A Hybrid Approach to Association Rule Hiding

S. Sangeetha,
Lecturer,
Department of information Technology,
PSG Polytechnic College, Coimbatore,
Tamil Nadu – 641004, India

R. Kiruba,
Student,
Department of Information Technology,
PSG Polytechnic College, Coimbatore,
Tamil Nadu – 641004, India

ABSTRACT

Data mining is a technique for summarizing and identifying similar patterns in data. Data mining can take different approaches and build different models depending upon the type of data involved and the objectives. In this Paper we follow the association rules approach for finding the correlation relationships among large set of data items. The rules are generated in order to hide the sensitive rules which are highly confidential by using DSR (Decrease support value of Right Hand Side) approach and PSO (Particle Swarm Optimization) approach. In this paper we propose a new algorithm called HYBRID algorithm. The objective of this paper is to reduce the side effects such as ghost rule and lost rule and number of modification and to increase the hiding ratio by hybrid approach which is achieved by combination of DSR & PSO. Experimental results of the proposed approach demonstrate the efficient information hiding with fewer side effects and modifications.

Keywords

Item sets, data mining, Association rules, privacy preservation, DSR approach, PSO approach, Hybrid.

1. INTRODUCTION

Data mining involves many different algorithms to accomplish different tasks. The purpose of these algorithms is to fit a model to the data. A data mining model can be either predictive or descriptive in nature. A predictive model makes prediction about values of data using known results found from different data and other historical data. The predictive models include classification, regression, time series analysis and prediction. A descriptive model identifies patterns or relationships in data. It explores the properties of the data being examined. It does not predict new values of the properties like predictive models. The descriptive models include clustering, summarization, association rules and sequence discovery[1].

Classification: Classification involves the predictive learning that classifies a data item into one of several predefined classes. It involves examining the features of an item and assigning to it a predefined class. Classification is a two-step process. First a model is built describing a predefined set of data classes and secondly, the model is used for classification.

Summarization: It is the abstraction or generalization of data. A set of relevant data is summarized and abstracted, resulting in smaller set which gives a general overview of the data and usually with aggregation information.

Association Rules: Association rules are if/then statements that help uncover relationships between seemingly unrelated data in a relational database or other information repository. An association rule has two parts, an antecedent (if) and a consequent (then). An antecedent is an item found in the data. A consequent is an item that is found in combination with the antecedent.

1.1 Association Rules

Association mining is about discovering a set of rules that is shared among a large percentage of the data. Association rules mining tend to produce a large number of rules. The goal is to find the rules that are useful to users. There are two criteria of measuring usefulness viz. support and confidence. Support is an indication of how frequently the items appear in the database. Confidence indicates the number of times the if/then statements have been found to be true.

The association rules problem is as follows: Let I= {i1, i2… in} be a set of literals call items. Let D be a set of all transactions where each transaction T is a set of items such that $T \subseteq I$. Let X, Y be a set of items such that $X, Y \subset I$. An association rule is an implication in the form $X \Rightarrow Y$, where $X \subset I$, $Y \subset I$, $X \cap Y=\emptyset$. Support: The rule $X \Rightarrow Y$ holds with support s if s% of transactions in D contains $X \cup Y$. Rules that have a s greater than a user-specified support are said to have minimum support. Confidence: The rule $X \Rightarrow Y$ holds with confidence c if c% of the transactions in D that contain X also contain Y. Rules that have a c greater than a user-specified confidence are said to have minimum confidence. The association between various data items can be found out by mining multilevel association rules, multidimensional association rules and/or quantitative association rules.

The left hand side of an association rule is called the antecedent, and the right hand side is the consequent. In the Cheese → Beer example Cheese is the antecedent and Beer is the consequent. The classic application of association rule mining is the market basket data analysis, which aims to discover how items purchased by customers in a supermarket (or a store) are associated. An example association rule is Cheese → Beer [support = 10%, confidence = 80%] The rule says that 10% customers buy Cheese and Beer together, and those who buy Cheese also buy Beer 80% of the time. The problem of mining association rules can be stated as follows: Let I = {i1, i2,……., im} be a set of items. Let T = (t1, t2, …, tn) be a set of transactions (the database), where each transaction ti is a set of items such that ti subset of I. An association rule is an implication of the form, X → Y, where and X (or Y) is a set of items, called an item set. Support: The support is the ratio (or percentage) of the number of item sets satisfying both antecedent and consequent to the total number of transaction [9].

The support of a rule, X → Y, is the percentage of transactions in T that contains, and can be seen as an estimate of the probability; the rule support thus determines how frequent the rule is applicable in the transaction set T. Let n be the number of transactions in T. The support of the rule X →Y is computed as follows:

$$support = \frac{|X \cup Y| * count}{N} \qquad \text{---------} \ (1)$$

Since N is constant (as it is the number of transactions in the given database). Support is a useful measure because if it is too low, the rule may just occur due to chance. Furthermore, in a business environment, a rule covering too few cases (or transactions) may not be useful because it does not make business sense to act on such a rule (not profitable).

Confidence: Confidence (strength or evidence) derives from a subset of the transaction in which two entities (or activities) are related. The confidence of a rule, X → Y, is the percentage of transactions in T that contain X also contain Y. It can be seen as an estimate of the conditional probability, Pr(Y | X). It is computed as follows:

$$\text{confidence} = \frac{|X \cup Y| * \text{count}}{|X|} \qquad \text{----------} \ (2)$$

Confidence thus determines the predictability of the rule. If the confidence of a rule is too low, one cannot reliably infer or predict Y from X. A rule with low predictability is of limited use.

Yi Hung wu [2] proposes an algorithm to hide rules with limited side effects but efficient approach to speed up the rule hiding process is not addressed. Ila Chandrakar [3] proposes a hybrid approach for rule hiding which reduces the scanning of the database but they haven’t addressed the side effects created by the proposed algorithm.

Dr .K. Duraisamy [5] proposes a new algorithm to sensitive rule hiding. The sensitive rule hiding algorithm clusters the sensitive rules and modifies the database to hide the rules. Further the clusters are converted into modifies database; however this approach has higher side effect of producing new rules called ghost rules. The HYBRID algorithm proposed in this paper addresses the problem and it does not create any ghost rule.

2. PROBLEM DESCRIPTION

The existing algorithms are being utilized for the purpose of sensitive item set hiding for a long time and across all the domains. Majority of the algorithms hides the sensitive information but has some implications on the data set like introduction of new rules, lost association rule and hiding failures. Algorithms are mainly focused on hiding the sensitive association rule without looking at the fact that how many databases they have to make while they compare the rules before applying the sensitive item set hiding. So, it is clear from the above discussion that there is scope that there should be some strategy which implements the association rule hiding while making the fewer changes to the database.

2.1. Proposed Approach

The progress of the proposed approach is represented in fig 1.

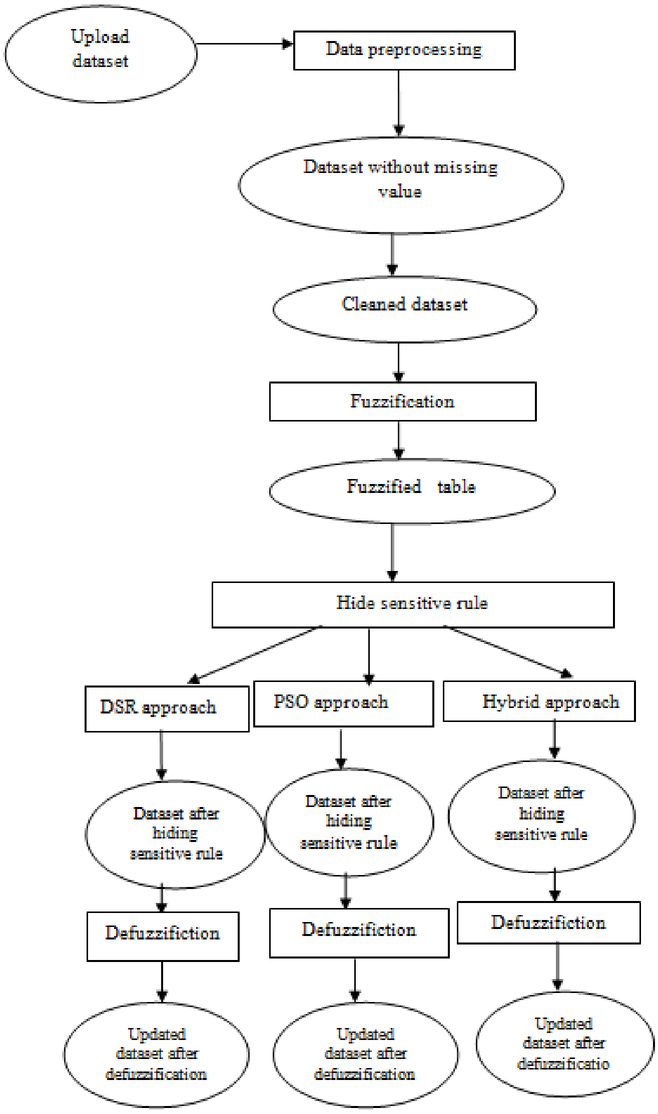
Upload dataset
Data preprocessing
Dataset without missing value
Cleaned dataset
Fuzzification
Fuzzified table
Hide sensitive rule
DSR approach
PSO approach
Hybrid approach
Dataset after hiding sensitive rule
Dataset after hiding sensitive rule
Dataset after hiding sensitive rule
Defuzzifiction
Defuzzifiction
Defuzzifiction
Updated dataset after defuzzification
Updated dataset after defuzzification
Updated dataset after defuzzificatio

Fig 1: Progress of the proposed approach

2.1.1 Data Collection

The dataset is taken from UCI Machine Learning repository [4].

2.1.2 Data Preprocessing

The database as in Table 1 is cleaned by substituting the unknown values by zero, and eliminating the redundant records.

Table 1. Sample data with 5 attributes

	A	**B**	**C**	**D**	**E**
T1	3	?	?	2	1
T2	14	5	10	4	2
T3	12	9	8	5	3
T4	10	8	10	6	4
T5	13	4	11	8	9

Table 2. Cleaned data

	A	B	C	D	E
T1	3	0	0	2	1
T2	14	5	10	4	2
T3	12	9	8	5	3
T4	10	8	10	6	4
T5	13	4	11	8	9

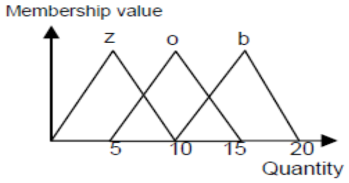

Fig 2: Triangular Member Function

Table 3. Fuzzified Transactional data

Transaction	**A**			**B**			**C**			**D**			**E**		
n	Az	Ao	Ab	Bz	Bo	Bb	Cz	Co	Cb	Dz	Do	Db	Ez	Eo	Eb
T1	0.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.0	0.0	0.2	0.0	0.0
T2	0.0	0.2	0.8	1.0	0.0	0.0	0.0	1.0	0.0	0.8	0.0	0.0	0.4	0.0	0.0
T3	0.0	0.6	0.4	0.2	0.8	0.0	0.4	0.6	0.0	1.0	0.0	0.0	0.6	0.0	0.0
T4	0.0	1.0	0.0	0.4	0.6	0.0	0.0	1.0	0.0	0.8	0.2	0.0	0.8	0.0	0.0
T5	0.0	0.4	0.6	0.8	0.0	0.0	0.0	0.8	0.2	0.4	0.6	0.0	0.2	0.8	0.0
Count	0.6	2.2	1.8	2.4	1.4	0.0	0.4	3.4	0.2	3.4	0.8	0.0	2.2	0.8	0.0

2.1.3 Fuzzification

The database after preprocessing is shown in table 2 which is fuzzified using triangular membership function given in equation (3). It is separated into 3 regions Z, O, B as shown in figure 3. The fuzzified data is shown in table 3.

$$\mu = Max\left(Min\left(\frac{x-a}{b-a},\frac{c-x}{c-b}\right),0\right) \qquad \text{-------} \ (3)$$

Where 'a' is the left end of the triangle, 'b' is the peak of the triangle and 'c' is the right end of the triangle (values are the corresponding x axis values).

2.1.4 Calculating support and confidence

Calculate the support count of each attribute region, R on the transactions data by summing up the fuzzy values of all the transactions in the fuzzified transaction data as in table 3.

Check whether count of each attribute is greater than or equal to the predefined minimum support value. If an attribute satisfies the above condition, put it in the set of large-2 itemsets (L2). Consider the minimum support is set to 2.3 and minimum confidence to 70%. The regions Bz, Co and Dz have their support value greater than minimum support, so they are considered in forming the rules and finding the corresponding confidence value. The rules can be Bz →Co, Co →Dz, Bz →Dz, Co →Bz, Dz→ Co, Dz→ Bz. Consider Bz →Co is a critical rule to be hidden and the support of the rule is calculated as Support(Bz →Co) = min(Bz,Co) as shown in table 4.

Table 4. Fuzzy values of Bz and Co

	Bz	**Co**	**Support**
T1	0.0	0.0	0.0
T2	1.0	1.0	1.0
T3	0.2	0.6	0.2
T4	0.4	1.0	0.4
T5	0.8	0.8	0.8
Count	2.4		2.4

For each 2 large item sets, based on user specified minimum confidence value, rules are extracted. Confidence value of A→B rule is computed as follows:

$$confidence(A \to B) = \frac{Support(AB)}{Support(A)}$$

The confidence value calculated for the rule Bz →Co

$$confidence(\mathrm{Bz} \ \to \mathrm{Co}) = \frac{2.4}{2.4} = 100\%$$

To hide a critical rule, its confidence value is decreased by decreasing support (AB). In order to hide the rule Bz →Co, the support (BzCo) is reduced by subtracting the transaction value of Co from 1 when the value of Co is greater than 0.5 and corresponding Bz's value. Using this procedure the support values of transaction T3 and T4 are reduced as shown in table 5.

Table 5. Modified T3 and T4

	Bz	**Co**	**Support**
T1	0.0	0.0	0.0
T2	1.0	1.0	1.0
T3	0.2	0.4	0.2
T4	0.4	0.0	0.4
T5	0.8	0.8	0.8
Count	2.4		2.0

Now the confidence is

$$confidence(\mathrm{Bz} \rightarrow \mathrm{Co}) = \frac{2.0}{2.4} = 92\%$$

As the confidence is still greater than minimum confidence, in those transactions that have Bz and Co value as 1, Co is replaced with 0 as shown in T2 of table 6.

Table 6. modified T2

	Bz	**Co**	**Support**
T1	0.0	0.0	0.0
T2	1.0	0.0	0.0
T3	0.2	0.4	0.2
T4	0.4	0.0	0.4
T5	0.8	0.8	0.8
Count	2.4		1.0

The confidence value after the modification is calculated as

$$confidence(\mathrm{Bz} \rightarrow \mathrm{Co}) = \frac{1}{2.4} = 42\%$$

As the confidence value is less than the predefined confidence value the rule Bz →Co is hided. The modified values replace the original fuzzified values in the fuzzification table as shown in table 7.

Table 7. Modified fuzzy table

Transaction	**A**			**B**			**C**			**D**			**E**		
n	Az	Ao	Ab	Bz	Bo	Bb	Cz	Co	Cb	Dz	Do	Db	Ez	Eo	Eb
T1	0.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.0	0.0	0.2	0.0	0.0
T2	0.0	0.2	0.8	1.0	0.0	0.0	0.0	0.0	0.0	0.8	0.0	0.0	0.4	0.0	0.0
T3	0.0	0.6	0.4	0.2	0.8	0.0	0.4	0.4	0.0	1.0	0.0	0.0	0.6	0.0	0.0
T4	0.0	1.0	0.0	0.4	0.6	0.0	0.0	0.0	0.0	0.8	0.2	0.0	0.8	0.0	0.0
T5	0.0	0.4	0.6	0.8	0.0	0.0	0.0	0.8	0.2	0.4	0.6	0.0	0.2	0.8	0.0
Count	0.6	2.2	1.8	2.4	1.4	0.0	0.4	1.2	0.2	3.4	0.8	0.0	2.2	0.8	0.0

2.1.5 DSR Approach

In DSR approach, confidence of a rule is decreased by decreasing the support value of right hand side (R.H.S.) of a rule.

In order to hide an association rule, X → Y, either decreases its support or its confidence to be smaller than user-specified MST and MCT. To decrease the confidence of a rule, either increase the support of X, the LHS of the rule, but not support of X U Y, or decrease the support of the itemset X U Y. For the second case, decrease the support of Y, the right hand side of the rule, it would reduce the confidence faster than simply reducing the support of XU Y.

To decrease support of an item, the system will modify one item at a time by changing from 1 to 0 or from 0 to 1 in a selected transaction. The DSR Approach is implemented by applying the DSR Algorithm.

2.1.6 Defuzzfication

Defuzzification using centroid method is done on the modified values to get back quantitative values using the equation (4). The defuzzified values are shown in table 8.

$$X = \frac{\sum_{i=1}^{n} Xi.\mu(Xi)}{\sum_{i=1}^{n} \mu(Xi)} \qquad \text{-------}\ (4)$$

X is the quantitative value

n is the number of regions

xi is the center point of that triangle

μ(xi),corresponding membership value in that triangle

Table 8. Deffuzified table

	A	B	C	D	E
T1	3	0	0	2	1
T2	14	5	0	4	2
T3	12	9	8	5	3
T4	10	8	0	6	4
T5	13	4	11	8	9

The deffuzification method is used after each and every approach.

2.1.7 PSO Approach

PSO is based on the sociological behaviour associated with bird flocking. The algorithm maintains a population of particles where each particle represents a potential solution to an optimization problem. Let s be the size of the swarm. Each particle 'i' can be represented as an object with several characteristics. The algorithm is stated below.

2.2 Hybrid Approach

The objective of the hybrid algorithm for privacy preserving data mining is to hide certain sensitive information so that they cannot be discovered through association rule mining techniques. The hybrid approach is achieved through combination of DSR & PSO. The datasets is taken from the UCI repository. The datasets is given to data preprocessing method then the cleaned data is fuzzified and the sensitive rules are hidden using the DSR algorithm before modifying or re-calculating the confidence the hidden rules is given as an input to the PSO algorithm and updates the database.

3. ALGORITHM

Algorithm for hiding fuzzy association rule using DSR

In a quantitative database, if a critical rule X→Y needs to be hidden; its confidence value is decreased to a value smaller than the minimum confidence value. One way of decreasing confidence value is decreasing the support value of an item Y at RHS, and the other way is increasing the support value of item X at LHS. Our approach decreases confidence value of a rule, by decreasing the support value of RHS item. If the value of item in RHS is greater than 0.5 and value of item in LHS then its value is subtracted from 1.

Abbreviations used in the proposed algorithm are given as follows:

C: Dataset with 'n' transactions

F: Fuzzified database

X: A set of predicting items

TL: Transactions belong to a LHS item

TR: Ttransactions belong to a RHS item

U: Rule

Rh: Sensitive rule

Input

(1) Cleaned dataset

(2) Minimum support value (min_support),

(3) Minimum confidence value (min_confidence).

Output

The values after processing is transformed to database D' so that useful fuzzy association rules cannot be mined.

Algorithm DSR

```
1. Dataset
2. Fuzzification of the cleaned database, C → F;
3. In fuzzified database F, calculate every item's support value where f→F;
4. IF all f (support) < min_support THEN EXIT; // there isn't any rule
6. Find large 2-itemsets from F;
7. FOR EACH X's large 2-itemset //find all rules
        Find R = {Rules from item set X};
        //for X= {i1, i2}, rules are i1 → i2, i2 → i1.
        Compute confidence of the rule U;
        IF confidence (U) > min_confidence and sensitive
      THEN
        Add the rule U to Rh;
        end//if
 end//end of FOR EACH
  //Hides all rules in Rh
8. FOR EACH U in Rh {//until no more rule can be hidden
        FOR EACH TR of rule{
                if TR >0.5 and TR > TL
            TR = 1 - TR
          end // if
        end // FOR EACH.
Re-calculate confidence value of rule U
        if rule U(confidence) > min_confidence
        FOR EACH TR of the rule
          if TR = 1.0
          TR = 0.0
          end// if
        end // FOR EACH
        else go to step 9
        end //if
9. Transform the updated database F to D' and output
updated D'
10. end.
```

Algorithm for hiding fuzzy association rule using PSO

X_i : The current position of the particle

V_i : The current velocity of the particle

Y_i : The personal best position of the particle

```
1.   Create and initialize an n-dimensional PSO: S
          Repeat:
2.   for each particle i [1,.....S] :
                    If f(S.Xi) < f(S.Yi)
Then S.Yi = S.Xi
                    If f(S.Yi) < f(S.Ŷ)
Then S.Ŷ = S.Yi
                    End for
3.   Perform PSO updates on S using equations 3 and 4
Until stopping condition is true
```

Algorithm for hiding fuzzy association rule using Hybrid approach

Abbreviations used in the proposed algorithm are given as follows:

C: Dataset with ‘n’ transactions

F: Fuzzified database

X: A set of predicting items

TL: Transactions belong to a LHS item

TR: Ttransactions belong to a RHS item

U: Rule

Rh: Sensitive rule

X_i: The current position of the particle

V_i: The current velocity of the particle

Y_i : The personal best position of the particle

Hybrid algorithm

```
1. Dataset
2. Fuzzification of the cleaned database, C → F;
3. In fuzzified database F, calculate every item’s support
value where f→F;
4. IF all f (support) < min_support THEN EXIT; // there isn’t
any rule
6. Find large 2-itemsets from F;
7. FOR EACH X’s large 2-itemset //find all rules
          Find R = {Rules from item set X};
          //for X= {i1, i2}, rules are i1 → i2, i2 → i1.
          Compute confidence of the rule U;
          IF confidence (U) > min_confidence and sensitive
       THEN
          Add the rule U to Rh;
          end//if
 end//end of FOR EACH
  //Hides all rules in Rh
8. Create and initialize an n-dimensional PSO: S
          Repeat:
9.for each particle i [1,.....S] :
                    If f(S.Xi) < f(S.Yi)
Then S.Yi = S.Xi
                    If f(S.Yi) < f(S.Ŷ)
Then S.Ŷ = S.Yi
                    End for
10.Perform PSO updates on S using equations 3 and 4
Until stopping condition is true
```

4. EXPERIMENTAL RESULTS

We conducted experiments based on the breast cancer dataset from UCI repository and the results are analyzed. The original dataset with missing values are cleaned using data preprocessing method. After cleaning and applying fuzzy method the sensitive rules are hidden using DSR, PSO, and HYBRID. The results are depicted and compared to show HYBRID is the best method for hiding.

The first experiment finds the relationship between number of total hidden rules, and number of transactions. The results generated for fuzzy, DSR, PSO, and HYBRID are depicted in Fig 3. The HYBRID gives the best hiding when compare to fuzzy, DSR, and PSO.

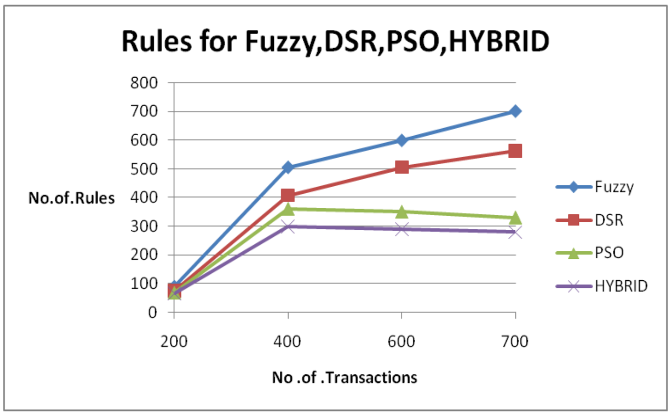

Fig 3: Rules for Fuzzy, DSR, PSO, HYBRID

The second experiment finds the number of modification between transactions for DSR, PSO, and HYBRID. The results are depicted in fig 4. Even though the number of modification for HYBRID is little more than PSO, it provides a best hiding.

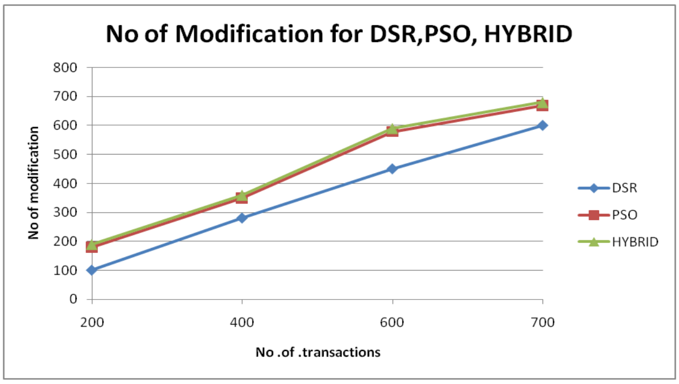

Fig 4: No of Modification for DSR, PSO, HYBRID

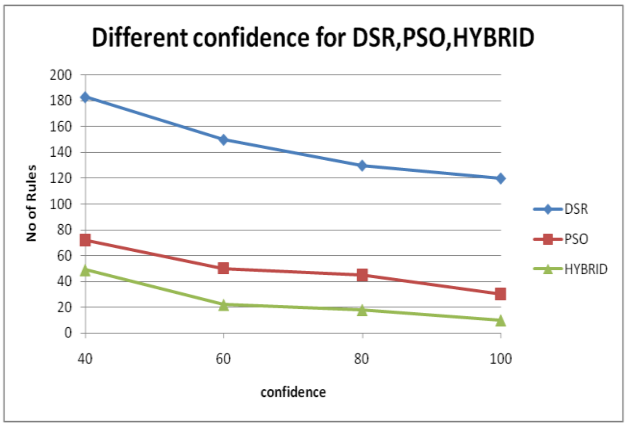

Fig 5: Different confidence for DSR, PSO, HYBRID

The fifth experiment finds Different confidence for DSR, PSO, and HYBRID. The results are obtained by varying the confidence value as 40%, 60%, 80%, and 100% for DSR, PSO, and HYBRID.

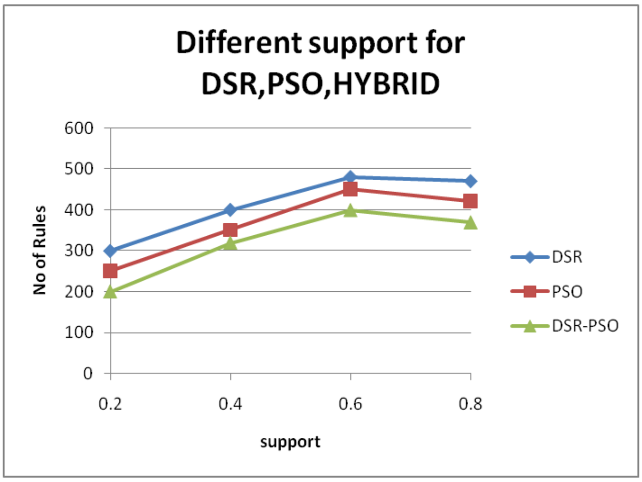

Fig 6: Different support for DSR, PSO, HYBRID

The sixth experiment finds Different support for DSR, PSO, and HYBRID. The results are obtained by varying the support value as 0.2, 0.4, 06, and 0.8 for DSR, PSO, and HYBRID.

With varying support and varying confidence HYBRID proves to be successful in hiding the maximum number of rules. Finally the ghost rule that is a side effect of hiding process which creates a new rule is not observed in our experimental results. Hence HYBRID approach does not have any ghost rules.

5. CONCLUSION

This paper proposes a hybrid algorithm which deals with hiding the association rules in database. The main advantage of proposed system is that it does hiding effectively than other approach with a minimal increase in the number of modifications performed by DSR and PSO approaches. However based on experimental results the number of rules hidden by HYBRID approach is higher than the other two approaches. An experimental result of the proposed approach demonstrates hidden rules for different values of support and confidence with minimum rules lost and no ghost rules generated. HYBRID can be enhanced to reduce the number of modifications. Our approach provides higher hiding with no ghost rule and minimum lost rules, hence the future scope is to reduce the lost rules and reduce the number of modifications.

6. REFERENCES

[1] Rajan Chattamvelli, “Data mining methods”, published by Narosa publishing house in the year 2009

[2] Yi-Hung Wu, Chia-Ming Chiang, and Arbee L.P. Chen, Senior Member, IEEE Computer Society, “Hiding Sensitive Association Rules with Limited Side Effects, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 19, NO. 1, JANUARY 2007

[3] Ila Chandrakar, Yelipe Usha Rani, Mortha Manasa and Kondabala Renuk “Hybrid Algorithm for Privacy Preserving Association Rule Mining” Department of Information Technology,VNR Vignana Jyothi Institute of Engineering and Technology, Hyderabad, India.

[4] https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Original)

[5] Dr. Duraiswamy. K, Dr. Manjula. D, and Maheswari. N “A New Approach to Sensitive Rule Hiding”, ccsenet journal, vol 1, No. 3, August, 107-111