ISSN 2320 - 2602

Volume 2, No.6, June 2013

International Journal of Advances in Computer Science and Technology

Available Online at http://warse.org/pdfs/2013/ijacst03262013.pdf

EFFICIENT DECISION TREE BASED PRIVACY PRESERVING APPROACH FOR UNREALIZED DATA SETS

Ms.S.Nithya[1,] Mrs. P.Senthil Vadivu[2]

[1]Research Scholar,Department of Computer Science, Hindusthan College of Arts and Science, Coimbatore, India.nithya.subi844@gmail.com

[2] Head, Department of Computer Application,Hindusthan College of Arts and Science, Coimbatore, India. sowju_sashi@rediffmail.com

ABSTRACT

Privacy preserving data mining (PPDM) is major issue in the areas of datamining and security. PPDM datamining algorithms are analysis the results of impact data privacy. Objective of PPDM is to preserve confidential information by applying datamining tasks without modifying the original data. Because privacy preservation is important for datamining and machine learning applications, it measures to measures designed to protect private information. Numerous datamining techniques are analysis the result of preserved data. But still it becomes loss of information and reduces the utility of training samples. In this research we introduce a decision tree based privacy preserving approach. In this approach the original dataset or data samples are converted into the unrealized dataset where the original data samples cannot be reconstructed if an unauthorized party were to steal some portion. It covers the application of privacy preserving approach with the ID3 decision tree learning algorithm. The problem in existing decision tree based privacy preservation approach is inadequate storage space and it can be implemented only for discrete-valued attributes. Improve the preservation accuracy of the system by building RIPPER learns a rule for a given class; here the examples of that class are denoted as positive instances, examples from the left behind classes are denoted as negative instances. RIPPER algorithm to maximize the information gain and increase the number of rules to covers the non negative rates as well as approach is compatible with other privacy preserving approaches, such as cryptography, for additional protection. In this research we analysis the performance of preserved confidential information with ID3 and RIPPER based approach. A proposed RIPPER result shows the better performance than the existing system.

Keywords : PPDM, decision tree algorithm, data mining, machine learning.

1. INTRODUCTION

Data mining is the procedure of extracting hidden information from large data sets. This process is achieved by combining methods from statistics and artificial intelligence (AI) with DBMS.Datamining (DM) is used in the investigation of activities by using mining techniques. It is widely used by researchers for business and science applications. Because the Data composed from individuals are key essential for making decision or recognition based applications.

Conventional data mining methods operate on the data warehouse representation model of collecting all information into the centralized database and then apply mining algorithm to that data warehouse. This representation of data warehouse works fine at the time the entire information is created by a single keeper who generate and uses a data mining representation without changing the results to any third party. Major problems in the privacy preservation of information from centralized data warehouse or cloud environment .The primary problem may be the reality of certain attributes of the data or a combination of attributes might be leak personal exclusive information. The final problem might be that the information or data is horizontally divided across multiple keeper nobody of which is allowed to transfer information to the other location. The information might be vertically partitioned in which case, different keeper own dissimilar attributes of the information and they have the same sharing restrictions. Finally they use the data mining model with some restrictions, here the rules generated by system might be restricted and a few rules might go ahead to person profiling in way which are forbidden by law.

Privacy offers emancipation from illegal entrance. The long term goal of the government statistical agencies and database security research community is how to secure the sensitive data against unconstitutional access. Privacy protection has become one of the major issues in data mining research. An essential constraint of privacy-preserving data mining is to safeguard the input data, yet still permit data miners to dig out the valuable knowledge models. Many numbers of privacy-preserving data mining techniques have newly been projected which take either a cryptographic or a statistical approach. Secure multi-party computation is used in the cryptographic approach which ensures strong privacy and accuracy. But, this approach typically suffers from its poor performance. The statistical approach has been used to extract the facts from association rules, clustering and decision trees. This approach is very popular because of its high performance. Privacy has become an important issue in Data Mining. Several methods have been bringing out to solve these issues. In order to protect the privacy information, the objective of privacy preserving data mining is to hide certain sensitive information so that they cannot be discovered through data mining techniques. The authors deal with the problem of association rule mining which preserve the confidentiality of each database. In order to avoided the privacy information broadcasted or been illegal used.

Privacy preserving data mining (PPDM) is a novel study area that examines the troubles which occurs after applying the data mining techniques. Privacy problems related to the application of data mining techniques are divided into two broad kinds, data hiding and knowledge hiding. Data hiding is the exclusion of confidential or sensitive information from the data before it is disclosed to others. Knowledge hiding is the results of data mining methods after the data analysis, these may find out the hidden knowledge. Such knowledge should be protected from others.

Privacy preserving data mining (PPDM) has emerged to address this issue. Most of the techniques for PPDM uses modified version of standard data mining algorithms, somewhere the modification are made using well known cryptographic techniques ensure the required privacy for the application for which the technique was designed. In most cases, the constraints for PPDM are preserving accuracy of the data and the generated models and the performance of the mining process while maintaining the privacy constraints. The numerous procedures used by PPDM can be summarized as below:

1. The data is changed before delivering it to the data miner.
2. The data is circulated between two or more locations which work together using a semi-honest protocol to learn global data mining results without revealing any information about the data at their individual sites.
3. While using a representation to classify data, the classification outcomes are only exposed to the selected party, who does not learn something else, further the classification results, but can check for existence of certain rules without revealing the rules.

Inorder to overcome the problems or issues of the existing privacy preservation approach propose a new perturbation and randomization based approach that protects centralized sample data sets utilized for decision tree data mining. Privacy protection is applied to sanitize the samples prior to their release to third parties in order to mitigate the threat of their inadvertent disclosure or theft. In contrast to other refinement methods, our privacy preservation approach doesn’t degradation of the accuracy of results. The decision tree using modified ID3 and proposed RIPPER algorithm can be built from the unrealized data sets, such that the originals dataset need not to be reconstructed. Likewise, this approach can be applied at any time during the data collection process so that privacy protection can be in effect even while samples are still being collected.

The major contribution of this paper are follows : first, as is the norm in data collection procedure, a sufficiently huge number of sample data sets have been collected to achieve significant data mining results covering the research objective Second the numeral amount of data sets leaked to potential attackers constitutes a small portion of the entire sample database. Third, identity attributes (e.g., social insurance number) are not considered for the data mining process because such attributes are not meaningful for decision making. Fourth the collected all data are discretized; continuous values can be represented via ranged value attributes for decision tree data mining using ID3,modified ID3 and RIPPER algorithm .The rest of this paper is defined in the following manner: the next section describes privacy preserving approaches that safeguard samples in storage. Section 3 introduces a privacy preservation approach via data set complementation. Decision tree based algorithm for unrealized dataset. Section 4 provides the experimental results .Section 5 concludes the results and further research of the system.

2. RELATED WORK

2.1 Privacy Preserving Data Mining

The randomization method is a technique [1] for privacy-preserving data mining in which noise is added to the data in order to mask the attribute values of records. The noise added is sufficiently large so that individual record values cannot be recovered. Consequently the techniques are designed to derive aggregate distributions from the perturbed records. Similarly the data mining techniques can be developed in order to work with these aggregate distributions. They will provide a broad overview of the different techniques for privacy-preserving data mining. They will provide a review of the major algorithms available for each method, and the variations on the different techniques. They will also discuss a number of combinations of different concepts such as k-anonymous mining over vertically- or horizontally-partitioned data. They will also discuss a number of unique challenges associated with privacy-preserving data mining in the high dimensional case.

In Privacy Preserving Data Mining: Models and Algorithms [4], Aggarwal and Yu classify privacy preserving data mining techniques together with data alteration and cryptographic approaches, query auditing and perturbation-based strategies. Query auditing and most cryptographic techniques are subjects beyond the focus of this research [4].

Secure Multi-Party Protocols

Privacy-preserving data mining solution [3] have the belongings that the only information learned by the dissimilar hospitals is the yield of the data mining algorithm. This difficulty, whereby dissimilar organizations cannot straightly share or pool their databases, yet must nevertheless carry out joint research via data mining, is quite common. For example, judge the interaction between dissimilar intelligence agencies. For security purpose this agency cannot allow each one other free access to their confidential information. It is much more possible that disbelieving behavior would be detected if the different agencies were able to run data mining algorithms on their combined data.

2.2 Trusted Third-Party service and Hiding Sensitive Predictive data

Sharing private data in a computation presents a paradox. How can two parties combine their private data in a computation without revealing their data to one another? Many systems, such as online auctions, solve this problem by introducing a trusted third party to run the computation. However, such systems are usually specialized to a single application and provide only vague guarantees on the system's ability to control information leaks. This paper presents the Trusted Execution Platform (TEP) [2],a new system that supports general-purpose multiparty computation with specific guarantees on information leaks. TEP satisfies its identification and isolation requirements. The computation initiates connections to the participants specified in the parameters. Participants run trusted local agents that accept connections from computations on TEP and handle the necessary authentication and policy decisions on their participant's behalf .An alternate model is to allow participants to initiate connections to computations on TEP. This has the convenience of following the standard client-server model and allowing participants to join long-running computations dynamically.

Association rule mining [ARM] [3] is an important data-mining method that finds interesting association amongst a large set of data items. Because it might disclose patterns and different kinds of perceptive knowledge that are difficult to find or else, it might pose a threat to the privacy of discovered confidential information. Such information is to be protected against unauthorized access. Much strategy had been proposed to hide the sensitive information. Some use distributed databases over numerous locations, data perturbation, data distortion techniques and clustering .The proposed approach uses the data distortion technique where the position of the sensitive items is altered but its support is never altered. The size of the database remains the similar. It uses the design of representative rules to prune the rules first and then hides the sensitive rules.

The Association rule mining [ARM] approach uses the data distortion technique where the position of the sensitive items is altered but its support is never changed. The dimension of the database remains the similar. It uses the idea of representative rules to prune the rules first and then hides the sensitive rules. Advantage of the Association rule mining [ARM] [3] approach is that it hides maximum number of rules still; the existing approach fail to hide all the preferred rules, which are supposed to be hidden in minimum number of passes. Strategies and a suit of algorithms for privacy preserving and hiding knowledge from data by minimal perturbing values. The proposed approach uses the data distortion technique where the position of the sensitive item(s) is altered but its support is never changed however the size of the database remains the same. The proposed heuristics use the idea of representative rules to prune the rules first and then hides the sensitive rules.

2.3 Anonymization based methods

Data modification techniques maintain privacy by modifying attribute values of the sample data sets. Basically the data sets are modified by eliminating the uncommon elements among all data sets. These related data sets act as mask for the others within the group because they cannot be distinguished from the others; every data set is loosely linked with a certain number of information providers. K-anonymity [5] is a data modification approach that aims to protect private information of the samples by generalize the attributes. K-anonymity tradeoffs privacy for utility.The protection of sensitive information when samples are given to third parties for processing or computing [1], [2], [3], [4], [5]. It is in the interest of research to disseminate samples to a wide audience of researchers, without making strong assumption about their dependability.

2.4 Secure multiparty computation (SMC)

SMC approaches employ cryptographic tools for collaborative data mining computation by multiple parties. Samples are distributed amongst diverse parties and they obtain part in the information computation and communication process. SMC investigation focus on protocol development [11] for protecting privacy among the involved parties [12] or computation efficiency [13]; However samples from centralized processing and storage space privacy is out of the scope of SMC.

3.UNREALIZED TRAINING SET AND DECISION TREE LEARNING ALGORITHM

We can build different decision trees from the same training set by using the procedure described in the previous section, because of the undetermined selection criteria of the test attribute in the recursive case. The efficiency of a test element or attribute can be determined by its classification of the training set. A perfect attribute splits the outcomes as an exact classification, which achieves the goal of decision-tree learning. Diverse criteria are used to select the “best” attributes, e.g. Gini impurity. Among these criteria, information gain is commonly used for measuring distribution of random events. Iterative Dichotomiser 3 (ID3) selects the test attribute based on the information gain provided by the test outcome. Information gain measures the change of uncertainty level after a classification from an attribute. Fundamentally, this measurement is rooted in information theory.

Input: set of training samples (T_S): R_1, R_2, …,R_m and set of Attributes a_1, a_2, …, a_m
Default: default value for the target predicate
Output: decision tree
Procedure build-tree(T_S, attributes, default)

```
1. if $T_S$ is empty then return default
2. default← Majority _ Value $T_S$
3. if $H_{a_i}(T_S)$then return default
4. else if attributes is empty then return default
5. else
6. best← Choose-Attribute (attribute, $T_S$ )
7. tree← a new decision tree with root attribute best
8. for each value $v_i$ of best do
9. $T_{Si}$←datasets inTS as best= $k_i$
10. subtree←Generate-Tree( attribute-best, $T_S$ , default)
11. connect tree and subtree with a branch labelled$k_i$
12. return tree
```

To unrealized the samples, we initialize both set of input sample dataset and perturbing datasetas empty sets, i.e. Unrealized training set is called. Consistent with the procedure described above, universal dataset is added as a parameter of the function because reusing pre-computed universal datasetis more efficient than recalculating universal dataset. The recursive function unrealized training-set takes one dataset in input sample dataset in a recursion without any special requirement; it then updates perturbing datasetand set of output training data sets correspondent with the next recursion. Therefore, it is obvious that the unrealized training set process can be executed at any point during the sample collection process.

Input: Unrealized training dataset
Output: Modified decision tree

```
If unrealized dataset is empty then return default
Default ← minority –Value
Else
Tree ← best highest value of information gain (Root)
Subtree← tree(root,best size)
Connect tree and subtree
Return tree
End
```

Similar to the traditional ID3 algorithm Choose Attribute selects the test attribute using the ID3 criterion based on the information entropies, i.e., select the attribute with the greatest information gain.Algorithm Minority-Value retrieves the least frequent value of the decision attribute, which performs the same function as algorithm Majority-Value of the tradition ID3 approach that is, getting the majority frequent value of the decision attribute of T_S. the decision attribute should be arbitrarily chosen and generate the decision tree by calling the function Generate-Tree.

Attributes: set of attributes
Default: default value for the target predicate
Output: tree, a decision tree

```
1. if $(T', T^p)$ is empty then return default
2. default←Minority _ Value$(T', T^p)$
3. if then return default
4. else if attributes is empty then return default
5. else
6. best← $choose - attribute'$(attributes,size,$(T', T^p)$)
7. tree← a new decision tree with root attribute best
8. size← size/number of possible values $k_i$ in best
9. for each value $v_i$ of best do
10.$T'_i$ ←dataset in $T'$as best =$k_i$
11.$T'_p$ ←dataset in $T^p$as best =$k_i$
12.subtree← Generate-tree (size,$T'$, $T^p$,attribute-best, default)
13. connect tree and subtree with a branch labelled$k_i$
14. return tree
```

RIPPER algorithm to maximize the information gain and increase the number of rules to covers the non negative rates. The RIPPER algorithm performs better than the ID3 algorithm. RIPPER algorithm performs two steps, adding the original rule R and after adding the condition R‘or candidate rule measure the information gain (R, R') at true positives. Then it performs until the coverage of negative positive and negative true samples in the data. Learning process, the training data is sorted by class labels in ascending order according to the corresponding class frequencies. Rules are then learned for the first m-1 classes, starting with the smallest one. Once a rule has been created, the instances covered by that rule are removed from the training data, and this is repeated until no instances from the target class are left. The algorithm then proceeds with the next class. Finally, when RIPPER finds no additional rules to discover, a default rule (with empty antecedent) is added for the last (and hence most frequent) class.

In Ripper, conditions are added to the rule to maximize an information gain measure

$$Gain(R', R) = s.(log_2 \frac{N'_+}{N} - log_2 \frac{N_+}{N})$$

Where

R : The original rule

R' : The candidate rule after adding a condition

N (N’): the number of instances that are covered by R (R')

N_+ (N'_+): the number of true positives in R (R')

s: the number of true positives in R and R' (after adding the condition)

Conditions are added to the rule until it covers no negative example.

p and n : the number of true and false positives respectively.

$$rvm(R) = \frac{p-n}{p+n} \approx 1$$

Outer loop adds one rule at a time to the rule base and Inner loop adds one condition at a time to the current rule. The information gain measure is maximized by adding the conditions to the rule.

Pseudo code for RIPPER algorithm

```
Ripper(Pos, Neg, k)
Rule Set ← LearnRuleSet(Pos, Neg)
For k times
RuleSet ← OptimizeRuleSet (RuleSet, Pos, Neg)
LearnRuleSet(Pos, Neg)
RuleSet ← ∅
DL ← DescLen(RuleSet, Pos, Neg)
Repeat
Rule ← LearnRule(Pos, Neg)
Add Rule to RuleSet
$DL'$ ← DescLen(RuleSet, Pos, Neg)
If $DL'$` >$DL$ + 64
PruneRuleSet(RuleSet, Pos, Neg)
Return RuleSet
If DL1 <DL ,   DL ←$DL'$
Delete instances covered from Pos and Neg
Until Pos = ∅
Return RuleSet
```

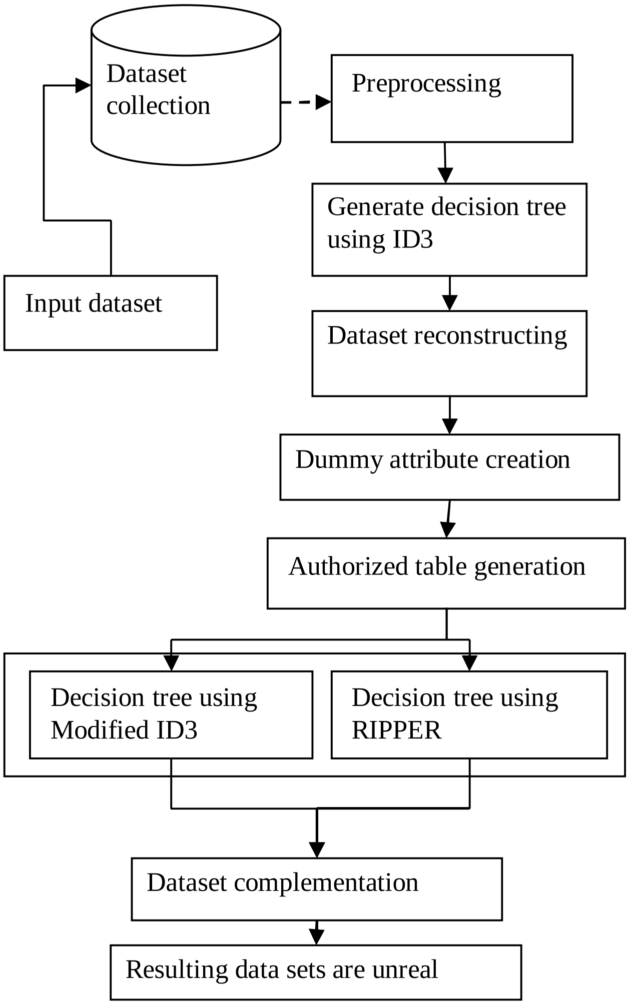
Dataset collection
Preprocessing
Generate decision tree using ID3
Input dataset
Dataset reconstructing
Dummy attribute creation
Authorized table generation
Decision tree using Modified ID3
Decision tree using RIPPER
Dataset complementation
Resulting data sets are unreal

Figure 1: Architecture of the Privacy preserving unrealized dataset

4. EXPERIMENTAL RESULTS

In this section we measure the performance of the system in terms of the precision, recall, Fmeasure with Decision tree learning algorithms for both realized dataset and unrealized dataset .Measuring these parameters show the results of the accuracy in terms of how privacy preservation is achieved better than the existing methods.

4.1 Precision

Precision value is calculated is based on the retrieval of information at true positive (TP) prediction, false positive (FP).In privacy preservation data precision is calculated the percentage of preserved data results returned that are relevant. Precision =TP/ (TP+FP)

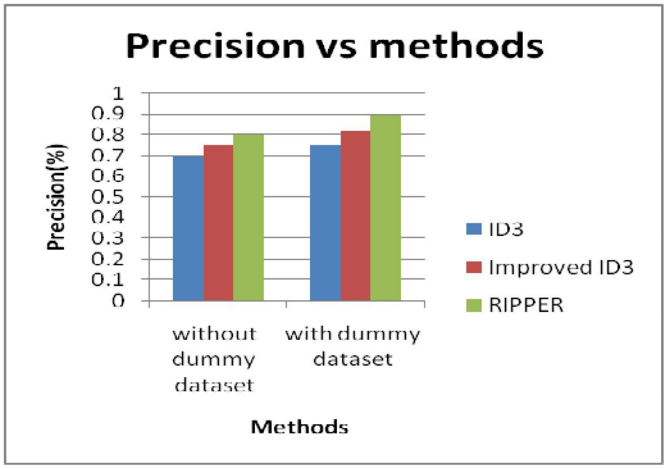

Figure 2: Precision Vs methods

In the above figure 1 measure precision results in two ways: After adding the Dummy dataset values and before adding dummy data to original dataset. Comparison of privacy preservation with three methods ,ID3 ,Improved ID3 ,RIPPER .Before adding the dummy dataset values to the making the decision tree are less results than the after adding the dummy dataset values to original dataset. So the results show that the X-axis defines the methods and the Y-axis measure precision accuracy in percentage. Precision value of the proposed privacy preservation with RIPPER algorithm show best results than the other decision tree learning methods.

4.2 Recall

Recall value is calculated is based on the retrieval of information at true positive (TP) prediction, false negative (FP). In privacy preservation approach the data precision is calculated with percentage of positive results returned that are recall in this context is also referred to as the True Positive Rate (TP). Recall is the fraction of relevant instances that are retrieved,Recall =TP/(TP+FN)

RIVAC

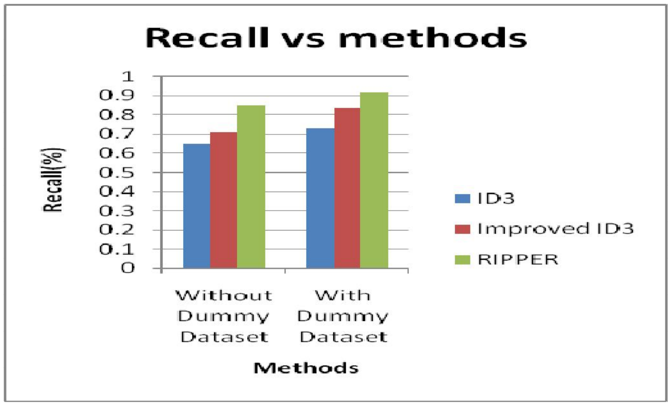

Figure 3: Recall Vs methods

In the above figure 3 measure recall results in two ways: After adding the Dummy dataset values and before adding dummy data to original dataset. Comparison of privacy preservation with three methods ,ID3 ,Improved ID3 ,RIPPER .Before adding the dummy dataset values to the making the decision tree are less results than the after adding the dummy dataset values to original dataset. So the results show that the X-axis defines the methods and the Y-axis measure precision accuracy in percentage. Precision value of the proposed privacy preservation with RIPPER algorithm show best results than the other decision tree learning methods.

4.3 Fmeasure

Fmeasure is a measure of a test's accuracy. It considers both the precisionp and the recallr of the test to compute the score: p is the number of correct results divided by the number of all returned results and r is the number of correct results divided by the number of results that should have been returned. The F Measure score can be interpreted as a weighted average of the precision and recall, where an F_1 score reaches its best value at 1 and worst score at 0.
Fmeasure = 2.Precision.recall /(precision + recall)

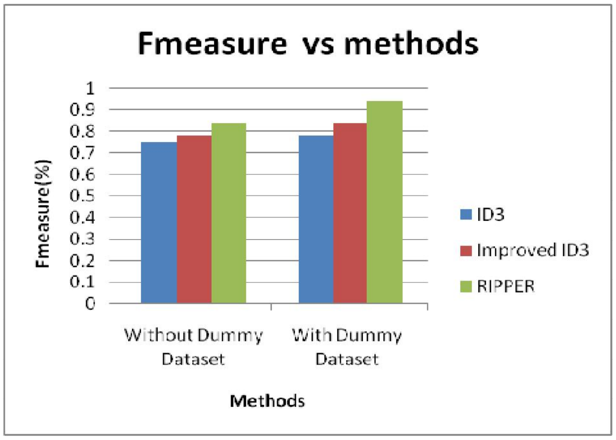

Figure 4: FmeasureVs methods

In the above figure 4 measure Fmeasure results in two ways: After adding the Dummy dataset values and before adding dummy data to original dataset. Comparison of privacy preservation with three methods ,ID3 ,Improved ID3 ,RIPPER .Before adding the dummy dataset values to the making the decision tree are less results than the after adding the dummy dataset values to original dataset. So the results show that the X-axis defines the methods and the Y-axis measure Fmeasure accuracy in percentage. Fmeasure value of the proposed privacy preservation with RIPPER algorithm show best results than the other decision tree learning .

5.CONCLUSION AND FUTURE WORK

Privacy preserving approach (PPDM) through data set complementation; it ensures the utility of training data sets for decision tree learning using ID3, Improved ID3. During the privacy preserving method, set of perturbed datasets is dynamically adapted. From the original data samples, these perturbed datasets are stored to permit a modified decision tree based datamining method. This method guarantees to provide the same datamining outcomes as the originals, which is proved mathematically and by a test using one set of sample datasets. From the viewpoint of privacy preservation, the original datasets can only be reconstructed in their entirety if someone has all perturbed datasets, which is not supposed to be the case for an unauthorized party. RIPPER algorithms which are very suitable for decision tree learning after completion of the unrealized dataset. RIPPER algorithm improvements have been created a rule learner and finally the results become unrealized dataset.

In our proposed system Privacy preservation of data set complementation fails if all training data samples are leaked since the data set restoration algorithm is common. So Further investigation is necessary to conquer the above limitation.As it is very easy to apply a cryptographic privacy preserving approach such as the antimonotone cryptographic construction, along with data set complementation.

REFERENCES

1. R. Agrawal and R. Srikant, "**Privacy Preserving Data Mining**," Proc. *ACM SIGMOD Conf. Management of Data (SIGMOD '00),* pp. 439-450, May 2000.
2. S. Ajmani, R. Morris, and B. Liskov, "**A Trusted Third-Party Computation Service**," *Technical Report MIT-LCS-TR-847*, MIT, 2001.
3. S.L. Wang and A. Jafari, "**Hiding Sensitive Predictive Association Rules**," *Proc. IEEE Int'l Conf. Systems, Man and Cybernetics*, pp. 164- 169, 2005.
4. C. Aggarwal and P. Yu, **Privacy-Preserving Data Mining:, Models and Algorithms**. *Springer,* 2008.
5. Q. Ma and P. Deng, "**Secure Multi-Party Protocols for Privacy Preserving Data Mining,**" *Proc. Third Int'l Conf. Wireless Algorithms, Systems, and Applications (WASA '08),* pp. 526-537, 2008.

6. L. Sweeney, "**k-Anonymity: A Model for Protecting Privacy**," *Int'l J. Uncertainty, Fuzziness and Knowledge-based Systems,* vol. 10, pp. 557-570, May 2002.
7. J. Gitanjali, J. Indumathi, N.C. Iyengar, and N. Sriman, "**A Pristine Clean Cabalistic Foruity Strategize Based Approach for Incremental Data Stream Privacy Preserving Data Mining**," *Proc. IEEE Second Int'l Advance Computing Conf. (IACC),* pp. 410-415, 2010.
8. Y. Zhu, L. Huang, W. Yang, D. Li, Y. Luo, and F. Dong, "**Three New Approaches to Privacy-Preserving Add to Multiply Protocol and Its Application**," *Proc. Second Int'l Workshop Knowledge Discovery and Data Mining, (WKDD '09),* pp. 554-558, 2009.
9. J. Vaidya and C. Clifton, "**Privacy Preserving Association Rule Mining in Vertically Partitioned Data**," *Proc Eighth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD '02),* pp. 23- 26, July 2002.
10. M. Shaneck and Y. Kim, "**Efficient Cryptographic Primitives for Private Data Mining**," *Proc. 43rd Hawaii Int'l Conf. System Sciences (HICSS),* pp. 1-9, 2010.