhance the privacy of business transactions with support count and confidence value. This GS-SRS technique first designed a conditional probability-based association rule selection that measures the support and confidence value based on the conditional probability which resulting in generation of association rules. In addition with the derived associative rule, Gravitational Search Based Sensitive Rule Selection Algorithm (GS-SRS) is designed to list the cohesive and non-cohesive items in each of the association rule generated and therefore reduces the execution time for selecting sensitive rules. Finally, the Gravitation Search model efficiently selects the sensitive rules for hiding and thereby preserving the privacy of transactional data base which resulting in improved accuracy of privacy preservation. The efficiency of GS-SRS technique is test with the metrics such as accuracy of privacy preservation, execution time and memory consumption using three datasets namely taxi trajectory, Tic-tac-toe dataset, shuttle and online retail dataset . With the experiments conducted, it is observed that the GS-SRS technique provided more accurate results as compared to state-of-the-art works. The experimental results demonstrate that GS-SRS technique is provides better performance with an improvement of accuracy of privacy preservation and reduction of execution time when compared to the state-of-the-art works.

REFERENCES:

[1] Jiye Liang, Feng Wang, Chuangyin Dang, and YuhuaQian, “A Group Incremental Approach to Feature Selection Applying Rough Set Technique”, IEEE Transactions on Knowledge and Data Engineering, Volume 26, Issue 2, February 2014, Pages 294-308.

[2] DimitriosKarapiperis and Vassilios S. Verykios, “An LSH-Based Blocking Approach with a Homomorphic Matching Technique for Privacy- Preserving Record Linkage”, IEEE Transactions on Knowledge and Data Engineering, Volume 27, Issue 4, April 2015, Pages 909-921.

[3] Gabriel Ghinita, PanosKalnis, and YufeiTao,“Anonymous Publication of Sensitive Transactional Data”, IEEE Transactions on Knowledge and Data Engineering, Volume 23, Issue 2, February 2011, Pages 161-174.

[4] Ahmed Abbasi, Stephen France, Zhu Zhang, and Hsinchun Chen, “Selecting Attributes for Sentiment Classification Using Feature Relation Networks”, IEEE Transactions on Knowledge and Data Engineering, Volume 23, Issue 3, March 2011, Pages 447-462.

[5] PradiptaMaji and Sankar K. Pal, “Feature Selection Using f-Information Measures in Fuzzy Approximation Spaces”, IEEE Transactions on Knowledge and Data Engineering, Volume 22, Issue 6, June 2010, Pages 854-867.

[6] ZahidPervaiz, Walid G. Aref , ArifGhafoor , NagabhushanaPrabhu, “Accuracy-constrained Privacy preserving Access Control Mechanism for Relational Data”, IEEE Transactions on Knowledge and Data Engineering, Volume 26, Issue 4, April 2014, Pages 795-807.

[7] Edoardo Serra and Francesca Spezzano, “An Effective GPU-Based Approach to Probabilistic Query Confidence Computation”, IEEE Transactions on Knowledge and Data Engineering, Volume 27, Issue 1, January 2015, Pages 17-31.

[8] Suyun Zhao, Eric C.C. Tsang, Degang Chen, and XiZhao Wang, “Building a Rule-Based Classifier—A Fuzzy-Rough Set Approach”, IEEE Transactions on Knowledge and Data Engineering, Volume 22, Issue 5, May 2010, Pages 624-638.

[9] Ke Zhou, Gui-RongXue, Qiang Yang, and Yong Yu, “Learning with Positive and Unlabeled Examples Using Topic-Sensitive PLSA”, IEEE Transactions on Knowledge and Data Engineering, Volume 22, Issue 1, January 2010, Pages 46-58.

[10] Chun-Wei Lin, Tzung-Pei Hong, and Hung-Chuan Hsu, “Reducing Side Effects of Hiding Sensitive Itemsets in Privacy Preserving Data Mining”, Hindawi Publishing Corporation, The Scientific World Journal, Volume 2014, April 2014, Pages 1-13.

[11] Chongjing Sun, Yan Fu, Junlin Zhou, and HuiGao, “Personalized Privacy-Preserving Frequent Itemset Mining Using Randomized Response”, Hindawi Publishing Corporation, The Scientific World Journal, Volume 2014, March 2014, Pages 1-11.

[12] Chun-Wei Lin, Tzung-Pei Hong, Jia-WeiWong, Guo-Cheng Lan, andWen-Yang Lin, “A GA-Based Approach to Hide Sensitive High Utility Itemsets”, Hindawi Publishing Corporation, The Scientific World Journal, Volume 2014, March 2014, Pages 1-13.

[13] Chun-Wei Lin, Binbin Zhang, Kuo-Tung Yang, and Tzung-Pei Hong, “Efficiently Hiding Sensitive Itemsets with Transaction Deletion Based on Genetic Algorithms”, Hindawi Publishing Corporation, The Scientific World Journal, Volume 2014, September2014, Pages 1-14.

[14] Yves-Alexandre de Montjoye, ErezShmueli, Samuel S. Wang, Alex Sandy Pentland, “openPDS: Protecting the Privacy of Metadata through SafeAnswers”, PLoS ONE, Volume 9, Issue 7, July 2014, Pages 1-9.

[15] Peng Cheng, Chun-Wei Lin Jeng-Shyang Pan, “Use HypE to Hide Association Rules by Adding Items”, PLoS ONE, Volume 10, Issue 6, April 2015, Pages 1-19.

[16] Tong Yi and Minyong Shi, “Privacy Protection Method for Multiple Sensitive Attributes Based on Strong Rule”, Hindawi Publishing Corporation, Mathematical Problems in Engineering, July 2015, Pages 1-15.

[17] Mohamed Ouda, Sameh Salem, Ihab Ali, and El-SayedSaad, “Privacy-Preserving Data Mining in Homogeneous Collaborative Clustering”, The International Arab Journal of Information Technology, Volume 12, Issue 6, November 2015, Pages 604-612.

[18] Yousra Abdul Alsahib S. Aldeen, Mazleena Salleh1 and Mohammad AbdurRazzaque, “A comprehensive review on privacy preserving data mining”, Aldeen et al. SpringerPlus , Volume 4, November 2015, Pages 1-36.

[19] Rahat Ali SHAH1, Sohail ASGHAR, “Privacy preserving in association rules using a genetic algorithm”, Turkish Journal of Electrical Engineering & Computer Sciences, Volume 22, Pages 1-17.

[20] Vassilios S. Verykios, Ahmed K. Elmagarmid, Elisa Bertino, YucelSaygin, Elena Dasseni, “Association Rule Hiding”, IEEE Transactions On Knowledge And Data Engineering, Vol. 16, No. 4, April 2004.

[21] E. Rashedi, H. Nezamabadi-pour, and S. Saryazdi, “GSA: A Gravitational Search Algorithm,” Information Sciences, vol. 179, pp. 2232-2248, 2009.

[22] Gkoulalas-Divanis, Aris, Verykios, Vassilios S. “Association Rule Hiding for Data Mining”, Springer Series: Advances in Database Systems, Vol. 41, 1stEdition., 2010, Pages 13.

[23] Fosca Giannotti, Laks V. S. Lakshmanan, Anna Monreale, Dino Pedreschi, and Hui (Wendy) Wang, “Privacy-Preserving Mining of Association Rules From Outsourced Transaction Databases”, IEEE Systems Journal, Volume 7, Issues 3, Pages 385-395, September 2013

[24] ianming Zhu, Zhanyu Li, “Privacy Preserving Association Rule Mining Algorithm Based on Hybrid Partial Hiding Strategy”, LISS, Springer, Pages 1065-1070, 2013

[25] Bettahally N. KeshavamurthyEmail authorAsad M. KhanDurga Toshniwal, “Privacy preserving association rule mining over distributed databases using genetic algorithm”, Neural Computing and Applications, Springer, Volume 22, Pages 351–364, May 2013

[26] Tamir Tassa, “Secure Mining of Association Rules in Horizontally Distributed Databases”, IEEE Transactions on Knowledge and Data Engineering, Volume 26, Issues 4, Pages 970-983, April 2014

[27] Zahid Pervaiz, Walid G. Aref Arif, Ghafoor, Nagabhushana Prabhu, “Accuracy-constrained Privacy-preserving Access Control Mechanism for Relational Data”, IEEE Transactions On Knowledge And Data Engineering, Volume 26, Issue 4, Pages 795 – 807, April 2014

[28] Pravin R. Ponde and S. M. Jagade, “Privacy Preserving by Hiding Association Rule Mining from Transaction Database”, IOSR Journal of Computer Engineering, Volume 16, Issue 5, Pages 25-31, 2014