A Hybrid Algorithm for Association Rule Hiding using Representative Rule

Vikram Garg
Sr. Lect., Deptt. of CSE
GGITM
Bhopal

Anju Singh
Asst. Prof., Deptt. of IT
BUIT, BU
Bhopal

Divakar Singh
HOD, Deptt. of CSE
BUIT, BU
Bhopal

ABSTRACT
In the recent years, data mining has emerged as a very popular tool for extracting hidden knowledge from collection of large amount of data. One of the major challenges of data mining is to find the hidden knowledge in the data while the sensitive information is not revealed. Many strategies have been proposed to hide the information containing sensitive data. Privacy preserving data mining is an answer to such challenge. Association rule hiding is one of the PPDM techniques to protect the sensitive association rule generated by Association rule mining (ARM). In this paper, the data distortion technique for hiding the sensitive information is used. The proposed approach uses the concept of Representative Rule (RR) which is used to prune the number of association rule. The proposed algorithm hides the more number of rules while making the fewer database scans.

KEYWORDS
Data Mining, Privacy Preserving, Association Rule Hiding, Representative Rule

1. INTRODUCTION
Privacy has become an increasingly important issue in many data mining application that deal with security, health care, financial and other type of data that is sensitive by nature. The knowledge discovered by various data mining techniques may contain some sensitive information about an individual or organization. For example, for some hospital maintaining database of a patient, it is useful to share the data related to disease, but at the same it also important is to maintain the patient's privacy. In another situation is where shopping malls are trying to understand the purchasing behaviour of the customer. In this case the data related to individual is not important, but the knowledge derived from the database is required to be protected.

Privacy preserving data mining (PPDM) provides solutions to the problem of maintaining the privacy of data as well as knowledge. It allows the extraction of knowledge and also prevents the sensitive data or information from disclosure. PPDM algorithms refer to the techniques used for the selective modification of the data. The selective modification will help us to achieve higher utility for modified data. Association rule hiding (ARH) is the PPDM technique used for hiding the sensitive association rule. All the ARH algorithms aim to modify the data set minimally and yet able to hide the sensitive association rule.

The next section explains the concept of association rule hiding. The Section III explains the related work that has been done in the field of ARH (Association Rule Hiding). The section IV describes the problem identified in literature survey. The next section will introduce the concept of Representative Rule (RR). The Section VI will have the proposed approach and algorithms which will be followed by section VII on experiments and results.

2. ASSOCIATION RULE HIDING
Let $I = \{i_1 \ldots, i_n\}$ be a set of items from a database. Let D be a set of transactions. Each transaction t Є D is an item set such that t is a proper subset of I. A transaction t supports A, a set of items in I, if A is a proper subset of t. An association rule of the form $A \rightarrow B$, where A and B are subsets of I and $A \cap B = \emptyset$. The support denoted as σ of rule $A \rightarrow B$ can be computed by the following equation: Support $(A \rightarrow B) = |AUB| / |D|$, where |AUB| denotes the number of transactions in the database that contains the item set AB, and |D| denotes the number of the transactions in the database D which means that σ% of the transactions in D supports item set AB. The confidence denoted as τ of rule $A \rightarrow B$ is calculated by following equation: Confidence $(A \rightarrow B) = |AUB|/|A|$, where |A| is number of transactions in database D that contains item set A which means τ% of the transactions in D that supports A also supports B. A rule $A \rightarrow B$ is strong if support $(A \rightarrow B) \geq$ min_support and confidence $(A \rightarrow B) \geq$ min_confidence, where min_support and min_confidence are two given minimum thresholds [12].

Association rule hiding algorithms prevents the sensitive rules from being disclosed. The problem of association rule hiding can be stated as follows: "Given a transactional database D with minimum confidence, minimum support and a set R of rules which have been mined from database D. A subset R_H of R is denoted as set of sensitive association rules which have to be preventing from being disclosed. The objective of association rule hiding is to transform D into a database D' in such a way that nobody will be able to mine association rule which belongs to R_H and all non sensitive rules in R should remain unaffected[12].

3. LITERATURE SURVEY
The concept of privacy preserving in data mining came in to existence in response to the concerns that were raised for preserving the private information which are produced as a result of data mining algorithms [2][1]. There are two types of privacy concern that were raised in reference to the data mining. The first type of privacy concern termed as output privacy is that the data is minimally altered so that the mining result will preserve privacy. Many techniques have been proposed for this type of output privacy [2][3].Techniques like blocking, perturbation, aggregation, swapping, and sampling are the example of output privacy. In output privacy for hiding a given specific rules or patterns, there are many proposed techniques available for hiding association rule, classification and clustering rules. For hiding the association rules, two approaches have been proposed. The first approach that has been proposed hides one rule at a time [12]. It first selects transactions that contain the items in a give rule. It then attempts to modify transaction by transaction until the

support or confidence of the rule fall below minimum support or minimum confidence. The modification is done by either deleting items from the transaction or adding new items to the transactions.

The second type of privacy concern which is related with the input privacy of the data is that the data is altered in such a way that the mining result is not affected or affected minimally [4], like cryptography-based techniques which allow users access to only a subset of data while global data mining results can still be discovered. The example includes multiparty computation. The second approach deals with groups of restricted patterns or association rules at a time [10]. It first selects the transactions that contain the intersecting patterns of a group of restricted patterns. After that on the basis of disclosure threshold supplied by users, it hides the restricted patterns by sanitizing the percentage of the selected transactions. In [7] authors summarize the advantages and limitations of associations hiding approaches.

In [1] the authors discussed three algorithms for hiding sensitive association rules. Algorithm 1 hides association rules by increasing the support of the rule's antecedent until the rule confidence decreases below the minimum confidence threshold. Algorithm 2 hides sensitive rules by decreasing the frequency of the consequent until either the confidence or the Support of the rule is below the threshold. Algorithm 3 decreases the support of the sensitive rules until either their confidence is below the minimum confidence threshold or their support is below the minimum support threshold. In algorithm 1 large number of new frequent item sets is introduced and, therefore, an increasing number of new rules are generated. Algorithm 2 and 3 affects number of no sensitive rules in database due to removal of items from transaction

In [15] authors proposed two algorithms, DCIS (Decrease Confidence by Increase Support) and DCDS (Decrease Confidence by Decrease Support) were introduced to automatically hide association rules without pre-mining and selection of hidden rules. In [14] the authors speak about ISL (Increase Support of LHS) and DSR (Decrease Support of RHS). Item sets are given as input to both the algorithms to automatically hide sensitive association rules without pre-mining and selection of hidden rules. The ISL and DCIS algorithms try to increase the support of left hand side of the association rule and algorithms DSR and DCDS try to decrease the support of the right hand side of the association rule. It is observed that the running time of ISL is more than DSR. Also both algorithm exhibit contrasting side effects. In [9] authors in their paper discussed a heuristic algorithm DSRRC (Decrease Support of R.H.S. item of Rule Clusters) which provides privacy for sensitive rules at certain level while maintains data quality. DSRRC algorithm clusters the sensitive association rules based on R.H.S. of rules and hides all possible rules at a time by modifying lesser number of transactions which helps maintaining data quality. DSRRC algorithm cannot hide rules having multiple RHS items.

In [13] an algorithm DSC (Decrease Support and Confidence) is proposed in which only one scan of database is required because pattern-inversion tree is used to store related information. The proposed algorithm can automatically sanitize informative rule sets without pre-mining and selection of a class of rules under one database scan.

In [11] the authors discussed about four heuristic algorithms: Algorithm Naïve, MinFIA (Minimum Frequency Item Algorithm), MaxFIA (Maximum Frequency Item Algorithm). The Naive Algorithm removes the entire items with the highest frequency in the database of selected transaction. In MinFIA (Minimum Frequency Item Algorithm) algorithm the item with the smallest support in the pattern chosen as a sensitive item and it removes that item from the sensitive transactions. Unlike the MinFIA, algorithm MaxFIA (Maximum Frequency Item Algorithm) selects the item with the maximum support in the pattern as a sensitive item and removes it from the transaction.

In [5] a Hybrid algorithm is proposed that uses the combination of ISL and DSR technique and hides the association rules by modifying the database transactions so that the confidence of the association rules can be reduced. Such approach will provide better result than using either ISL or DSR. In [6] the support &confidence of the association rules remains unchanged because the transactional database is not modified. It prunes more number of hidden rules and scans the database less number of times. In [16] the authors introduced an efficient algorithm known as FHSAR (Fast Hiding Sensitive Association Rules), for fast hiding of sensitive association rules. The algorithm is capable of hiding any given sensitive association rule by scanning database once, which reduces the execution time significantly. The IGA [11] algorithm (Item Grouping algorithm) groups restricts the patterns in groups of patterns sharing the same item sets so that all sensitive patterns in the group will be hidden in single step.

4. PROBLEM DISCRIPTION

All the algorithms that have been discussed in the above sections are being utilized for the purpose of sensitive item set hiding for a long time and across all the domains. Majority of the algorithms from the above section hides the sensitive information but has some implications on the data set like introduction of new rules, lost association rule and hiding failures. Algorithms discussed in the previous section mainly focussed on hiding the sensitive association rule without looking at the fact that how many database scans 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 databases scans.

5. REPRESENTATIVE RULE

A set of representative association rules with respect to minimum support S and minimum confidence C will be denoted by RR(S,C) and defined as follows:

$$RR(s,c) = \{ r \in AR(S,C) \mid \neg \exists r' \in AR(S,C), r' \neq r \text{ and } r \in C(r') \}$$

If S and C are understood then RR(S,C) will be denoted by RR. Each rule in RR is called a representative association rule. By the definition of RR no representative association rule may belong in the cover of another association rule [8].

A notion of a cover operator for deriving a set of association rules from a given association rule without accessing a database. The cover C of the rule $X \Rightarrow Y$, $Y \neq \emptyset$ is defined as follows:

$$C(X \Rightarrow Y) = \{ X \cup Z \Rightarrow V \mid Z,V \subseteq Y \text{ and } Z \cap V = \emptyset \text{ and } V \neq \emptyset \}$$

Each rule in C ($X \Rightarrow Y$) consists of a subset of items occurring in the rule $X \Rightarrow Y$. The antecedent of any rule r covered by $X \Rightarrow Y$ contains X and perhaps some items from Y, whereas r's consequent is a non-empty subset of the remaining items in Y. It was proved that each rule r in the cover C(r'), where r' is an

association rule having support s and confidence c, belongs in AR(S, C). Hence, if r belongs in AR(S,C) then every rule r' in C(r) also belongs in AR(S,C). The number of different rules in the cover of the association rule X⇒Y is equal to $3^m - 2^m$, where m = |Y|.

Let T1 = {A,B,C,D,E}, T2 = {A,B,C,D,E,F}, T3 = {A,B,C,D,E,H,I}, T4 = {A,B,E} and T5 = {B,C,D,E,H,I} are the only transactions in the database D. Let r: (B ⇒CDE). TABLE 1 contain all rules belonging in the cover C(r) along with their support and confidence in D. The support of r is equal to 40% and its confidence is equal to 80%. The support and confidence of all other rules in C(r) are not less than the support and confidence of r.

Table 1: The cover of the rule r: (B⇒CDE)

#	Rule r' in C(r)	Support of r'	Confidence of r'
1.	B ⇒CDE	80%	80%
2.	B ⇒CD	80%	80%
3.	B ⇒CE	80%	80%
4.	B ⇒DE	80%	80%
5.	B⇒C	80%	80%
6.	B⇒D	80%	80%
7.	B⇒E	100%	100%
8.	BC ⇒ DE	80%	100%
9.	BC⇒D	80%	100%
10.	BC⇒E	80%	100%
11.	BD ⇒ CE	80%	100%
12.	BD⇒C	80%	100%
13.	BD⇒E	80%	100%
14.	BE ⇒CD	80%	80%
15.	BE⇒C	80%	80%
16.	BE ⇒D	80%	80%
17.	BCD ⇒ E	80%	100%
18.	BCE ⇒ D	80%	100%
19.	BDE ⇒C	80%	100%

Given minimum support S = 40% and minimum confidence C= 80%, the following representative rules would be found for the database D from Example stated above:

RR (40%, 80%) = {B⇒CDE}.

6. PROPOSED APPROACH

This section discusses the proposed methodology that will be used to improve the performance of the association rule hiding process. The following diagram will explain the working of the proposed methodology.

In the figure there is sample database on which the frequent item set algorithm is applied. After generation of frequent item set the representative rules are generated. Then the association rule hiding algorithm is applied which is a hybrid algorithm and is a combination ISL and DSR. After applying the algorithm the modified database is produced.

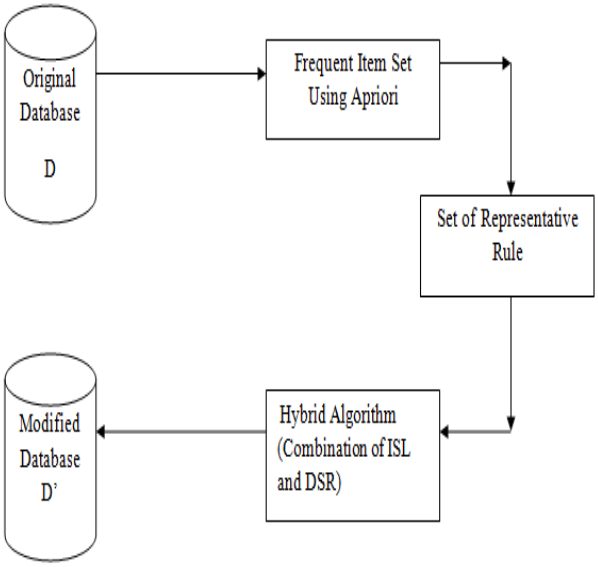

Fig 1: Working of Proposed Algorithm

Hybrid algorithm Using Representative Rules(HRR):

Input:
1. A source database D,
2. A minimum support min_support,
3. A minimum confidence min_confidence,
4. A set of hidden items *A*.

Output:
A transformed database D, where rules containing *A* on Left Hand Side (LHS) or Right Hand Side (RHS) will be hidden.

Steps of algorithm:

1. Generate all Frequent Item Set F_k from A Using Apriori Algorithm;
2. For all ($Z \in F_k$, k >= 2) do begin
3. Max_Sup = max({sup(Z') l $Z \subset Z \in F_{k+1}$} ∪{0});
4. if Z. sup ≠Max_Sup then begin
5. A_1 = {{Z[1] }, {Z[2] } .....Z[k]]}; //create 1-Antecedents
6. for (i = I; (Ai ≠ Ø) and (i< k); i++) do begin
7. for all X ∈ Ai do begin
7.1 find Y∈ Fi such that Y = X;
7.2 X.Count = Y. count;
7.3 if (Z. count/XCount > c) then begin
7.4 if (Max_Sup/X.Count < c) then
7.5 print(X ⇒" Z\X with support: ", Z. count, and confidence: " Z.count / X.Count");
7.6 Ai = Ai \ {X}
7.7 End if
7.8 End for
7.9 A_{i+1} = Apriori_Gen(Aj)
8. End for
9. End if
10. End for
11. Compute confidence of all the Representative rules.
12. for each hidden item h
13. For each rule containing h, compute confidence of rule R
14. For each rule R in which h is in RHS
14.1.1 If confidence (R) < min conf, then Go to next RR;
14.1.2 Else go to step 6
15. Decrease Support of RHS i.e. item h.
15.1 Find T = t in D | t fully support R;

15.2 While (T is not empty)
15.2.1 Choose the first transaction t from T;
15.2.2 Delete the item set which is in RHS Item of RR;
15.2.3 End While
15.3 Compute confidence of R;
15.4 If T is empty, then h cannot be hidden;
End For
16. For each rule R in which is in LHS
17. Increase Support of LHS;
17.1 Find T = t in D | t does not support R;
17.2 While (T is not empty)
17.2.1 Choose the first transaction t from T;
17.2.2 ADD the item set which is in LHS
Item of RR;
17.2.3 End While
17.3 Compute confidence of R;
17.4 If T is empty, then h cannot be hidden;
End For
End Else
End For
18. Output updated D, as the transformed D;

The proposed algorithm can be illustrated with the following examples for the given transactional data set given in TABLE 2

Table 2: Transaction Dataset for Analysis

TID	Item Set
T_1	ABCDE
T_2	ABCDEF
T_3	ABCDEHI
T_4	ABE
T_5	BCDEHI

After the proposed algorithm is applied to the given data set and let the minimum support S=40% and minimum confidence C= 80%. The number of association rule generated by the above transactional database is 93. The representative rules for the given data set using the cover operator is 9. The RR (Representative Rule) generated by the given data set are as follows

Table 3: Representative Rule

Min Support and Min Confidence	Representative Rule
RR (40%, 80%)	AC⇒BDE, AD⇒BCE, B⇒CDE, C⇒BDE, D⇒BCE, E⇒BCD, A⇒BE, B⇒AE, E⇒AB

The support and confidence for the given database is 40% and 80% respectively. Let sensitive item set H= {C}. Now choose the representative rules containing the 'C' in RHS. From the set of RR, one can find there are

Table 4: RR with C in R.H.S.

RR	Support(%)	Confidence (%)
AD⇒BCE	60	100
B⇒CDE	80	80
D⇒BCE	80	100
E⇒BCD	80	80

From the above rules select AD⇒BCE and from the transactional data set find the transactions which fully supports rule which are {T_1, T2, T_3}. Now delete the sensitive item 'C' from all the transactions and the transactional data set will be modified.

Table 5: Modified Data set 1

TID	Item Set
T_1	ABDE
T_2	ABDEF
T_3	ABDEHI
T_4	ABE
T_5	BCDEHI

Now choose the representative rules containing the 'C' in LHS. From the set of RR, one can find there are:

Table 6: RR with C in L.H.S.

RR	Support(%)	Confidence (%)
C⇒BDE	80	100

Now delete the item 'C' from transaction T_5, which fully supports the representative rule. After applying hiding using DSR and ISL, the modified database is as follows:

Table 7: Modified Data set

TID	Item Set
T_1	ABDE
T_2	ABDEF
T_3	ABDEHI
T_4	ABE
T_5	BCDEHI

Now from the above database if the representative rule algorithm is applied, than the rules containing sensitive item 'C' will not be displayed. So the sensitivity of item 'C' is maintained using the provided algorithm.

7. EXPERIMENTS & RESULTS

The proposed algorithm is implemented in java using eclipse framework. The algorithm implemented on windows 8 64 bit platform (Intel(R) Core(TM) i5-3230M CPU @2.60GHz) and 4GB RAM. The algorithm is tested on various dataset of different size.

The following datasets that is used for the analysis the performance and behavior of the proposed algorithm were taken from the UCI datasets and PUMSB. This data set is used for the analysis with Minimum Support as 40% and Minimum Confidence as 50%.

- chess (5KB)
- mushroom (15KB)
- Hepatitis (25KB)
- Pumsb (30KB)
- Accident (35KB)

The proposed algorithm is applied to these datasets and parameters like number of association rule, time taken and the memory requirement of the proposed and existing algorithm (apriori with ISL and DSR) are recorded. Using the data for different parameters, different graphs are constructed.

The figure 2 shows the number of association rule generated by existing algorithm and the proposed algorithm using data set of various domains and of various sizes. From the graph it is clear that the numbers of representative rules are always less as compared to association rule.

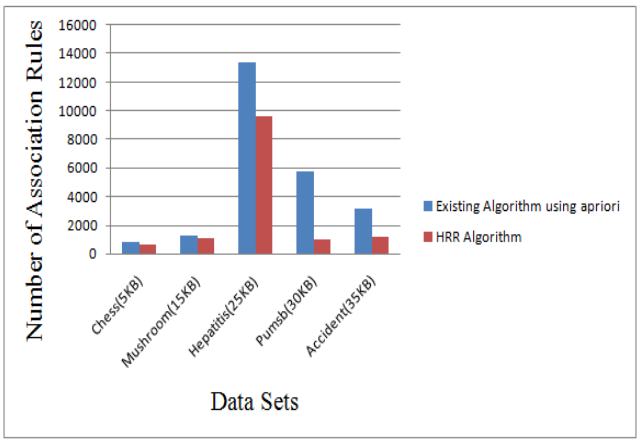

Fig 2: Result Analysis of Existing Algorithm & HRR Algorithm

The figure 3 shows the execution time by the existing algorithm and the proposed algorithm. The execution time for both the algorithm is measured in milliseconds. From the graph it is found that the execution time by proposed algorithm is always less than the existing algorithm.

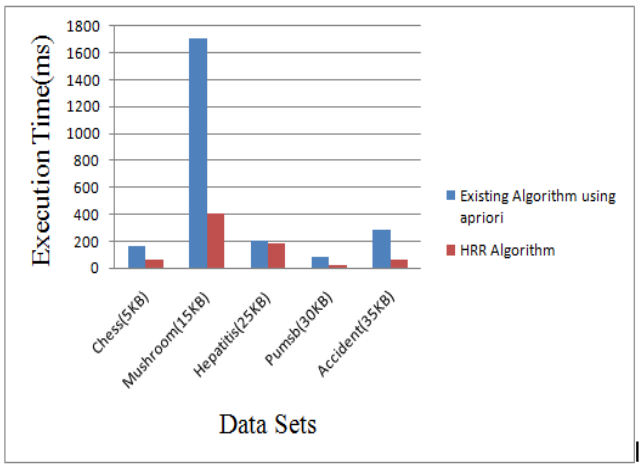

Fig 3: Result Analysis of Existing Algorithm & HRR Algorithm

The figure 4 shows the memory requirement for both the algorithm and it is measured in megabyte (MB). From the graph it is found that the memory requirement by proposed algorithm is always uses more memory than the existing algorithm.

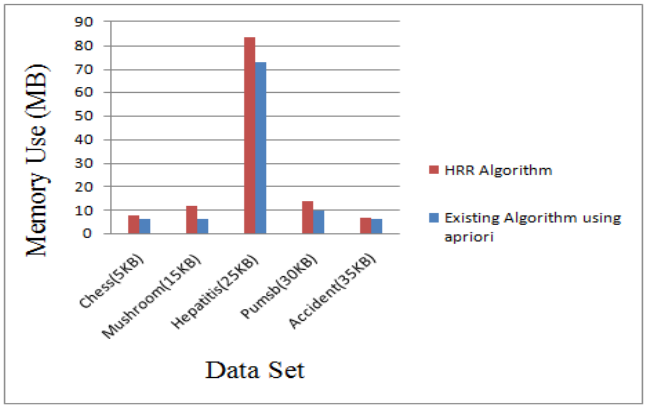

Fig 4: Result Analysis of Existing Algorithm & HRR Algorithm

8. CONCLUSIONS

The result in the previous section shows that the proposed algorithms works better as compared to the Hybrid algorithms without representative rules. The proposed algorithm performs better in terms of number of rules and time taken by the algorithm. The memory utilization of the algorithm increases.

In future, this algorithm can be applied in Secure Multiparty Computation. The algorithm can also be implemented by Hadoop using Map Reduce Framework. This algorithm can be converted into distributed algorithm so it can be used in the distributed environment.

9. REFERENCES

[1] Agrawal R. & Srikant R., "Privacy preserving data mining" , In ACM SIGMOD conference on management of data, pp. 439-450, 2000.[11]

[2] Clifton C., "Protecting against data mining through samples" In Proceedings of the thirteenth annual IFIP WG 11.3 working conference on database security, pp. 193-207, 1999.[1]

[3] Clifton C., "Using sample size to limit exposure to data mining", Journal of Computer Security, Vol.8, pp. 281-307, 2000.[2]

[4] Dasseni E. , Verykios V. , Elmagarmid A. & Bertino E., "Hiding association rules by using confidence and support" In Proceedings of 4th information hiding workshop, pp. 369-383,2001.[5]

[5] Dhutraj Niteen ,Sasane Siddhart,Kshirsagar Vivek, "Hiding Sensitive Association Rule for Privacy Preservation", IEEE Transactions on Knowledge And Data Engineering, 2013.[9]

[6] Gulwani Padam,"Association Rule Hiding by Positions Swapping of Support and Confidence", International journal of Information Technology and Computer Science, Vol. 4, pp. 54-61, 2012.[10]

[7] Jadav Khyati B.,Vania Jignesh,Patel Dhiren R. , "A Survey on Association Rule Hiding Methods", International Journal of Computer Applications, Vol. 82, pp. 20-25, 2013.[6]

[8] Kryszkiewicz, M., "Representative Association Rules and Minimum Condition Maximum Consequence Association Rules", PKDD, Vol. 1510, pp. 361-369, Springer, 1998.[8]

[9] Modi C.N. , Rao U.P. & Patel D.R., "Maintaining privacy and data quality in privacy preserving association rule mining", International Conference on Computing Communication and Networking Technologies (ICCCNT),pp. 1-6, 2010.[3]

[10] Oliveira S. & Zaiane O., "Algorithms for balancing privacy and knowledge discovery in association rule mining" , In Proceedings of 7th international database engineering and applications symposium (IDEAS03), pp. 54-63, 2003.[13]

[11] Oliveira Stanley R. M.,Osmar R. Zaiane, "Privacy Preserving Frequent Itemset Mining", IEEE International Conference on Data Mining Workshop on Privacy, Security, and Data Mining,Vol. 14, pp. 19-26, 2002.[12]

[12] Shah Komal ,Thakkar Amit ,Ganatra Amit," A Study on Association Rule Hiding Approaches", International

Journal of Engineering and Advanced Technology (IJEAT), 2012.[7]

[13] Wang Shyue-Liang,Maskey Rajeev,Jafari Ayat & Hong Tzung-Pei,"Efficient sanitization of informative association rules" ,ACM, Expert Systems with Applications: An International Journal, 2008.[16]

[14] Wang Shyue-Liang,Parikh Bhavesh,Jafari Ayat, "Hiding informative association rule sets", ELSEVIER, Expert Systems with Applications, pp. 316-323, 2007.[14]

[15] Wang Shyue-Liang,Patel Dipen,Jafari Ayat & Hong Tzung-Pei, "Hiding collaborative recommendation association rules" ,Springer Science+Business Media, LLC 2007.[15]

[16] Weng Chih-Chia ,Chen Shan-Tai & Lo Hung-Che "A Novel Algorithm for Completely Hiding Sensitive Association Rules" ,IEEE Intelligent Systems Design and Applications, Vol. 2, pp. 202-208 2008.[4]