International Journal of Innovation and Scientific Research
ISSN 2351-8014 Vol. 24 No. 1 Jun. 2016, pp. 13-19

http://www.ijisr.issr-journals.org/

Improving Data Precision and Accuracy using Efficient Techniques

P. Suganthi[1], K. Kala[2], and C. Balasubramanian[3]

[1]PG Scholar, Department of CSE, P.S.R. Rengasamy College of Engineering for Women, Sivakasi, Tamilnadu, India

[2]Asst Professor, Department of CSE, P.S.R. Rengasamy College of Engineering for Women, Sivakasi, Tamilnadu, India

[3]Prof and head of CSE, P.S.R. Rengasamy College of Engineering for Women, Sivakasi, Tamilnadu, India

ABSTRACT: The need for privacy preserving is sharing of sensitive information occurs in many different ways. In order to maintain privacy in database, the confidential data should be protected in the form of modifying the sensitive data items. Protecting sensitive data is an important issue in the government, public database. It used protected the sensitive numerical data item in the form of modifying the original data item using the proposed techniques. There are various anonymization technique provides privacy protection which can be used such as data encryption, Randomization and k-anonymity. The existing system uses commutative encryption scheme to improve data privacy of the database and provides security of data by using AES algorithm. To enhancing other cryptography techniques (such as RSA) are used in the user database for secure their database from unauthorized user. The multiple users have to access the data from a database with the permission of administrator.

KEYWORDS: Privacy preserving, Anonymization, Classification, RSA algorithm.

1 INTRODUCTION

Data mining examine deals with the extraction of useful information from large collections of data with a variety of application. The individual information is collected by administrator. The mined information can be the form of patterns, rules, clusters or classification models. During the whole process of data mining, which contain sensitive information such as personal details, medical information, often get exposed to several parties including data owners, users and miners. Privacy preserving data mining deals with protecting the privacy of individual data. In cryptography system, multiple parties may wish to share aggregate private data, without leaking any sensitive information. The field of privacy preserving has seen fast advances in recent years because of the increases in the ability to store data. In particular, advances in the data mining have led to increased concern about privacy. People have become well aware of the privacy on their personal data and are very unwilling to share their sensitive information. In current years, the area of privacy deals with the ability to access personal data. The aim of privacy preserving data mining algorithms is to gather correct information from huge amounts of data while protecting at the same time considerate information.

Privacy preserving handles the data in modified form such as sensitive data like customer's name, addresses and the like should be modified from the original database. The recipient of the data not to be able to compromise other person's data. The main concern of privacy preserving algorithm is to preserve the privacy of person's sensitive data.

2 RELATED WORK

Most techniques for privacy preserving use some form of modified format of data. Normally, such methods diminish the granularity of exhibition in order to reduce the privacy. Many research on classification exploited decision tree and ensemble methods. The goal of classification algorithm was to achieve higher accuracy.

Corresponding Author: P. Suganthi

Panda et.al [2] compared the performance of Naïve Bayesian, Id3 algorithms for network intrusion detection. According to the authors NB performs better than Id3 with respect to overall classification accuracy. However, authors add that Decision Tree are robust in detecting new intrusion/attacks, in comparison to NB.

Ya-Qin [4] describes the ensemble bagging with decision tree(C5) to predict the living expectancy of breast cancer patients.

Asha and Natarajan et.al [5] compared the use of three different ensemble methods, including Bagging, id3 , and Random Forest. The experiment showed that Random Forest was the weakest method while Bagging was the strongest method in classification of tuberculosis (TB).

Abdul. Elminaam et al. [5] compare the evaluation of six of the most common encryption algorithms. A comparison has been conducted at different settings for each algorithm which means each data block have different key size and speed of encryption/decryption. There is no noteworthy difference when the results are displayed either in hexadecimal base encoding or in base 64 encoding. To change the data type such as image instead of text, it was found that DES has disadvantage over other algorithms in terms of time consumption. Also, AES still has low performance compared to algorithm DES. Finally -in the case of changing key size (possible only in AES algorithms) it can be seen that higher key size leads to change in the battery and time consumption.

Pavithra describes [8] the compares of performance evaluation of various cryptographic algorithms. On the origin of parameter taken as time in various cryptographic algorithms are evaluated on different files. Different files are having different processing speed on which various size of file are processed.

Cufoglu A., [3] discuss NB and IB1 classifiers have the same classification accuracy results. ID3 identifies attributes that differentiate one class from another. All attributes must be identified in advance, must also be either continuous or selected from a set of known values. The limitation of ID3 is that it is sensitive to attributes with a large number of values.

3 SYSTEM DESCRIPTION

The efficient classification algorithm used for improving accuracy level of data. The combination of privacy preserving and access control mechanism is worked together in the system.

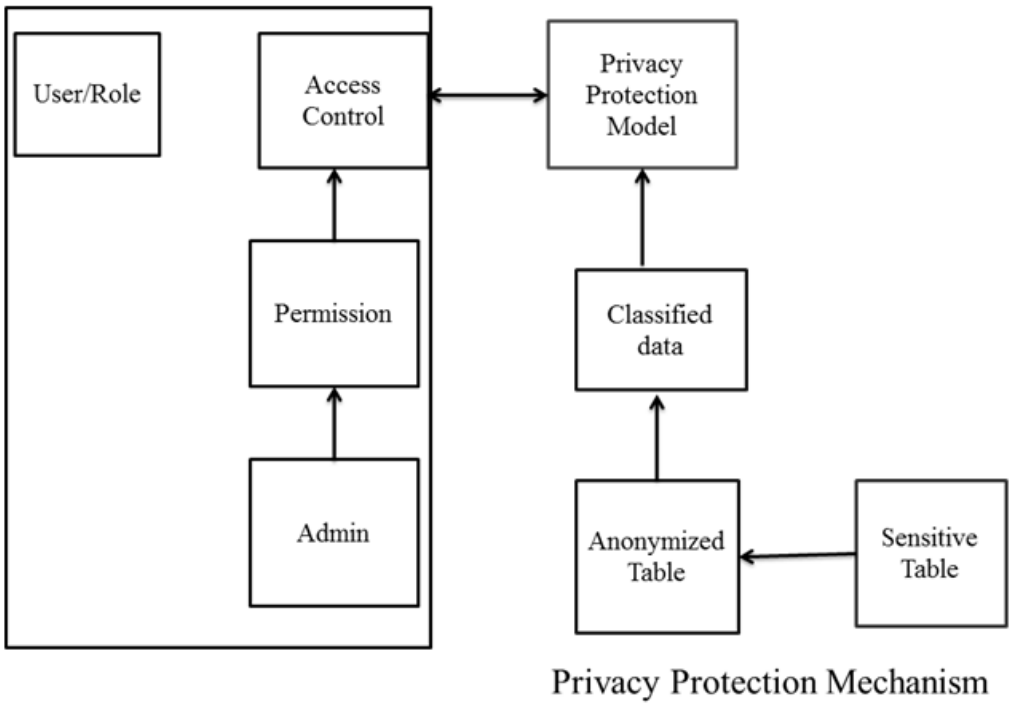

Fig 1: System Description

Access Control describes the security features that control how users and system communicate and interact with one another system. The privacy protection mechanism ensure privacy of sensitive data accomplish by using RSA algorithm.

Role Based Access Control (RBAC) is a centrally administered controls used for determine how subjects and objects interact together with respect to their roles. Access control components specify authorized accesses of a system.

USERS- Person

ROLES- An organizational work function with a clear definition of inherent responsibility and authority. Example: Developer, Director.

OPERATIONS (ops)- An execution of an a program specific function that's invocated by a user. Example: In Database- Insert, Delete, Update is specified operations.

OBJECTS (obs)- An entity that contains or receives information, or details about system resources. Example: Database columns, rows.

The access control mechanism describe as:

- A user has access to an object based on the assign role.
- Roles are defined based on user assignment.
- A role contains smallest amount of permissions to instantiate an object.
- A user is assigned to a role that allows him or her to perform only what's required for that role.
- Permissions are defined based on assignment and responsibilities within object.
- Operations of an object are invocated based on the permissions.
- The object is apprehensive with the user's role and not the user.

4 IMPLEMENTATION

This section describes the implementation of proposed work. The proposed system implemented with the following modules:

- Privacy Protection Mechanism
- Classification Analysis
- Multiuser Access

PRIVACY PROTECTION MECHANISM

The Rivest Shamir Adleman (RSA) method has since that time reigned supreme as the most widely accepted. The public key encryption schemes are rarely used to essentially encrypt messages; they are typically used to encrypt a symmetric key for future bulk encryption.

RSA is a block chipper in which the plaintext and cipher text are integers between 0 and n-1. A size for n is 1024 bits, that means n is less than 2^1024.

Plaintext is encrypted in blocks, with each block having a binary value which less than n. Encryption and decryption are of the form, for some plaintext block are represented by M and cipher text block C. This is a public-key encryption algorithm with a public key of PU={e, n} and a private key of PR ={d, n}.

Steps:

With the definitions of d and e as public and private key, the modulus n must be selected in such a manner that the following guaranteed:

$$(M^e)\mathrm{d}(\mathrm{mod\ n}) = \mathrm{M}^{\mathrm{ed}(\mathrm{mod}\ \varphi(\mathrm{n}))} = \mathrm{M}(\mathrm{mod\ n})$$

This guarantee because C = Me mod m is the encrypted form of the message integer M and decryption is carried out by C^d mod n.

Input: Plaintext M < n

Output: Ciphertext C

1. Generate two different prime p and q.
2. Compute n = p × q for some prime p and q
3. Calculate totient $\varphi(n)$ = (p-1)(q-1)

4. Select for public exponent an integer e such that 1 < e < $\varphi(n)$ and GCD (φ(n), e) = 1
5. Calculate for the private exponent a value for d such that d = e−1 mod φ(n)

CLASSIFICATION ANALYSIS

Classification is the process of construction a model of classes from a set of records that contain class labels. It can be used to extract models describing main data classes and predict future data trends. The basic techniques describes how to build decision tree classifiers, rule based classifiers, Bayesian classifiers.

Data classification is a two-step process such as learning and classification. Learning denotes training data are analyzed based on rules. Classification denotes test data are used to estimate the accuracy of the classification rules. J48 allows classification via either decision trees or rules generated from them.

Decision Tree Algorithm is to discover the way the attributes-vector behaves for a number of instances. It based on the binary classification tree. This algorithm generates the rules for the prediction of the intention variable. Decision trees are the most powerful approaches in knowledge discovery and data mining. It includes the knowledge of research large and complex size of facts in order to determine useful patterns. The idea is essential because it enables modeling and knowledge extraction from the large size of data available. All specialists are frequently searching for techniques to make the process more proficient, cost-effective and accurate. Decision trees are highly efficient tools in many areas such as data and machine learning, text mining, information extraction, and pattern recognition.

In this paper, using j48 classification algorithm approach for implement anonymized table. By doing so the accuracy rate of the J48 algorithm has increased to large extent as compared to the accuracy of the different classification algorithm to be used. J48 is an open source Java execution of the C4.5 algorithm in the weka data mining tool.

The confusion matrix is a helpful tool for analyzing the classifier can distinguish tuples of different classes. From the confusion matrix to analyze the performance criterion for the classifiers in detecting accuracy, sensitivity and specificity have been computed.

Accuracy is the percentage of predictions that are correct. The sensitivity is the measure of the capability of a prediction model to select instances of a certain class from anonymized data set. The specificity corresponds to the true negative rate which is commonly used in two class problems. There are four terms used in computing evaluation measures.

True positives(TP):The positive tuples that were correctly labeled by the classifier.

True negatives(TN):The negative tuples that were correctly labeled by the classifier.

False positives(FP): The negative tuples that were incorrectly labeled as positive.

False negatives(FN): The positive tuples were mislabeled as negative.

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+FP+FN+TN}} \quad (1)$$

$$\text{TN} = \text{Total instance} - (\text{TP} + \text{FP} + \text{FN})$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP+FN}} \quad (2)$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN+FP}} \quad (3)$$

Algorithm for j48 Decision Tree

Input: Anonymized Table(Trained data)

Output: Decision Tree(T)

Steps:

1. Assign T= Create root node and label with splitting attribute.

2. T= Add arc to root node for each splitting node.

3. For each arc do

D= Database created by applying splitting predicate to training data.

4. If stopping point is reached, then T'= Leaf node

Else

T'= Build Decision Tree(D)

T= Add T' to arc.

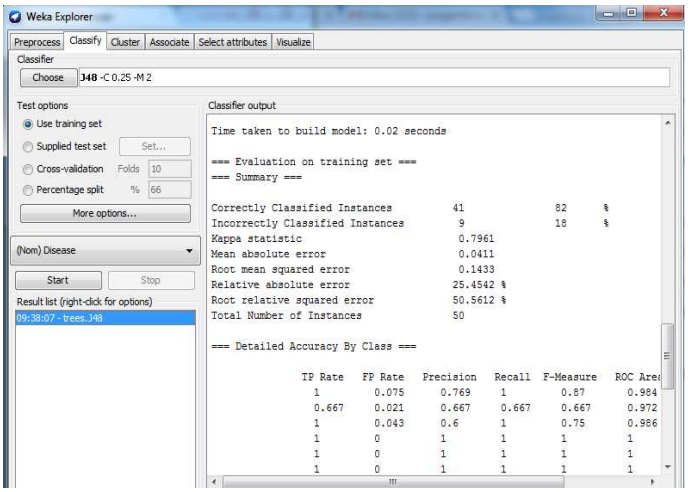

Fig 2- Classified Instance using j48 classifier

```
=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  i  j  k   <-- classified as
 10  0  0  0  0  0  0  0  0  0  0 |  a = Flu
  0  2  0  0  0  0  0  0  0  1  0 |  b = fever
  0  0  3  0  0  0  0  0  0  0  0 |  c = asthma
  0  0  0  3  0  0  0  0  0  0  0 |  d = brain tumour
  0  0  0  0  3  0  0  0  0  0  0 |  e = Typhoid
  0  0  0  0  0  3  0  0  0  0  0 |  f = Acidity
  1  0  2  0  0  0  3  0  0  0  1 |  g = cancer
  0  0  0  0  0  0  0  3  0  0  0 |  h = jaundice
  0  1  0  0  0  0  0  0  2  0  0 |  i = diabetes
  1  0  0  0  0  0  0  0  0  3  0 |  j = Hepatitis
  1  0  0  0  0  0  0  0  0  1  6 |  k = cholera
```

True positive rate = diagonal element/ sum of

relevant row

False positive rate = non-diagonal element/ sum of

relevant row

Average TP rate = 0.842

Average FP rate = 0.026

4.1 MULTIUSER ACCESS

A multiple connections describes a single application might have to make multiple connection to a single database. Many access applications migrate from single user application to an application which provides multiuser access.

5 RESULT AND ANALYSIS

This section describes the result and analysis of proposed method work.

Table I- Time Taken To Classified Instances

Algorithm	L-diversity Dataset(sec)	Anonymized Dataset(sec)
J48	0.03	0.02
Naive Bayes	0.08	0.02
Simple cart	0.42	0.39
Bagging	0.13	0.05

Table I describes the time taken to classify the instances by using different classification algorithm. It shows that the comparison of time taken to form a classified instances using two different dataset such as l-diversity and anonymized table.

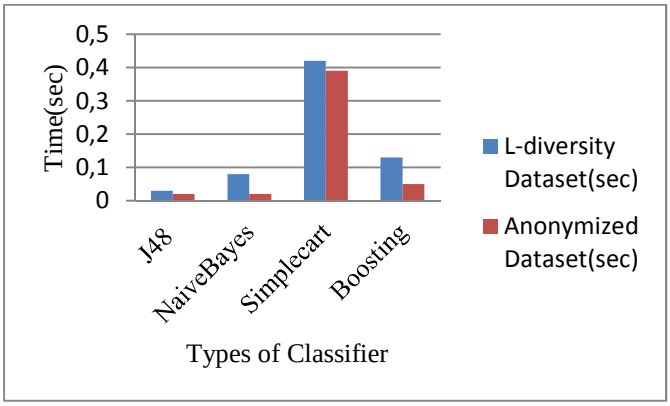

Fig 3- Graphical representation of Time taken for Classified Instances

Table II shows that the comparison of accuracy of anonymized dataset using different classification algorithm and also compare the performance criteria. Figure 4 shows that better accuracy of anonymized dataset compare than other dataset.

Table II- Performance Criteria for Anonymized Dataset

Classifier	Accuracy	Sensitivity	Specificity
J48	84	100	91.43
Naïve Bayes	62	40	90
Simple cart	70	70	93
Bagging	68	80	94.12

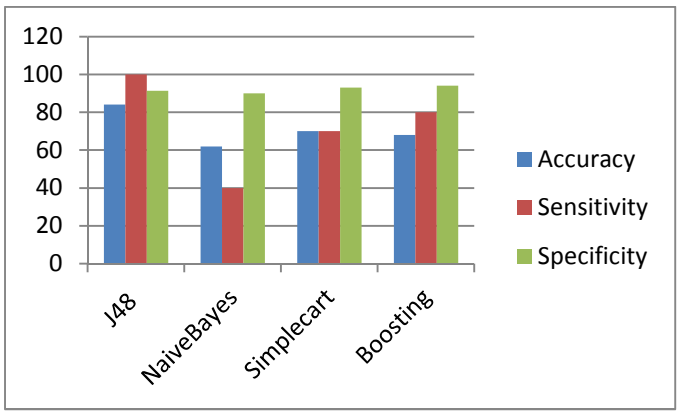

Fig 4-Graphical representation of performance Criteria

6 CONCLUSION

The privacy preserving algorithm provides protecting sensitive data item which are in the form of anonymized table. The privacy preserving techniques anonymized the data to acquire requirements. In this, we use cryptography based anonymization which ensure that the resulting details should be anonymous. Based on the experimental results j48 was able to improve the accuracy of anonymized dataset from 81.6129% to 85.1613%.

REFERENCES

[1] Zahid Pervaiz, Walid G. Aref, " Accuracy- Constrained Privacy-Preserving Access Control Mechanism for Relational Data",id no:10.1109/TKDE.2013.71,2014.

[2] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam,"L- Diversity: Privacy Beyond k-anonymity,"ACM Trans.Knowledge Discovery from Data, vol. 1, no. 1, article 3, 2007.

[3] Panda M.; Patra R. M. (2008) "A comperative study of data mining algorithms for network intrusion detection". Paper appears in ICETET 2008, 1st International Conference on Emerging Trens in Engineering and Technology.

[4] Xiaoxun Sun, Min Li Hua, Wang Ashley Plank. 2008. An efficient hash-based algorithm for minimal Computer Science Conference (ACSC2008), Wollk-anonymity. Australasian ongong, Australia.Conferences in Research and Practice in Information Technology (CRPIT), Vol. 74.

[5] Cufoglu A., Lohi M. and Madani K. (2008) "Classification accuracy performance of Naïve Bayesian (NB), Bayesian Networks (BN), Lazy Learning of Bayesian Rules (LBR) and Instance-Based Learner (IB1)- Comperative Study". Paper appears in the International Conference on Computer Engineering and Systems, IEEE, ICCES 2008, in press.

[6] Diaa Salama Abdul. Elminaam, Hatem Mohamed Abdul Kader and Mohie Mohamed Hadhoud, "Performance Evaluation of Symmetric Encryption Algorithms", IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.12, pp. 280-286, December 2008.

[7] Ya-Qin, L., Cheng, W., & Lu, Z. (2009). Decision Tree Based Predictive Models for Breast Cancer Survivability on Imbalanced Data. Bioinformatics and Biomedical Engineering, 2009. ICBBE 2009. 3rd International Conference on (pp. 1-4). IEEE.

[8] Asha, T., Natarajan, S., & Murthy, K. N. B. 2010. Diagnosis of Tuberculosis Using Ensemble Methods. Computer Science and Information Technology (ICCSIT), 2010 3rd IEEE International Conference on (Vol. 8, pp. 409-412). IEEE.

[9] Slava Kisilevich, Lior Rokach, Yuval Elovici, Bracha Shapira,"Efficient Multi-Dimensional Suppression for K-Anonymity", in proceedings of IEEE Transactions on Knowledge and Data Engineering, Vol. 22, No. 3. (March 2010), pp. 334-347, IEEE 2010.

[10] A. Parmar, U. P. Rao, D. R. Patel, "Blocking based approach for classification Rule hiding to Preserve the Privacy in Database", in proceedings of International Symposium on Computer Science and Society, IEEE 2011

[11] G. Mathew, Z. Obradovic," A Privacy-Preserving Framework for Distributed Clinical Decision Support", in proceedings of 978-1-61284-852-5/11/$26.00 ©2011 IEEE.

[12] Alberto Trombetta, Wei Jiang, Member, IEEE, Elisa Bertino,Fellow, IEEE, and Lorenzo Bossi. July/August 2011. Privacy-Preserving Updates to Anonymous and Confidential Databases.IEEE Transactions On Dependable And Secure Computing, Vol. 8,No.2011.

[13] S. Lohiya and L. Ragha, "Privacy Preserving in Data Mining Using Hybrid Approach", in proceedings of 2012 Fourth International Conference on Computational Intelligence and Communication Networks, IEEE 2012.

[14] S. Pavithra and Mrs. E. Ramadevi, "Performance Evaluation of Symmetric Algorithms", Journal of Global Research in Computer Science, Volume 3, No. 8, pp. 43-45, August 2012.