DOI: http://dx.doi.org/10.26483/ijarcs.v8i9.5042

ISSN No. 0976-5697

Volume 8, No. 9, November-December 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

ENHANCE DATA SECURITY IN CLOUD COMPUTING USING MACHINE LEARNING AND HYBRID CRYPTOGRAPHY TECHNIQUES

Kiran
Department of Computer Engineering and Technology
Guru Nanak Dev University
Amritsar, India-143005

Dr. Sandeep Sharma
Department of Computer Engineering and Technology
Guru Nanak Dev University
Amritsar, India-143005

Abstract: In Cloud computing, the user's can access their important data via internet that is being stored on the remote servers. As technology is growing day by day, there is a rapid increase in the personal and crucial data. This brings up the need of securing the users data. Data can be of any type and each required different degree of protection. In this paper, we have proposed a model that classifies the data according to its security parameters. The performance of the existing KNN is improved by appending it with ensemble learning technique. The basic algorithms of ensemble learning i.e., base level-0 and meta level-1 are modified. This will improve prediction capability and classification accuracy of existing KNN technique. Also to secure sensitive data, the hybrid cryptography technique is used. The quantitative analysis show that the accuracy of ensemble learning when combines with existing KNN is 73.5% whereas the accuracy of KNN was 65.5%.

Keywords: Cloud Computing, Data Classification, Security, Machine learning technique, Cryptography

1. INTRODUCTION

Cloud computing is an emerging virtual distributed environment that uses the ideas of sharing, power processing, storing, connectivity, and virtualization. Communicating through broad network i.e., Internet cloud facilitate a large pool of resources, storage media and sharing media that helps to supply on-demand services. This will help the end users in complying with the ideas of isolation, elasticity, security and distribution [1].

Security issues are the foremost difficult problem in cloud domain and hence, the most vital obstacle for aggrandize of IT-based companies that provide users on-demand services. These security issues can be visualize at application phase, network phase, authentication/authorization phase, information storage and at virtualization phase. These challenges or threats are still an obstacle within the complete success path of cloud computing. One reason is that consumers and plenty of organization keep their information on cloud database, so the main focus is that the user's information should be safe, and the vital information shouldn't get drift and tampered when travelling from one place to a different across the network. Thus, it is essential that I(integrity), C(confidentiality) , and A(availability) of user information ought to be ensured. Another reason is that, the unauthenticated user tries to access the authenticated user's data.

We can apply cryptographic algorithms in cloud servers to solve these threats. However when a user is revoked, using a single cryptography algorithm is not adequate to assure the confidentiality of data and managing the access control methods in cloud computing environment. For data security these techniques are applied on encryption. Encrypting complete data can turn to be very expensive in terms of memory as well as for time. So, to solve this problem it would be better if we first separate our sensitive data and then apply encryption algorithms. It would address reliable results if we classify the data according to its sensitivity level[2].

In machine learning field, the data classification is a method of distinguishing the category of unclassified data sample set with the help of build classifier. The classifier is constructed by constructing a training set of familiar data samples. To create an appropriate classifier, a large range of justified training data samples are required. This advancement invites a new paradigm of services where data classification is offered by servers in a cloud to its various clients/users. Specifically, the server will process the data automatically and hence, categorise the clients' data samples present on remote private servers. However untrusted third-party-servers can access the private data. Moreover, any vital detail or training data set specifications may not be disclosed by the servers even if it provides the classification services to its client. Thus, a mechanism that ensures the privacy of the server's training set and client data samples is required. Hence, a re-encryption model is essential requirement to forestall the revoked user from accessing the encrypted information as well as to generate reliable keys for valid users. Hence, in this paper a hybrid re-encryption model based on index classification which will categorise the data on the basis of sensitivity.

2. RELATED WORK

The users of cloud get great benefit and advantages in terms of accessibility, availability, confidentiality of data through Internet. Also the data is stored at different servers so that the user can access data from anywhere and at anytime without bringing the physical storage devices. Shared data can be accessed by clients thereby enhancing the collaborative efforts. As there is no need to buy expensive hardware's, accessing through cloud storage can be economical. Moreover cloud facilitates recoveries from backup, disaster and archival. Despite these benefits of cloud there exist some major limitations. As in public cloud the information is present on remote servers and is not in

control of legal users. Due to this the data can be accessed globally by unauthorised users and makes cloud computing model unsecure.

Cloud computing issues from consumers' perspective was discussed in[3]. The main security issues discussed are: securing the information and secrecy protection. Different techniques like Airavat, to protect the data had been suggested. The approach used was not practical and also rely on different methods for model execution.

To secured information mechanism in cloud, AES technique was used in [4] which guarantees the security of the users' information present on the cloud servers.

The matter of fine-graininess, information confidentiality and quantifiability in cloud was discussed in [5]. The different techniques like KP-ABE, re-encryption techniques (proxy and lazy) was used in the paper.

Cryptography and steganography techniques was used for storing the data and information by author in [6]. The model has represented a 3 step data security model to secure the users' data. The steps involves: using cryptography algorithms with RSA, hide the data with stenography technique and in finally the accessing the data by decrypting via RSA algorithm.

A data classification model that result in minimum overhead and processing time was discussed by author in [7]. In this paper, different security mechanism with varying key length that provides data confidentiality was discussed. The model was evaluated with different encryption algorithms that show better result in terms of reliability and efficiency. The drawback of the paper was that the automatic classification of the data was needed. Moreover for higher degree of security and confidentially more secured cryptographic algorithms like RSA, Elliptic curve cryptography was required.

A Confidentiality based data classification model for cloud computing was suggested in [2]. In this paper K-NN (K Nearest Neighbour) technique was used for the classification of data. The main focus was to decide what kind of data need to be secured and which data should be kept as public. K-NN was used for classification of data. The data was categorised into sensitive data and non-sensitive data. The confidential data was secured using RSA algorithm and the basic data was accumulated on the cloud servers. Drawback of the paper was that only RSA algorithm was used for sensitive data. Moreover automatic recognition for the categories (basic, confidential, highly confidential) by K-NN model was not implemented.

Most of these techniques used above were used for encrypting the data without considering its sensitivity level. Encrypting the whole data would be very expensive to its users' so to reduce the cost, first categorize the data into different classes (highly confidential, confidential or basic) and then apply the encrypting techniques only to most sensitive data. This would result in saving the encryption/ decryption time but also become economical for its users.

3. PROPOSED WORK

Data classification is the task of identifying data sets with respect to its data value. These values are based on usage consumption of data by its users and restrictions on access control methods [8]. KNN technique is used in machine learning artificial intelligence that helps to categorise the class of unclassified data with the help of build classifier. It is constructed by employing a training set of familiar data samples.

In this proposed work, the modified Ensemble Learning Technique is used to enhance the execution of existing KNN technique.

Ensemble learning method comprises a set of different models are group together to improve the prediction and stability power of any model. It has two levels: base level-0 and Meta level-1 as shown in fig: 2. At base level a no. of algorithms can run ie, AdaBoost and bagging algorithm. At Meta level, also known as decision making algorithm, random forest tree is used. To utilize the training sets of data which is given by KNN model, the training set is computed with Euclidean distance function. To reduce the computational density, we improved the basic algorithm of ensemble learning.

How modified model algorithm works:

The classifier is evaluated using a cross validation (K-fold) technique. Each layer is trained as:

- The dataset is split into two sets: training set and testing set.
- For each base level layers classifiers are generated using Training set.
- The predictions generated by the base level is assembled and forwarded to meta level. At meta level, validation process is performed and provides the final predicted results.
- Finally, by using complete layer's datasets (not the use of simplest out of k folds of the dataset) we can train each layers of the training set.
- Now to secure highly confidential data we combine RSA encryption technique with HMAC. The HMAC(hashed message authentication code) is a cryptographic checksum. It is stored at local machine as well as appended with the cipher text and sent to the cloud. The HMAC+ cipher text must be matched with the HMAC checksum i.e, already stored at the local machine of user. The hacker finds it difficult to guess the HMAC checksum while hacking the data. Thus, results in increasing the security of user's HC data by using RSA encryption technique.
- HMAC is calculated by following formula:
  HMAC(Key,msg)=HMAC(KeyXORout)||H((Key XORin)||msg))
  Where Key is the original key, m is the message that needs to be authenticated.

Cloud simulation environment:

The cloudsim as shown in fig-1 is used for the proposed model that solves the problem of confidentiality and classification of data. Firstly we run the simulation then data centres' are build. The virtual machine manager (VMM) is used to handle the VMs and for allocating VMs to cloudlets i.e., the cloud task. Authentication process is performed so that only authenticated users can access the data. Then classification KNN is combined with modified Ensemble learning technique. This will improve the prediction capabilities and accuracy of existing KNN.

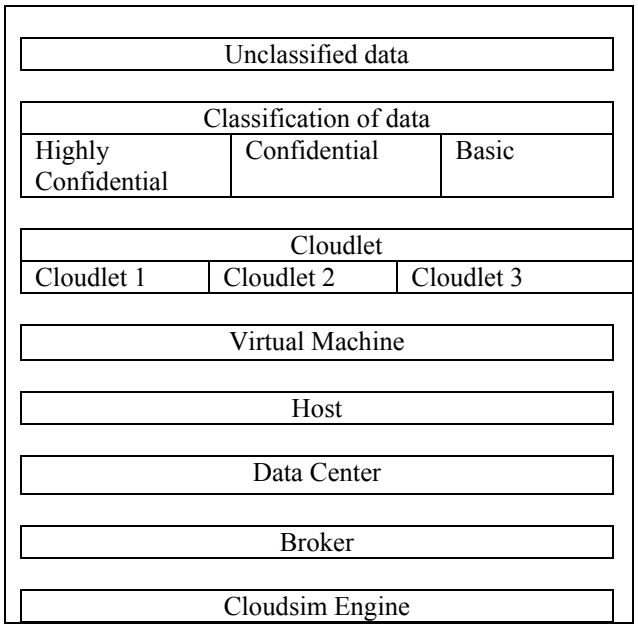


Figure 1: Proposed Simulation Environment

STEP 1: To select Base and Meta layers

I/p: o.ds: DS, folds: Int
DS←o.ds
AoC as classifiers array that consist AdaBoost and Bagged Decision Tree
For lr= 0 to 2 do:
For each fold in folds do:
If lr ≠2 do:
AOC←Trn_S.Lr (lr, t.s, folds)
In_New←Classify (test-set,AoC)
Adding In_New to ds [lr+1]Else
Trn_S.Lr (lr, ts, folds)
Lr = lr + 1
Trn-S.Lr (lr 0, o-ds)

STEP 2: Single layer classification

Trn-S.Lr I/p: Lr
No: Int, ds: DS, folds: Int
O/p: Scr-Ds: DS
Scr DS ← emp Grp
For each fold in folds do:
B.C (Lr No, ts)
For each In. in ts
Produce probabilities-vector by applying In. on current layer's classifiers.
build a new In. from probabilities-vector
Add the new In. to Scr-DS
Ret Scr-DS

Where:
o.ds: Original dataset
DS: Dataset
Int: Integer
AoC: Array of classifier
Lr: Layer
Ts: Training set
In: Instances
Scr: Successor

Properties and description Cloud Service Model:

Before starting the simulation it is necessary to set the properties of SaaS(software as a Service), PaaS(Platform as a Service) and IaaS(Infrastructure as a Service). The properties are as follows:

Table 1:The properties of SaaS model is deployed with Virtual machines in simulation environment. ID=identification no of specific cloudlet. Length= Cloudlet size. I/O file sizes are measured in bytes.

Table 1: SaaS model attributes

ID	Cloudlet size	Size of input file	Size of output file
0	4000	158	158
1	3000	139	139
0	4000	158	158

Table 2: The property of PaaS model in application deployment layer contains the properties of VM. The strength of the basic node is given to the VM layer. MIPS: machine instruction per second, BW: bandwidth, Pr. Number: processor numbers used in VM.

Table 2: PaaS model attributes (VM management)

VM ID NO.	MIPS	Size of image	BW	Pr.No	Virtual m/c manager
0	100	1000	1000	1	Xen
1	100	1000	1000	1	Xen

Table 3: Shows the properties of IaaS model. Here data centres' are assigned to VM.

Datacenter ID	RAM in Mb	Storage limit	Architecture of data	Operating system	BW
2	2048	10000000	X86	Linux	10000
3	2048	10000000	X86	Linux	10000

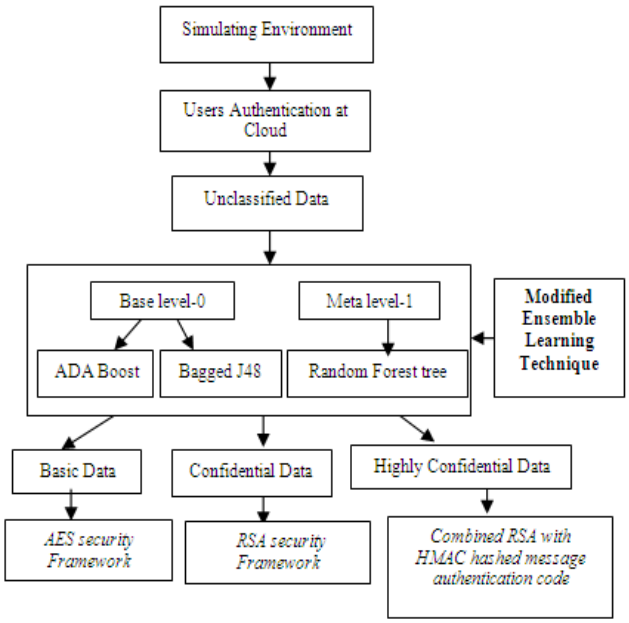


Figure 2: Proposed Model

4. RESULTS AND DISCUSSIONS

The results of our proposed system has been demonstrated and discussed in this section. Comparison of existing work with the proposed work has been shown. According to the following analysis, it has been observed that the system proposed in this research work is giving improved and reliable results. For higher degree of confidentiality and security different cryptographic techniques are required. The performance of the proposed system has been evaluated by following parameters i.e., Accuracy, Classification details, Error Rates, Encryption Time and Decryption Time.

1. Accuracy: It is the measure of correctly classified instances to the total number of correctly and incorrectly classified instances.
   TP: True Positive, TN: True Negative, FP: False Positive, FN: False Negative

$$\text{Accuracy}=\frac{TP+TN}{TP+FP+TN+FN}$$

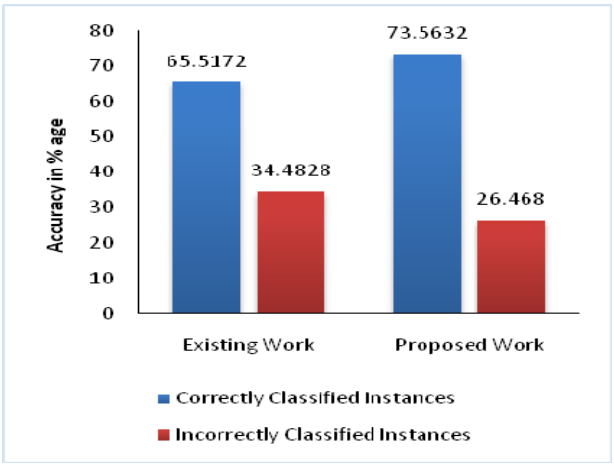


Figure 3: Comparison of accuracy b/w the proposed and existing model

2. Encryption Time: Encoding information or any message in such a way that only authorised party can be able to access it.

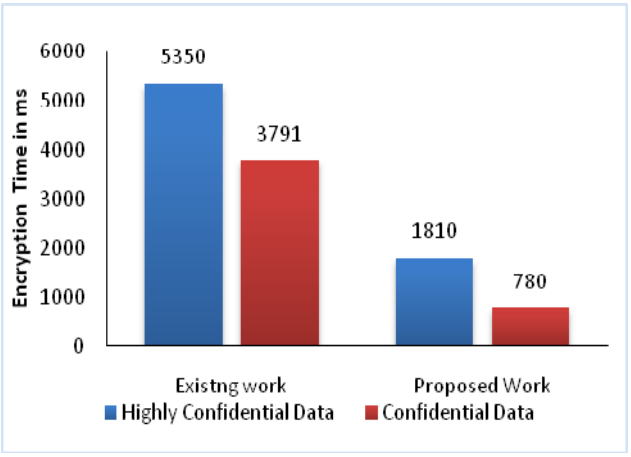


Figure 4: Comparison of encryption time b/w the proposed and existing model

3. Decryption Time: It is the process of reconverting the encrypted data or message into its original form.

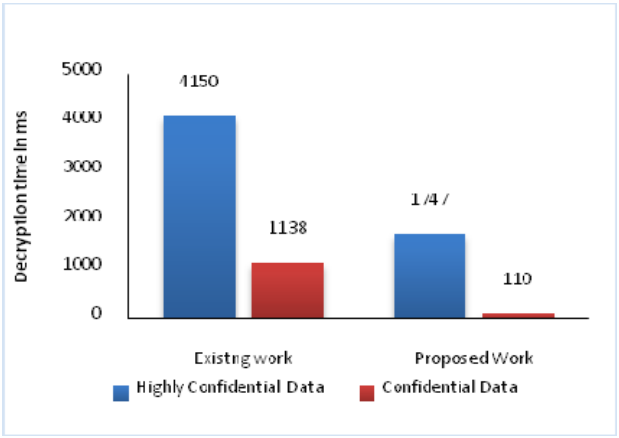


Figure 5: Comparison of decryption time b/w the proposed and existing model

4. Classification details:
   - Precision and recall: Both are used for evaluating execution capability and in data analysis field like retrieving data and data mining. Precision measures the exactness and recall measures the completeness of data.

$$\text{Precision}=\frac{TP}{TP+FP}$$

$$\text{Recall}=\frac{TP}{TP+FN}$$

   - F-measure: It is the Harmonic mean of precision and recall.

$$\text{F measure}=\frac{2*recall*precision}{precision+recall}$$

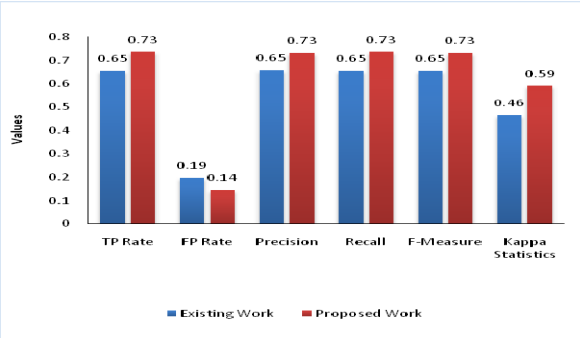
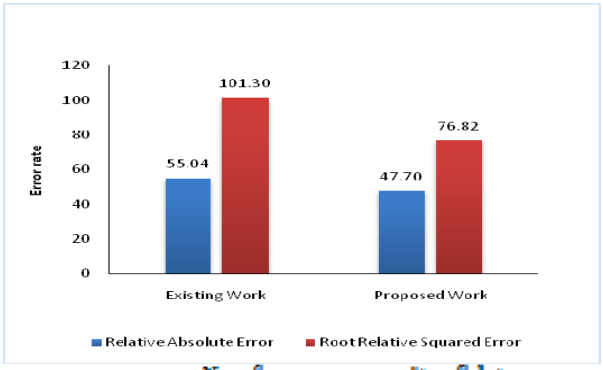


Figure 6: Comparison of classification details b/w the proposed and existing model

5. Error Rate: The measurement of the effectiveness of a communications channel.



$$\text{Error rate} = \frac{\text{No. of erroneous units of data}}{\text{Total data}}$$

Figure 7: Comparison of error rates b/w the proposed and existing model

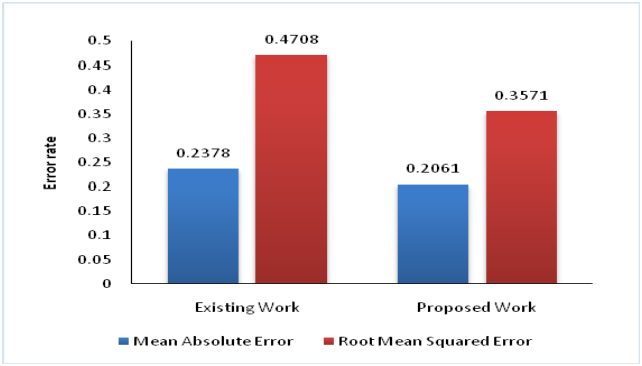


Figure 8: Comparison of error rates b/w the proposed and existing model

5. CONCLUSIONS AND FUTURE SCOPE

This paper aims at achieving data confidentiality, data access control management and provides an automatic classification of data in cloud. On the contrary, the existing system was lacking in providing automatic classification of data. In this paper the functionality and performance of existing KNN technique is improved with modified Ensemble Learning technique. The data will be automatically classified by the machine on the basis of data security parameters. Also to emphasis more focus on highly confidential data HMAC function has been appended with the existing RSA encryption algorithm. The proposed system proves to be more accurate and economical. And also saves the user time for encrypting/decrypting different classes of data (basic, confidential and highly confidential).

REFERENCES

[1] F. F. Moghaddam, M. Vala, M. Ahmadi, T. Khodadadi, and K. Madadipouya, "A reliable data protection model based on re-encryption concepts in cloud environments," Proc. - 2015 6th IEEE Control Syst. Grad. Res. Colloquium, ICSGRC 2015, pp. 11–16, 2016.

[2] M. A. Zardari, L. T. Jung, and N. Zakaria, "K-NN classifier for data confidentiality in cloud computing," 2014 Int. Conf. Comput. Inf. Sci. ICCOINS 2014 - A Conf. World Eng. Sci. Technol. Congr. ESTCON 2014 - Proc., 2014.

[3] D. Chen and H. Zhao, "Data Security and Privacy Protection Issues in Cloud Computing," 2012 Int. Conf. Comput. Sci. Electron. Eng., no. 973, pp. 647–651, 2012.

[4] N. Surv, B. Wanve, R. Kamble, S. Patil, and J. Katti, "Framework for client side AES encryption technique in cloud computing," Souvenir 2015 IEEE Int. Adv. Comput. Conf. IACC 2015, pp. 525–528, 2015.

[5] S. Yu, C. Wang, K. Ren, and W. Lou, "Achieving secure,scalable ,and fine-grained data access control in cloud computing.pdf," Ieee Infocom, pp. 1–9, 2010.

[6] V. K. Pant, J. Prakash, and A. Asthana, "Three step data security model for cloud computing based on RSA and steganography," 2015 Int. Conf. Green Comput. Internet Things, pp. 490–494, 2015.

[7] L. Tawalbeh, N. S. Darwazeh, R. S. Al-Qassas, and F. AlDosari, "A secure cloud computing model based on data classification," Procedia Comput. Sci., vol. 52, no. 1, pp. 1153–1158, 2015.

[8] R. Shaikh and M. Sasikumar, "Data classification for achieving security in cloud computing," Procedia Comput. Sci., vol. 45, no. C, pp. 493–498, 2015.