ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 1, Issue 7, September 2012

Design and Implement an Association Rule hiding Algorithm for Privacy Preserving Data Mining

R.Natarajan[1], Dr.R.Sugumar[2], M.Mahendran[3], K.Anbazhagan[4]

Research Scholar, Dept. of CSE, CMJ University, Shillong, Meghalaya, India[1]

Associate Professor, Dept. of CSE, VelMultitech Dr. RR Dr.SR Engineering College, Chennai, India[2]

Research Scholar, Dept. of CSE, CMJ University, Shillong, Meghalaya, India[3]

Research Scholar, Dept. of CSE, CMJ University, Shillong, Meghalaya, India[4]

ABSTRACT: Privacy preserving data mining is a research area concerned with the privacy driven from personally identifiable information when considered for data mining. This paper addresses the privacy problem by considering the privacy and algorithmic requirements simultaneously. The objective of this paper is to implement an association rule hiding algorithm for privacy preserving data mining which would be efficient in providing confidentiality and improve the performance at the time when the database stores and retrieves huge amount of data. This paper compares the performance of proposed algorithm with the two existing algorithms namely ISL and DSR.

Keywords: **Confidence, Support, association rules, Item sets**

I. INTRODUCTION

In data mining, association rule learning is a popular and well researched method for discovering interesting relations between variables in large databases. Piatetsky et al.(2007) describes analysing and presenting strong rules discovered in databases using different measures of interestingness. Based on the concept of strong rules, Agrawal et al. (2008) introduced association rules for discovering regularities between products in large scale transaction data recorded by point-of-sale (POS) systems in supermarkets. Association rule hiding refers to the process of modifying the original database in such a way that certain sensitive association rules disappear without seriously affecting the data and the non sensitive rules. The objective of the proposed Association rule hiding algorithm for PPDM is to hide certain information so that it cannot be discovered through association rule mining algorithm.

Association rules are usually required to satisfy a user-specified minimum support and a user-specified minimum confidence at the same time. Association rule generation is usually split up into two separate steps:

1. First, minimum support is applied to find all frequent item sets in a database.
2. Second, these frequent item sets and the minimum confidence constraint are used to form rules (Yıldız et al. 2010).

There are three types of Association rule hiding algorithms namely border-based approaches, exact approaches and heuristic approaches. Heuristic approaches are very efficient and fast algorithms that modify the selected transactions from the database for hiding the sensitive knowledge. There are two types of heuristic approaches exist namely data blocking method and data distortion method. This chapter describes about the types of heuristic approaches and compares the performance with the proposed weight based sorting distortion algorithm in terms of hiding failure and data quality (Yıldız et al. 2010).

Given a rule r and calculate minconf(r), maxconf(r) as
minconf(r) = minsup(r)*100 / maxsup (lr) ---------- (1)
maxconf(r) = maxsup(r)*100 / minsup (lr) ---------- (2)
Where lr denotes the rule antecedent.

Considering the support interval and the minimum support threshold have the following cases for an itemset A:

(i) A is hidden when maxsup(A) is smaller than MST
(ii) A is visible with an uncertainty level when

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 1, Issue 7, September 2012

minsup (A) ≤ MST ≤ maxsup(A)
(iii) A is visible if minsup(A) is greater than or equal to MST

The Figure 4.1 shows the architecture of association rule hiding. A rule hiding process takes place according to two different strategies: decreasing its support or its confidence. In this method, the adopted alternative strategies aim at introducing uncertainty in the frequency or the importance of the rules to hide. The two strategies reduce the minimum support and the minimum confidence of the item sets generating these rules below the Minimum Support Threshold (MST) and Minimum Confidence Threshold (MCT) correspondingly by a certain Safety Margin Threshold (SMT) fixed by the user. In order to reduce the support of the large itemset generating a sensitive rule, Algorithm 1 replaces 1's by " ? " for the items in transactions supporting the itemset until its minimum support goes below the minimum support threshold MST by the fixed safety margin SM (Fovino et al. 2008).

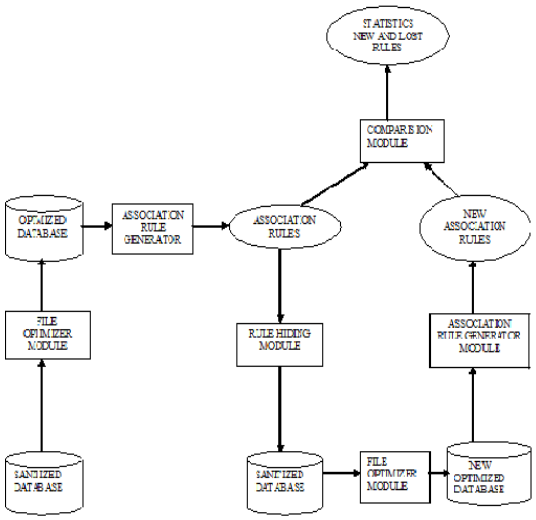


Figure 4.1 Architecture of Association Rule Hiding

The first rule decreases the minimum support of the generating itemset of a sensitive rule by replacing items of the rule consequent with unknown values. Whereas the second rule increases the maximum support value of the antecedent of the rule to hide via placing question marks in the place of the zero values of items in the antecedent. All the algorithms hide a sensitive rule with an uncertainty level by decreasing the minimum support or confidence values below the resulting thresholds, MST-SM and MCT-SM (Kshitij Pathak et. al 2009).

II.DATA BLOCKING METHOD

Data-Blocking is a type of heuristic method for association rule hiding. Instead of making data distorted (part of data is altered to false), blocking approach is implemented by replacing certain data items with a question mark "?". The introduction of this special unknown value brings uncertainty to the data, making the support and confidence of an association rule become two uncertain intervals respectively. At the beginning, the lower bounds of the intervals equal to the upper bounds. As the number of "?" in the data increases, the lower and upper bounds begin to separate gradually and the uncertainty of the rules grows accordingly. When either of the lower bounds of a rule's support interval and confidence interval gets below the security threshold, the rule is deemed to be concealed (Griffith et. al 2007).

The introduction of this new question mark "?" in the dataset, imposes some changes on the definition of the support and confidence of an association rule. In this regard, the minimum support and minimum confidence will be altered into a minimum support interval and a minimum confidence interval correspondingly. As long as the support and/or the confidence of a sensitive rule lie below the middle in these two ranges of values, then it expect that the confidentiality of data is not violated.

In the Blocking based algorithms the idea is to substitute the value of an item supporting the rule with a meaningless symbol. The algorithm describes the results of a blocking algorithm which reduces loss of data and minimizes the undesirable side effects by selecting the items in the appropriate transactions to change, and maximize the desirable side effects. To modify the database in a way that an adversary cannot recover the original values of the database. The data blocking method produces more side effects whenever the MCT and MST are increased and decreased (Griffith et. al 2007).

III. TYPES OF HIDING STRATEGIES

The hiding strategies heavily depend on finding transactions that fully or partially support the generating item sets of a rule. Because if a rule has to be hidden, need to change the support of some part of the rule, that is, should decrease the support of the generating item set. Again, as mentioned in the previous section, the changes in the database introduced by the hiding process should be limited,

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 1, Issue 7, September 2012

in such a way that the information loss incurred by the process has reduced. So, the system tries to apply small changes in the database at every step of the hiding algorithms (Verykios et. al 2004).

The decreasing the support of an item set S can be done by selecting a transaction t, that supports S and by setting to 0 at least one of the non-zero values that represent items in S. The increase in the support of an item set S can be accomplished by selecting a transaction t that partially supports it and setting to 1 (Bertino et. al 2006).

If the formulas are analyzed for determining support and confidence values (mentioned in previous section), the system can find that there can be two ways to reduce the support and confidence of a rule. Both the confidence and the support are expressed as ratios of supports of item sets that support the two parts of a rule or its generating item set. In this way, if the value of a ratio has to be decreased, the system uses one of the following options:
(a) Decrease the numerator, while keeping the denominator fixed, or
(b) Increase the denominator while keeping the numerator fixed.
Considering the case of decreasing support value, the support S of a rule X => Y is given by

$$\frac{|X \cup Y| * 100}{N} > S \quad \text{---------- (1)}$$

Since N is constant (as it is the number of transactions in the given database), the only option left for this is to change the numerator value (option (a)) and decrease the support of any rule by decreasing the support of the generating item set of the rule.
Considering the case of decreasing confidence value, Confidence C of a rule X => Y is given by

$$\frac{|X \cup Y| * 100}{|X|} > S \quad \text{---------- (2)}$$

Now, analyze each of the options separately to check which of them (or both) works in the current context.

The first option implies to decrease the numerator (which is the support) of the generating item set of the rule, while the support of the item set in the left hand side of the rule remains fixed. In order to do that, decrease the support of the generating item set of the entire rule by modifying the transactions that support this item set, making sure that it hides items from the consequent or the right hand side of the rule. This will decrease the support of the generating item set of the rule, while it will leave unchanged the support of the left hand side or else the denominator.

The second option implies to increase the denominator (which is the support of the item set in the antecedent) of the rule, while the support of the generating item set of the rule remains fixed. This option is also applicable, since increase the support of the rule antecedent while keeping the support of the generating item set fixed by modifying the transactions that partially support the item set in the antecedent of the rule but do not fully support the item set in the consequent.

So briefly state the strategies as, given a rule X => Y on a database D, the support of the rule in D expresses the probability to find transactions containing all the items in X U Y. The confidence of X => Y is, instead, the probability to find transactions containing all the items in X U Y that contain X.

- Decrease the confidence of a rule by increasing the support of the rule antecedent X through transactions that partially support it or by decreasing the support of the rule consequent Y in transactions that support both X and Y

- Decrease the support of a rule by decreasing the support of either the rule antecedent X or the rule consequent Y, through transactions that fully support the rule.

IV. DESIGN OF ASSOCIATION RULE HIDING ALGORITHM

The objective of association rule hiding algorithm is to hide certain confidential data so that they cannot be discovered through data mining techniques. In this research work, it is assumed that only sensitive items are given and propose one algorithm to modify data in database so that sensitive items cannot be deduced through association rules mining algorithms. More specifically, given a transaction database D, a minimum support, a minimum confidence and a set of items H to be hidden, the objective is to modify the database D such that no association rules containing H on the right hand side or left hand side will be discovered.

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 1, Issue 7, September 2012

The proposed association rule hiding algorithm is based on two algorithms namely ISL (Increase Support of Left hand side) and DSR (Decrease Support of Right hand side) to hide useful association rule from transactions data with binary attributes. In ISL method, confidence of a rule is decreased by increasing the support value of Left Hand Side (LHS) of the rule.

For this purpose, only the items from LHS of a rule are chosen for modification. In DSR method, confidence of a rule is decreased by decreasing the support value of Right Hand Side (R.H.S.) of a rule. For this purpose, only the items from R.H.S. of a rule are chosen for modification.

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 increases the support of X, the LHS of the rule, but not support of X U Y, or decrease the support of the item set 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 X U 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.

Based on these two concepts, a new association rule hiding algorithm for hiding sensitive items in association rules has been proposed. In the proposed algorithm, a rule X → Y is hidden by decreasing the support value of X U Y and increasing the support value of X. That can increase and decrease the support of the LHS and RHS item of the rule correspondingly.

This algorithm first tries to hide the rules in which item to be hidden i.e., X is in right hand side and then tries to hide the rules in which X is in left hand side. For this algorithm t is a transaction, T is a set of transactions, R is used for rule, RHS (R) is Right Hand Side of rule R, LHS (R) is the Left Hand Side of the rule R, Confidence (R) is the confidence of the rule R, a set of items H to be hidden.

INPUT**:** A source database D, A minimum support min_support (MST), a minimum confidence min_confidence (MCT), a set of hidden items X.

OUTPUT**:** The sanitized database D, where rules containing X on LHS or RHS will be hidden.

1. Begin
2. Generate all possible rule from given items X;
3. Compute confidence of all the rules for each hidden item H, compute confidence of rule R.
4. For each rule R in which H is in RHS
   4.1 If confidence (R) < MCT, then
   Go to next 2-itemset;
   Else go to step 5
5. Decrease Support of RHS item H.
   5.1 Find T=t in D fully support R;
   5.2 While (T is not empty)
   5.3 Choose the first transaction t from T;
   5.4 Modify t by putting 0 instead of 1 for RHS item;
   5.5 Remove and save the first transaction t from T; End While
6. Compute confidence of R;

Table 4.1 Association Rule Hiding Algorithm

Consider the Table 4.2 as a database, MST=33%, MCT=70%, each element has value 1 if the corresponding item is supported by the transaction and 0 otherwise. Size means the number of elements in the list having value 1.

Transaction id	Items	Rule	Size
T1	ABC	111	3
T2	ABC	111	3
T3	ABC	111	3
T4	AB	110	2
T5	A	100	1
T6	AC	101	2

Table 4.2 Database D with MST=33% and MCT=70%

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 1, Issue 7, September 2012

Transaction id	ABC	Size
T1	011	2
T2	011	2
T3	011	2
T4	010	2
T5	100	1
T6	101	2

Table 4.3 Confidence and rules of table D

In the table 4.3, suppose if item A has to be hidden, first consider rule in which A is in RHS. These rules are B→A and C→A both has greater confidence from MCT. Then consider rule B→A search for transaction which support both B and A, B=A=1. There are four transactions T1, T2, T3, T4 with A=B=1. Now update table put 0 for item A in all four transactions. Now calculate confidence of B→A, it is 0% which is less than MCT so now this rule is hidden.

Now consider rule C→A, search for transaction in which A=C=1, only transaction T6 has A=C=1, update transaction by putting 0 instead 1 in place of A. Now assume the rules in which A is in LHS. There are two rules A→B and A→C but both rules have confidence less than MCT so there is no need to hide these rules. So Table 4.4 shows the modified database after hiding item A.

Rule	Confidence
A→B	66.6 %
A→C	66.6%
B→A	100%
B→C	75%
C→A	100%
C→B	75%

Table 4.4: Updated database after hiding item A

V. SIMULATION AND RESULTS

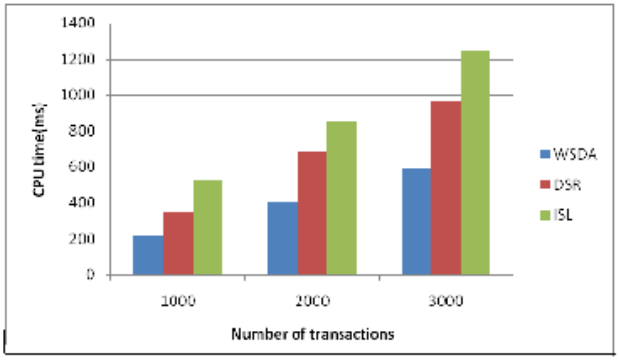
The proposed system has performed the experiments using the ARMADA tool (James Malone et al. 2008). The system has performed four different experiments to compare the performance of proposed
algorithm with ISL and DSR algorithm. For each data set, various sets of association rules are generated under various minimum supports and minimum confidences. The minimum support range is from 10% to 30%. The minimum confidence range is from 40% to 70%. The first experiment shows the relationship between CPU time, number of modified entries and number of transactions. Table 4.5 shows the experimental results. In this experiment, the minimum confidence value is set 70% and minimum support values are taken as 10, 20, and 30 for 500, 1000 and 1500 transactions respectively.

Number of Transactions	CPU Time (milliseconds)			Number of Modified Entries		
	ISL	DSR	WSDA	ISL	DSR	WSDA
1000	842	688	425	683	575	372
2000	1655	1337	827	1297	982	764
3000	2567	2153	1273	1980	1442	1127

Table 4.5 Experimental results of ISL, DSR and WSDA

The Table 4.5 shows the experiment results of ISL, DSR and WSDA for MCT = 70% and 1000, 2000, 3000 transactions. It is clear that WSDA method's CPU time has reduced to 37%, 39%, 41% for 500, 1000, 1500 transactions compared to the existing methods ISL and DSR.

Table 4.6 Comparison of hidden rules for ISL, DSR and WSDA



The Figure 4.2 shows the experiment results of ISL, DSR and WSDA for MCT = 70% and 1000, 2000, 3000 transactions. It is clear that WSDA method's CPU time has reduced to 37%, 39%, 41% due to MCT value changed for 1000, 2000, 3000 transactions compared to the existing methods ISL and DSR.

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 1, Issue 7, September 2012

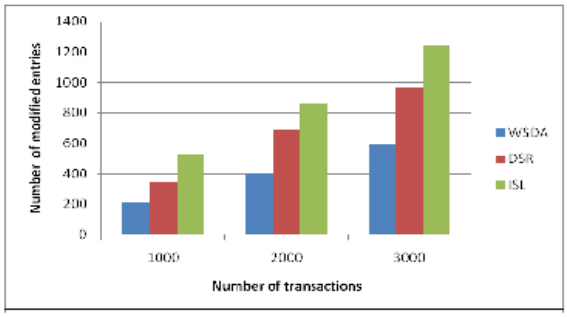


Figure 4.3 Performance comparison of WSDA with DSR and ISL for modified entries with number of transactions

The Figure 4.3 shows the experiment results of ISL, DSR and WSDA with number of modified entries for MCT = 70% and 1000, 2000, 3000 transactions. It is clear that WSDA method's number of modified entries has reduced to 34%, 37%, 39% due to MCT value changed for 1000, 2000, 3000 transactions compared to the existing methods ISL and DSR.

Number of Transactions	Number of Rules Hidden		
	ISL	DSR	WSDA
1000	5	11	19
2000	12	17	28
3000	21	26	37

The Table 4.6 shows the experiment results of number of rules ISL, DSR and WSDA. The numbers of hidden rules are increased to 42%, 60%, 64% due to increase of confidence rules for 1000, 2000, 3000transactions compared to DSR. The Figure 4.4 shows the number of hidden rules for proposed WSDA compared with DSR and ISL.

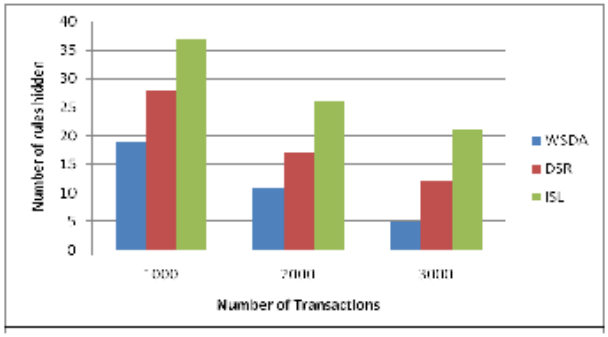


Figure 4.4 Comparative analysis of hidden rules for ISL, DSR and WSDA

The first characteristic has observed the total number of rules hidden for different values of support and confidence. Table 4.6 shows the relationship between number of hidden rules and number of transactions, and shows the relationship between the numbers of hidden rules for different values (10, 20, and 30) of minimum support. From this experiment results, it can be easily seen that the proposed algorithm hides more rules in comparison to previous work for different value of minimum support and minimum confidence value.

From these experimental results, it can be easily seen that the proposed algorithm hides more rules in comparison to previous work for different user specified value of MST and MCT. In this algorithm, a rule $X \rightarrow Y$ is hidden by decreasing the support value of X U Y and increasing the support value of X. That can increase and decrease the support of the LHS and RHS item of the rule correspondingly. Whereas in the ISL algorithm a rule $X \rightarrow Y$ is hidden by increase the support value of X, the LHS of the rule but not support count X U Y. In DSR algorithm a rule $X \rightarrow Y$ is hidden by decreasing the support count of the item set X U Y in the transactions contain both X and Y, if the support value of Y has decreased. Also, the condition used by ISL algorithm allows only a small number of transactions to be modified for the rule under hidden. Therefore, the proposed algorithm hides more number of rules in comparison to previous work.

The second characteristic has been observed the database effects. Table 4.5 shows the relationship between total number entries modified and number of transaction, different values (10, 20, and 30) of minimum support. The proposed algorithm modified a few numbers of entries for hiding a given set of rules in all the datasets.

The last characteristic has observed is the CPU time requirement. Table 4.5 shows the relationship between total CPU time for number of entries modified and number of transaction, different values (10, 20, and 30) of minimum support. The proposed algorithm modified a few CPU time for hiding rule and modified entries are given set of rules in all the datasets.

VI.CONCLUSION

The purpose of proposed WSDA association rule hiding algorithm for PPDM is to hide certain crucial information so they cannot discovered through association rule. In this chapter, an efficient association rule hiding

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 1, Issue 7, September 2012

algorithm for PPDM has been proposed. This is based on association rule hiding approaches namely ISL and DSR and modifying the database transactions so that the confidence of the association rule can be reduced. The proposed algorithm hides the generated crucial association rule on the both side (LHS and RHS) correspondingly, so it has reduced the number of modification and hides more rules in less time. The performance of WSDA algorithm is compared with ISL and DSR for 1000, 2000, 3000 transactions database. The proposed method has increased 19% of hidden rules compared to ISL and DSR approach.

REFERENCES

1. Abedelaziz Mohaisen and Dowon Hong (2008), 'Privacy Preserving Association Rule Mining Revisited', Journal of the Computing Research Repository, pp. 1-16.
2. Agrawal, R., and Srikant (2007), 'Privacy Preserving Data Mining', Proceedings of the 19th ACM International Conference on Knowledge Discovery and Data Mining, Canada, pp. 439-450.
3. Agrawal R. and R. Strikant (2004), 'Fast algorithms for mining association rules', Proceedings of the International Conference on Very Large Data Bases, pp. 487-499.
4. Bertino E., Nai Fovino and Parasiliti Provenza (2005), 'A Framework for Evaluating Privacy Preserving Data Mining Algorithms', Journal of Data Mining and Knowledge Discovery, pp. 78-87.
5. Bikramjit Saikia and Debkumar Bhowmik (2009), 'Study of Association Rule Mining and different hiding Techniques', PhD thesis, Department of computer Science Engineering, National Institute of Technology, pp.55-63.
6. Bo Peng and Xingyu Geng (2010), 'Combined Data Distortion Strategies for Privacy-Preserving Data Mining', Proceedings of the 3rd International Conference on Advanced Computer Theory and Engineering, pp. 241-253.
7. Charu C. Aggarwal and Philip S. Yu (2008), 'A Survey of Association Rule Hiding Methods for Privacy', A hand book of Privacy-preserving data mining: models and algorithms, Kluwer Academic Publishers, London, pp.32-39.
8. Chen M.S and Yu P.S. (2004), 'Data Mining: An Overview from a Database Perspective', IEEE Transactions on Knowledge and Data Engineering, Vol. 8, 866-883.
9. Chih Chia Weng, Shan-Tai Chen and Hung-Che Lo (2009), 'A Novel Algorithm for Completely Hiding Sensitive Association Rules', Proceedings of the 8th International Conference on Intelligent Systems Design and Applications, pp.202-208.
10. Dasseni E., Verykios V. and Bertino E. (2004), 'Hiding Association Rules by using Confidence and Support', Proceedings of the 4th Information Hiding Workshop, pp. 369-383.
11. Du W. and Zhan Z (2010), 'Using randomized response techniques for privacy-preserving data mining', Proceedings of the 9th ACM International Conference on Knowledge Discovery and Data Mining, Washington, pp. 142-157.
12. Du W. and Atallah M. J. (2007), 'Secure MultiParty Computation Problems and Their Applications: A Review and Open Problems', Proceedings of the Workshop on New Security Paradigms, pp. 13-22.
13. Durgesh Kumar, Mishra Purnima Trivedi and Samiksha Shukla (2009), 'A Glance at Secure Multiparty Computation for Privacy Preserving Data Mining', International Journal on Computer Science and Engineering, Vol.1, pp.171-175.
14. Elisa Bertino, Dan Lin and Wei Jiang (2008), 'A Survey of Quantification of Privacy Preserving Data Mining Algorithms', Journal of Computing, Volume 34, pp. 183-205.
15. Elisa bertino and Igor fovino (2005), 'A Framework for Evaluating Privacy Preserving Data Mining Algorithms', International Journal of Data Mining and Knowledge Discovery, pp.121-154.
16. Evfimevski A., Gehrke J., and Srikant R.(2007), 'Limiting Privacy Breaches in Privacy Preserving Data Mining', Proceedings of the International Conference on Databases, pp. 171-182.
17. Evfimievski A., Srikant R., Agrawal R. and Gehrke J. (2007), 'Privacy Preserving Mining of Association Rules', Proceedings of the 8th ACM International Conference on Knowledge, Discovery and Data Mining, pp.217-228.
18. Evfimievski M. (2008), 'Randomization in Privacy Preserving Data Mining', Proceedings of the 9th ACM International Conference on Knowledge Discovery and Data Mining, pp.43-48.
19. Felty A. P. and Matwin S. (2007), 'Privacy-Oriented Data Mining by Proof Checking', Proceedings of the 6th European Conference on Principles of Data Mining and Knowledge Discovery, pp.138-149.
20. Guo Y. (2007), 'Reconstruction-Based Association Rule Hiding', Proceedings of the Workshop on Innovative Database Research, pp.511-543.
21. Joseph Domingo (2008), 'A survey of inference control methods For privacy-preserving datamining', International Journal of datamining, pp.53-80.
22. Juggapong Natwichai, Xingzhi Sun and Xue (2008), 'A Heuristic Data Reduction Approach for Associative Classification Rule Hiding', Proceedings of the 10th Pacific Rim international conference on artificial intelligence, pp. 140-151.