ISSN 2278-3083

Volume 3, No.1, January - February 2014

International Journal of Science and Applied Information Technology

Available Online at http://warse.org/pdfs/2014/ijsait05312014.pdf

AN IMPROVED PRIVACY PRESERVING WITH RSA AND C5.0 DECISION TREE LEARNING FOR UNREALIZED DATASETS

P.SENTHIL VADIVU[1], S.NITHYA[2]

[1]HEAD, DEPARTMENT OF COMPUTER APPLICATIONS, HINDUSTHAN COLLEGEOF ARTSAND SCIENCE, COIMBATORE, INDIA.
sowju_sashi@rediffmail.com

[2]RESEARCHSCHOLAR,DEPARTMENT OF COMPUTERSCIENCE,HINDUSTHAN COLLEGEOF ARTSAND SCIENCE, COIMBATORE,INDIA.
nithya.subi844@gmail.com

ABSTRACT

PPDM methods have been observed in various areas to preserve privacy for each data .Earlier work of privacy preserving data categorized into two ways perturbation based splitting and classification of the those data. It conform effectiveness of exercise data sets for decision tree learning. This work covers the purpose of new privacy preserving move toward through the decision learning ID3 algorithm. A major issue of the work is insufficient storage space method and this ID3 simply be capable of implementing for discrete-valued attributes simply. Perturbation dataset methods becomes less privacy because of the training data samples are leakages in restoration procedure, to overcome these leakage problem proposed secure multiparty communication methods with the use of cryptographic methods like RSA encryption methods and to support continuous values attribute proposed an C5.0 decision tree based classification .As a proposed work RSA algorithm can be included in Dataset completion approach with an encryption-decryption. Encryption scheme, the original data is encrypted before performing Perturbation of training data. A certified party, conversely, is capable to decipher the training data if the data leakage problem occurs at the restoration procedure using a decryption with specified secret key. It fundamentally discovers the most excellent splitting attribute and the most excellent splitting position of the numeric continuous attributes. C5.0 algorithm to construct whichever a decision tree or a rule set. A C5.0 representation mechanism by splitting the example based on the field with the intention of provides the highest Information Gain (IG). In this research evaluate the performance of the decision learning algorithm for both discrete and continuous attributes result, preserves the privacy data result was enhanced than existing works.

Keywords: Privacy Preserving Data Mining (PPDM), Encryption and Decryption schema, Decision Tree Algorithm, C5.0 Decision Tree Algorithm **.**

1. INTRODUCTION

The major principle of data mining is to extract information from large database or source information from database .many of the earlier data mining methods extract specific information in exact manner, but in order to preserve user data it not work well. The majority of presented data mining algorithms are conceded away in the statement to every the data might be obtainable at a solitary middle site. Whereas two or additional parties, who don't comprise sufficient self-confidence in every additional have a widespread aspiration to mine information beginning every one of their confidential data, the privacy problems approach up.

To preserve the user data PPDM objective to making privacy and data mining coexist under the conditions. Outstanding to growing concerns correlated to confidentiality, a variety of PPDM techniques have been proposed to address dissimilar confidentiality problems [1]. These techniques typically work below a variety of postulation and make use of dissimilar methods.

In this paper, we determinations major focal point on the perturbation method, it is broadly second-hand in PPDM. In this research a novel technique to construct data mining representation straightforwardly from the perturbed data not including difficult to resolve the universal data distribution restoration as an intermediary step. Supplementary accurately, recommend a C5.0 decision tree classifier that be capable of compact by means of disturbed numeric continuous attributes. Our privacy preserving decision tree C5.0 classifier uses perturbed preparation data, and construct a decision tree representation, which might be second-hand to categorize the unique or perturbed data sets.

According to my study on existing literatures, numerous confidentiality security approaches protect confidential information of example datasets, other than not accuracy of data mining outcomes. Therefore, the usefulness of the sanitary datasets is reduced. Several approaches are appropriate transformation functions to disinfect the samples, and utilize opposite functions of individuals transformation functions to get better the original datasets. Conversely, safety issues of these function also increased, since off to recover original data samples of user private information.

It has two major contributions:Initially to prevent user or private information and improve helpfulness of data sample for decision tree C5.0 learning as well as ID3 learning algorithm .It converts the original data sample or training data samples into unrealized dataset first ,after that only proceed the data for other process. Then it is applied to data mining process to measure the unrealized dataset results. Our proposed schema is a C5.0 decision tree algorithm and agrees to noise preservative technique. Our approach is appropriate for the situation where several parties would like to carry out data mining, However every of them solitary have a little section of the information.

To obtain the comprehensive data mining sample, a variety of parties should distribute their data, however every party has its confidentiality and safety concern.Our move toward is the result for such a condition. Each and every data sample perturbed by adding additional data sample and then it send to classification or decision learning procedure, data mining method gathers perturbed data set sample from unrealized dataset sample for each user data or some party private information such as voter list hospital data, adult dataset samples etc. From this result data miner classify the unrealized dataset into categorization process using C5.0 algorithm ,before that more security issues occurs during transformation process so encryption algorithm is proposed on original data , after that only convert into unrealized data samples for each data party.In this case, every party solitary distinguish its individual information and the classifier, and it does not have some information concerning the information of others. This way of process reduces the privacy preservation issues that is improves the result of preserved privacy result information during the communication among the parties and estimation costs for every party are reduced.

2. BACKGROUND STUDY

In earlier work privacy preservation of data becomes two major issues: One of them is how to improve the customer private information preservation data and another one of them is how to reconstruct the original information in equivalent manner that is without any loss of information or data .In first issues overcome by perturbing data set values in [3].

This procedure random method is followed that is randomly noise data are added to original dataset sample to alter susceptible values, and the allocation of the indiscriminate data is second-hand to produce a novel data allocation not including illuminating the unique data standards. The resultant of these data samplers is used to redistribute the original data samples without making any changes of original private information or data and it is applied to data mining methods such as classification, association rule mining to regain the original private information.

Later modification of this move toward has tightened assessment of unique values based on the indistinct information [2]. The information deformation move toward has also been useful to Boolean values in study effort [1,4,5].Since a number of methods might obtain confidential information from the restoration step [6]. Dissimilar to the unique noise preservative technique in [3], numerous characteristic perturbation techniques have been anticipated. One of the major categories of this method is multiplicative perturbation technique.In the examination of arithmetic belongings of the information, multiplying the unique information values with an unsystematic noise matrix is to turn around the unique information matrix; consequently it is moreover called rotated based perturbation.

Rotation in variant Classifiers developed by author [7] to demonstrate several DM methods are directly applied to rotated based perturbation information. Improved privacy protection result was analyzed the same procedure was proposed by Liu et al [8] .Some of the interesting methods and techniques also proposed in earlier work such as matrix decomposition [9] and so on. Without proceeding of reconstruction step also privacy are preserved using data mining with perturbation based approaches [10].Decide the appropriate techniques are resolute by the technique which noises have been introduced.

To our knowledge, extremely a small number of works focal point on altering the original information to assemble the perturbation information requirements.The additional approaches employ cryptographic methods to create a data mining methods or techniques. For instance, the objective of the security is preserved by constructing ID3 decision tree classifier [11] where the training set is dispersed among two parties. Dissimilar solution was specified to deal with dissimilar data mining problems using cryptographic procedures (e.g., [12-14]).It treats PPDM as an extraordinary casing of protected multi-party calculation and not only objective to preserve individual user information also detects the leakage information data other than the concluding consequence. However at what time the numbers of parties turn into larger, the communiqué and calculation costs develop exponentially.

3. UNREALIZED TRAINING SET WITH RSA AND DECISION TREE LEARNING C5.0 ALGORITHM

In order to provide secure multiparty among the communication between the parties in public area network data specific cryptography methods requires with encryption and decryption procedure. The original private information among two parties is encrypted using the private key for each party. Then the user data is decrypted using the decryption key the encrypted data [15]. Encryption is an essential implement for the security of responsive information. The attitude to make use of encryption is confidentiality to preventing discovery of information in communications.Encryption is a move toward of discussion to an important person while the other public are listening; though such the additional people cannot recognize what you are say [15].Encryption algorithms participating a very important responsibility in provided that information safety alongside malevolent attacks.In mobile devices protection is very important and diverse types of methods are second-hand to avoid malicious attack on the broadcasted information.

It can be categorized into two ways both symmetric keys and asymmetric key [16]. In Symmetric keys based encryption schema only one key is used for both encryption and decryption procedure for private information.Whereas asymmetric Keys generation encryption schema there are two keys are generated i.e. is private and public keys. Public key is used to encrypt the original information and private key is used to decrypt the original information or private information of individuals.

Here the original private information such as adult dataset or voter list information data are first encrypted before performing the unrealized dataset conversion and then perform unrealized dataset conversion by adding the additional data into encrypted data ,then perform the decision tree learning for both discrete and continuous attributes dataset .To perform encryption for original information using RSA algorithm .

RSA stands for Rivest, Shamir and Adleman. RSA is a frequently accepted public key cryptography algorithm.The principal and quiet the majority frequently used method for RSA as asymmetric algorithm. It is used in several areas of software based security and may be used to exchange the information from private information among the key values are generated with variable size key or attribute based key size. The key-pair (p, q) is derived from a extremely huge numeral n, that is the end result of two prime numbers (p, q) are selected according to exacting rules.RSA has been expansively second-hand for begin protected communication channel and for validation the individuality of service provider over lacking confidence communication intermediate.RSA absorb a public key and a private key. The public key is able to be well-known to everyone and it is second-hand for encrypting unique information beginning the communiqué. Individual private information can be encrypted by means of the public key be able to merely be decrypted in a reasonable quantity of instance by means of the private key.

The keys for the RSA algorithm are generated the subsequent technique:

1. Choose two distinct prime numbers p and q for original private user information.
    - For protection principle, the integer's p and q be supposed to be selected at random, and be supposed to be of comparable bit-length. Prime integers can be professionally establish using a mainly test.

2. Calculate $n = pq$. n is used as the modulus for both the public and private keys. Its length, usually expressed in bits, is the key length.
3. Compute $\varphi(n) = \varphi(p)\varphi(q) = (p-1)(q-1)$, where φ Euler's totient function for original information
4. Choose an integer e such that $1 < e < \varphi(n)$and $gcd(e, \varphi(n)) = 1$; i.e. e and $\varphi(n)$ are coprime. $e$is released as the public key exponent.
5. Determine $d$as $d-1 \equiv e\ (mod\ \varphi(n))$, i.e., d is the multiplicative inverse of $e\ (modulo\ \varphi(n))$.
    - This is more clearly stated as solve for d given $de \equiv 1\ (mod\ \varphi(n))$
    - $d$is kept as the private key exponent.

The research work accessible at this time believes the C5.0 Algorithm for data mining. The improvement of C5.0, which shows the improved performance as compared to the other existing ID3 and Improved ID3 algorithms.C5.0 algorithm to construct moreover a decision tree or creation of a rule set for unrealized user dataset with encryption key. It splits the user unrealized dataset by calculation of maximum information gain ratio $InformationGain(S, V)$. Every one subsample distinct by the primary split is after that split once more, frequently based on a dissimilar field, and the procedure replicate in anticipation of the subsamples cannot be split several additional. In conclusion, lesser value of the data splits is evaluated again and again to achieve better splitting results else it is removed or eliminated for original data.

A decision tree is a basic explanation of the splits established by the algorithm. Every one terminal node in the tree denotes a specific set of training data samples with encrypted format and each case in the training data are in the right place to precisely individual terminal node in the tree. It creates a rule set to predict individual information with privacy protected. Rule set are created by using the attribute that are related to user records and, in a technique, correspond to a simplify description of the information established in the C5.0 decision tree. Since of the manner rule sets effort, they do not contain the similar belongings as decision trees.

The mainly significant dissimilarity is with the intention of by means of a rule set, additional than single rule might be relevant for every exacting record. If numerous rules be relevant, every regulation obtain a weighted values are related by rule, and the concluding calculation is determined by combination of different weight values result for all rules with individual records or information. If no other rule matches for prediction of the individual record or information then randomly select the rule which are highest gain value matches to records as selected, the generalized form of rule is the collection of many attribute values that belong to the same class is defined as generally in the following manner

$$\text{“}if\ A\ and\ B\ and\ C\ and\ \ldots then\ class\ X\text{”},$$

where rules for every target class either positive or negative class relevant attributes are grouped jointly.

Particular user specified data that relevant to any case A is classified if the similar B, C attributes are matched to every rule; else if no rule matched then considered default class for case A. Every case goes to single of a little number of equally restricted classes.Belongings of each case so as to might be applicable to its class are presented, even though a number of cases might contain unidentified values for a number of characteristic. C5.0 is able to arrangement with some numeral of attributes. Rule sets are usually easier to appreciate than trees because every rule illustrates a detailed background connected with a class. One more improvement of rule set C5.0 classifiers is to facilitate they are frequently supplementary precise predictors than decision trees.

Each ruleset consists of:

- Rule number defined by user that is quite random and serves simply to recognize the rule.
- Statistics that review the presentation of the rule. Correspondingly to a leaf node for decision tress, n is the numeral of training samples covered through the rule and m, if it becomes visible, demonstrate how numerous of them do not fit into the class expected by the rule. Performance result of each and every rule is evaluated by using the Laplace ratio $(n - m + 1)/(n + 2)$.
- One or supplementary conditions that should everyone be contented if the rule are to be appropriate.
- Each of the class is predicted by using rule ,the value corresponds to every class is either zero or one.

In numerous applications, rule sets are favored suitable to the information with the intention of they are simpler and easier to appreciate than decision trees. Decision tree learning of C5.0 method offers better improvement result than the ID3 methods like the parameters are compared to ID3 are Speed, Memory usage, Smaller rule sets, Accuracy , Weighting , Winnowing. In C5.0 decision tree is created using $GainRatio$. $GainRatio$is to determine integrating entropy. $Entropy$ $(E(S),)$ procedures how unordered the data set is proceeds . It is indicated by the subsequent equation while there are classes $C_1, \dots\dots C_N$in data set S where $P(s_c)$is the probability of class C happening in the data set S:

$$E(S) = -\sum_{c=1}^{N} P(s_c) * \log_2 P(s_c)$$

Information Gain is a assess of the development in the quantity of regulate.

$$Gain(S,V) = E(S) - \sum_{V \in values(V)} \frac{|S_v|}{|S|} E(S_v)$$

Gain has a preconception towards variables with numerous values that separation the data set into slighter ordered sets. In regulate to decrease this bias, the entropy of every variable in excess of its m variable ideals is considered as $SplitInfo$:

$$SplitInfo(S,V) = \sum_{i=1}^{m} -\frac{|s_i|}{|S|} * \log_2 \frac{|s_i|}{|S|}$$

$GainRatio$is calculated by dividing $Gain(S,V)$by $SplitInfo(S,V)$so that the bias error values of the m variables are reduced for both minimal and larger rule set.

$$GainRatio(S,V) = \frac{Gain(S,V)}{SplitInfo(S,V)}$$

A concluding decision tree is altered to a set of rules by changing the path interested in conjunctive rules and prunes them to get better classification accurateness.

PSEUDOCODE

```
$Input: unrealizeddataset$
$Output: perturbingset$
$readtheinputdataset$
$defineclassesC_1, C_2 \dots C_n$
$TheprobabilityofclassC$
$occurringinthedatasetS: p(s)$
$EntropycalculationE(s) \leftarrow p(s)$
$E(s) = -\sum_{c=1}^{N} P(s) * log_2 P(S_c)$
$InformationGain \leftarrow Gain(S,V)$
        $= E(S) - \sum_{V \in values(V)} \frac{|S_v|}{|S|} * E(S_v)$
$calculatetheentropyfor$
$eachvariableoveritsmvariable$
$valuesiscalculate$d as SplitInfo
$\mathrm{SplitInfo}(S,V) = \sum_{i=1}^{m} -\frac{|S_i|}{|S|} * \log_2 \frac{|S_i|}{|S|}$
$\mathrm{Gain\ ratio}(S,V) = \frac{\mathrm{Gain}(S,V)}{\mathrm{SplitInfo}(S,V)}$
return tree
```

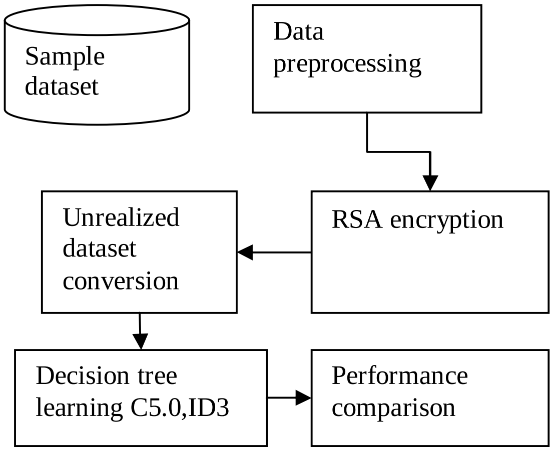

Figure 1: Architecture of the Privacy preserving unrealized dataset

4. EXPERIMENTAL RESULTS

In this section evaluate the performance of decision tree learning methods without cryptographic and with cryptographic methods as well as existing decision tree learning ID3 algorithm in terms of accuracy for realized dataset and unrealized dataset. Measuring these parameters show the results of the accurateness in terms of how privacy protection is achievedimproved than the presented methods.

The efficiency concerns the time required to find preserved data , the effectiveness is related to the quality of the privacy data for unrealized dataset for both ID3 and C5.0 learning algorithm and experimentally evaluated. The proposed method has been estimated using the following measures:

$$Accuracy = (TP + TN) / (TP + TN + FP + FN)$$

Where

TP = Number of true postive cases for learning
FP = number of false postive cases for learning
TN = Number of true negative cases for learning
FN = Number of false negative cases for learning

The following graph 1,Graph 2 shows the corresponding results in terms of accuracy with and without cryptography for ID3 and C5.0 learning algorithms .The accuracy values are tabulated in Table 1 and results are shown in Figure 1 ,Figure 2

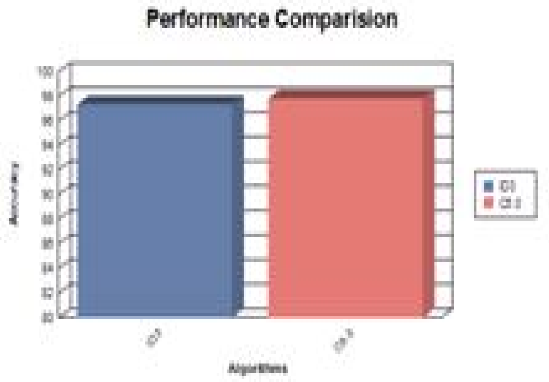

Figure 1: Performance comparison of decision tree without cryptography

In the figure1 shows the performance comparison of decision tree without cryptography to original dataset. It compares privacypreservation results with two methods, ID3 and C5.0 learning algorithm. Consequently the resultsshow that the X-axis defines the decision tree learning algorithms and the Y-axismeasure accuracy in terms of percentage. It shows that C5.0 has improved privacy accuracy than the ID3 algorithm

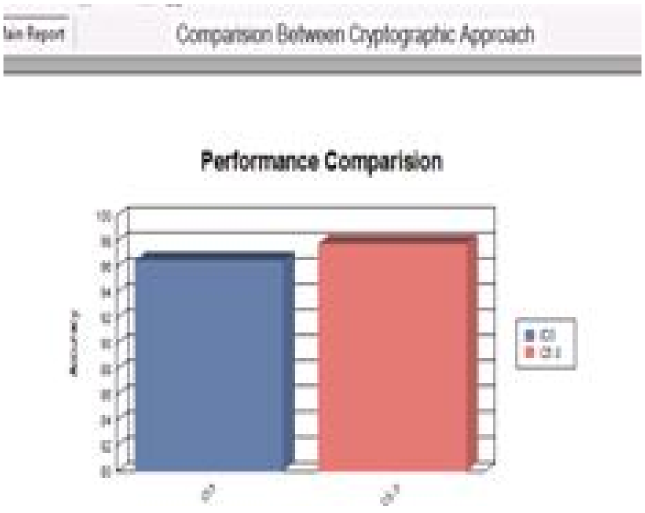

Figure 2: Performance comparison of decision tree withcryptography

In the figure2 shows the performance comparison of decision tree with cryptography to original dataset. It compares privacypreservation results with two methods, ID3 and C5.0 learning algorithm. Consequently the resultsshow that the X-axis defines the decision tree learning algorithms and the Y-axismeasure accuracy in terms of percentage with cryptography. It shows that C5.0 has improved privacy accuracy than the ID3 algorithm without /with privacy methods.

Table 1 : Performance comparison of decision tree with and without cryptography

Methods /Accuracy	**ID3**	**C5.0**
Without Cryptography	96.37	97.30
With Cryptography	97.5	97.85

5. CONCLUSION AND FUTURE WORK

In this work we presented a privacy preservation approach with decision tree learning for continuous and discrete attributes C5.0 and cryptographic methods via dataset complementation, which removeeveryexamplebeginning a perturbing set for each datasets. During the preservation of privacy for unrealized dataset perturbed datasets is parallely altered or changed and also stored in database .Proposed C5.0 learning algorithm assurance to present the similar data resulting as the originals, which is demonstratedprecisely and by anexaminationof sample dataset from original dataset. From the point of view of confidentialityprotection, the original datasets preservesimplyprivacy by reconstruction of original samples by adding noise data samples during unrealized dataset conversion.In future in regulate to get better the privacy preservation and the mining effectiveness, an effectual privacy preserving distributed mining techniques with association rules can be planned and improved version of decision tree learning algorithm also modified to improve performance as well as role based user provide privacy also another important direction to improve privacy level in terms of utility and accuracy.

REFERENCES

1. B. M. Thuraisingham,"**Privacy constraint processing in a privacy-enhanced database management system**", *Data Knowl. Eng.*, 55(2):159–188, 2005.
2. D. Agrawal and C. C. Aggarwal," **On the design and quantification of privacy preserving data mining algorithms**", *In PODS. ACM*, 2001.
3. R. Agrawal and R. Srikant," **Privacy-preserving data mining**", *In SIGMOD Conference*, pages 439–450, 2000
4. W. Du and Z. Zhan**," Using randomized response techniques for privacy-preserving data mining",***In KDD*, pages 505– 510, 2003.
5. A. V. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke ," **Privacy preserving mining of association rules**" *In Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 217–228, 2002.
6. H. Kargupta, S. Datta, Q. Wang, and K. Sivakumar**," On the privacy preserving properties of random data perturbation techniques**", *In ICDM*, pages 99–106. IEEE Computer Society, 2003.
7. K. Chen and L. Liu, **" Privacy preserving data classification with rotation perturbation",***In ICDM*, pages 589–592, 2005.
8. K. Liu, H. Kargupta, and J. Ryan**," Random Projection-Based Multiplicative Data Perturbation for Privacy Preserving Distributed Data Mining"**, *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 18(1):92–106, January 2006.
9. S. Xu, J. Zhang, D. Han, and J. Wang **,"Singular value decomposition based data distortion strategy for privacy protection",***Knowl. Inf. Syst.*, 10(3):383–397, 2006.
10. L. Liu, M. Kantarcioglu, and B. Thuraisingham**," The applicability of the perturbation based privacy preserving data mining for real-world data",***Data and Knowledge Engineering Journal*, 2007.
11. Y. Lindell and B. Pinkas**," Privacy preserving data mining",***In M. Bellare, editor, CRYPTO, volume 1880 of Lecture Notes in Computer Science*, pages 36–54. Springer, 2000.
12. C. Clifton, M. Kantarcioglu, J. Vaidya, X. Lin, and M. Y. Zhu**," Tools for privacy preserving data mining",***SIGKDD Explorations*, 4(2):28–34, 2002.
13. M. Kantarcioglu and C. Clifton**," Privately computing a distributed k-nn classifier",** In J.-F. Boulicaut, F. Esposito, F. Giannotti, and D. Pedreschi, editors, PKDD, volume 3202 of Lecture Notes in Computer Science, pages 279–290. Springer, 2004.
14. J. Vaidya and C. Clifton**," Privacy-preserving -means clustering over vertically partitioned data",***In KDD*, pages 206–215, 2003.
15. NeetuSettia ," **Crypt analysis of modern cryptography Algorithms**", .*International Journal Of Computer Science and Technology*. December 2010.
16. MohlyMohamadHadhoud ,"**Evaluation the problem of Symmetric Encryption algorithms**", *International journal of network security* vol.10 May 2010.
17. P. SenthilVadivu, S. Nithya, "**An Improved Privacy Preserving With RSA and C5.0 Decision Tree Learning For Unrealized Datasets".**