Secure Classification Of Encrypted Cloud Storage By Using SVM

Ashwini Sarode
Department of Computer Engineering
SKN College of Engineering
Pune, India

Mrs. Saudagar Barde
Department of Computer Engineering
SKN College of Engineering
Pune, India

***Abstract*—with the efficient development and popularity of cloud service providers, huge amount of data is available. This data should be mine. This mine data can be used for various kind of real time and other applications including medicine, scientific research, banking etc. There is a various data mining approaches are available to mine the data available on cloud servers. Among the data mining approaches, classification is very important and challenging task. A classifier is used to define the suitable class for each text document based on the input algorithm used for classification. In past decades, various classification schemes have been proposed. But these schemes are not efficiently applicable for encrypted data, mean they are not able to classify the encrypted data store on cloud servers. The data store on cloud is encrypted, because most of the users store their data on cloud in encrypted format to preserve the security and privacy. In this paper, we have proposed a classification approach to tackle this kind of issue. A SVM classifier is implemented to classify the encrypted data stored at cloud servers and also we remove two disadvantages of SVM by calculating reachability and cover ability on encrypted data to magnify the data. it makes them useful for various applications also maintained the security of user query and their data access patterns. Experimental evaluation of the system prove that the classification of encrypted data is more accurate and efficient with SVM classifier than the KNN classifier. The proposed protocol secures the encrypted data, assure the client's query and hides the data access patterns. Proposed algorithm is memory efficient as well as time efficient due to every process regarding data is done on the cloud.**

Keywords—Cloud computing, data mining, classification, encrypted cloud storage, security and privacy of encrypted data, SVM, reachability and cover ability.

I. Introduction

Today's digital system allows a user's to store and retrieve their data. In fact user can store their data on remote servers and perform various operations on that data. These remote servers are managed by third party people, this is known as a cloud service providers in cloud computing area. Formally cloud computing is defined as, a type of computing based on Internet-based that allows a sharing of processing resources and data on demand. Computing resources includes, networks, servers, storage, applications and services). Cloud system are cost effective, flexible and reducing the administration overhead of various organizations also provide the various operations over data stored in remote servers.

Cloud storage online space where you can store your data or keep backup of files, through cloud service providers by selecting some data storage package. At this point, cloud storage is physical storage including, hard drives or sub flash devices etc. With this storage system, security and reliability is become a challenging issues. It is very important to keep data secure and original. Data should not be modified. The confidential data store at cloud might be leak or modified by third party unauthorized entities like hackers. To keep security of data over cloud, various techniques are available such as,

- Encryption: It is the technique, used to hide the original information using various encryption algorithm.
- Authentication: In this technique, every valid user has its own user name and password.
- Authorization: In this technique, only valid authenticated users can be able to access the cloud storage data.

In this paper, to maintain the security of data stored on cloud storage, we have focused on the encryption technique. In this technique, the users can encrypt their data before storing on cloud. Various encryption schemes are available with some advantages and disadvantages. Encryption schemes includes, AES, ECC, RSA, RC4 etc.

In our proposed system, we have used Paillier cryptosystem, to encrypt the data file, user queries to server and their responses to query. It will provide more security than other encryption algorithms. This is the holomorphic and probabilistic public key encryption. A security is based on the assumption known as decisional composite residuosity assumption, in paillier algorithm.

This is all about security of data, but there is another challenging task, to use this encrypted data in various real time applications. To improve the security, authorized trusted users are able to query the data hosted at the cloud service provider. Due to data encryption, the process of query evaluation over encrypted data becomes challenging. The data mining on such encrypted data is very hectic and difficult process. The classification is a main task of data mining, used in various applications.

Since the recent improvement of cloud computing, now, users can outsource their data, in encrypted form, as well as the data mining collaborated with the cloud, the existing privacy preserving classification methods are not valid, such as KNN. There is a need of advanced classification system. In this paper we will implement the Support Vector Machine (SVM) classifier for classifying the user encrypted query over encrypted dataset. Also remove two disadvantages of SVM by calculating reachability and cover ability on encrypted data to magnify the data.

SVM is based on the concept of decision planes use to define decision boundaries. This decision plane is use to separate the object set with different class membership. In our project it is used to classify the encrypted query by testing it against trained encrypted data store on cloud. SVM is more beneficial because of its silent feature that is margin maximization and the classification via kernel tricks, which is more effective in many real-world applications.

In this paper further we will see: Section II talks about related work studied till now on topic. Section III current implementation details, introductory definitions and documentations and in addition formally expresses the proposed work undertakings tended to by this paper. Section IV describes results of experimental evaluation and this paper is concluded in section V.

II. LITERATURE SURVEY

In paper [1], authors implemented a system for cloud storage which provides security for the data while outsourcing as well as retrieving data from client to making sure of integrity of the data. By making use of SSL system makes sure that data privacy must be encrypts before outsource. Also Predicate Based Encryption method is utilized for encrypting the data from client. This technique concentrates on multiple data owner scenario and splits the data based upon the user severity rating (SR) such as low, and medium and high severity level. The test results showed the system gives a high performance in terms of privacy performance of client data.

In paper [2], authors concentrated on sorting out the issue of classification problem on encrypted data. For being specific authors presents a secure k-NN classifier on encrypted data in the cloud which will secure the importance of data, privacy of query given by user's also makes data access patterns are hidden. Authors also analyze the efficiency of proposed protocol experimentally by making use of a dataset of real-world under various settings of the parameters.

In paper [3], authors suggested a classification algorithm which is improved and a workflow in case of detection on encrypted data structures on a storage device. The second part i.e. the classification of workflow is depending on statistical tests. To reduce the number of falsely classified data which is not encrypted data structure in the important goal keeping investigators in to the focus. Authors given and also evaluate a tool for automated analysis of storage devices which develops a multitude of statically tests for detection of encrypted data, compared to both the application of only one such test and the calculation of entropy. According to authors the tool they developed is able to sort the high-entropy file format such as DOCX, JPG, PDF etc. from encrypted file.

In paper [4], authors have created a framework which is known as CUMMA used for classifying service usages of mobile messaging Apps by modeling user behavioral patterns commonly, characteristics of network traffic also temporal dependencies. They also built the clustering Hidden Markov Model (HMM) method to find mixed dialogs from hackers and retrieve mixed dialogs of single-type usage. CUMMA also allows mobile analysts to find the service usages as well as study end-user in app behaviors also in internet traffic which is encrypted. Test results show the effectiveness and efficiency of system author proposed.

In paper [5], Authors developed a technique for calculating the likeness in two encrypted data points. They also make some improvement in Jaccard similarity function accompanying Private Equality Test protocol assisting the process of a semi honest 3rd party to take an equality test. The developed system is efficient method for calculating the likeness with reduced cost of communication for data mining.

In paper [6], authors proposed the new Protocol for Outsourced SVM (POS). POS allows the user and the cloud to do combine operation on the data which is encrypted and outsourced data without affecting data privacy added by every user. They also confirmed that POS is accurate and secure.

In paper [7], authors made an analysis on a privacy-preserving (PP) data classification method in which server is not able to study any knowledge related to the clients input data at the same time server side classifier is kept hidden form the clients while classification process. Also they developed a client-server data classification protocol which is first of its kind by making the use of support vector machine. Presented protocol does PP classification for two-class as well as multi-class problems. The protocol makes use of properties of Paillier homomorphic encryption for securing two-party computation.

In paper [8], authors given a new and unique order-preserving encryption (OPE) based ranked search method on encrypted cloud data, which makes use of the encrypted keyword frequency for ranking the result and gives a perfect result by way of two-step ranking strategy. In the first step documents are ranked with the measure of coordinate matching, in next step, a fine ranking process is executed for every category found in previous step by adding up the encrypted score.

In paper [9], authors gives a holistic and efficient way which is consist of a secure traversal system as well as encryption scheme dependent on privacy homomorphism. The system can be merged to big datasets by using index-based approach. Also many techniques used for optimization are also tends to improve the efficiency of the query processing protocols. Presented technique is confirmed by theoretical analysis and performance study.

In paper [10], authors gives a design and developed a protocols having security saying that those are much stronger than are possible with semi-honest protocols at minimum cost. They have made use of previously done work by Mohassel and Franklin the main concept in protocols developed is to carry two runs of a semi-honest, garbled-circuit protocol along with the parties swapping roles, take the place of less costly secure tested equality. Also authors have proposed some heuristic improvements to reduce the overall information a cheating adversary learns.

III. IMPLEMENTATION DETAILS

In this section discussed about the proposed system in detail. In this section discuss the system overview in detail, proposed algorithm, mathematical model of the proposed system,

A. Problem Definition

Implement a paillier algorithm to encrypt the data before storing at cloud and also encrypt the requesting query to the server to provide more security and confidentiality of data.to calculate reachability and cover ability on encrypted data to magnify the data. Also perform the classification of encrypted data store on cloud server, to make that data useful in real time applications.

B. System Overview

This paper overcome the Data Mining over Encrypted Data (DMED) issue in cloud computing. For encryption purpose system uses Paillier cryptosystem technique. Using this algorithm, data as well as user query is encrypted. So that the confidentiality is maintained also no one knows the original query of users. The main aim of the system is to solve the classification problem over encrypted data. To achieve this, SVM classifier is implemented. Which performs better than KNN classifier.

Overall system working is depicted in Fig. 1. Figure shows the system with 2 clouds namely cloud 1 and cloud 2, organizational user to store their dataset and third party users to request a query for particular information. To provide authentication and authorization, user must be login to system. Only authenticated users can be able to fetch the encrypted information store on cloud server.

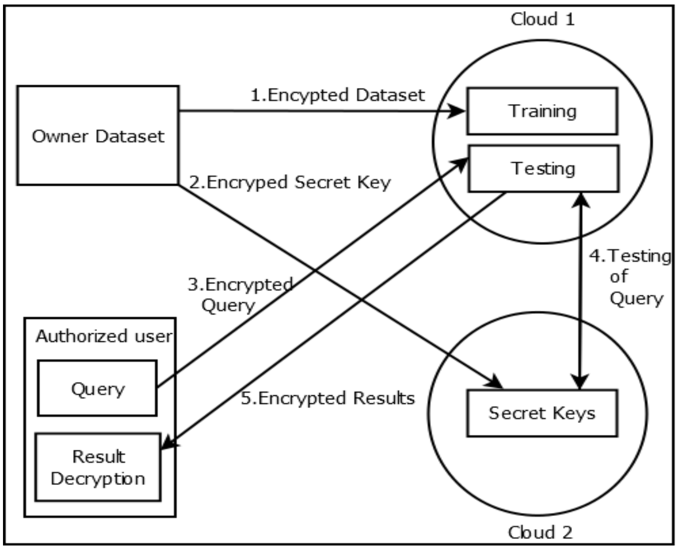

Fig. 1. Classification over Encrypted Cloud Storage

Suppose first User want to store their data on cloud 1 in encrypted format. He has public and private key pair. This user encrypt its data with public key. This encrypted data is store on cloud 1 and on cloud 2 private or secret key is stored. This secret key is further used by cloud to decrypt the data and in the process of query result generation. Now suppose that, if another user wants to fetch some data from dataset already store on server. At that time, that user must be the part of same organization and he must be authenticated. Then this user now querying to cloud 1 in encrypted format. That is user queries to cloud for data classification. Cloud 1 has some data, but its secret key to decrypt that data, is store on cloud 2. Therefore, cloud 1 and 2 are communicate with each other and perform data classification by applying SVM classifier. After results generation, cloud 1 encrypt the result and send it to query user. At user side, he receives that encrypted results and decrypt it using its key. Before this we calculate reachability and cover ability by using algorithm on encrypted data to magnify the data. Which help us to remove two disadvantages of SVM mainly performance issue in large dataset and second when data contain noisy data.

B. Algorithm

Algorithm 1

Require 1. Alice/Organizer database D of n records t1…tn and m+1 attributes
2. authorized user Bob/client query with query q = <q1,…,qm >
3. Encrypted dataset Epk (Di')
4. Encrypted query Epk (qi')
5.public key Pk
6.secrate Key Sk

```
1   Alice/Organizer does :-
        a) Get database D with n records t1…tn
           and m+1 attributes.
        b) Encrypt database
           Compute Epk ( Di' ), for 1<j<m ←
           paillier Cryptosystem ( Di )
        c) Send Epk ( Di' ) = < Epk ( Di1'
           ),…………., Epk ( Dim' )> to C1
        d) Send Sk to C2
        e) Alice detach from procedure.
2   Bob/client query does :-
        a) Get query q with q1,….qn attribute to
           evaluate on encrypted dataset Epk (
           Di' )
        b) Encrypt query
           Compute Epk ( qi' ), for 1<j<m ←
           paillier Cryptosystem ( qi )
        c) Send Epk ( qi' ) = < Epk ( qi1'
           ),…………., Epk ( qim' )> to C1
3   C1 and C2 :-
        a. C1 receive Epk ( qi' )
        b. Assign weight to each instance
        c. For S=1 to k do
           Generate id for instance
           Compute k(I,j)=Epk(Drka(si,j)*t)
                  4.   C1
        d. compute
           T ←reachability and coverabilty (Epk (
           Di' )) and for Epk ( qi' )
        e. get encrypted weight vector
           w={wa+wb}
        f. get T
               5.   C1 and C2
        g. for 1<i<n
           compute Epk(class_lable)←over(Epk (
           Di' )) and for Epk ( qi' )
        h. return Epk(class_lable) to C1,
                  6.   C1
        i. return Epk(class_lable) to bob i.e query
           client
  7.    Bob
        j. Get Epk(class_lable)
        k. Decrypt Epk(class_lable) ←Sk
```

Algorithm 2

reachability and coverabilty (weight)

RequireAlice and Bob's encrypted attributes (Epk (Di')) and Epk (qi') , public keys pk

```
for all x belongs to T do
    if x classified incorrectly by k nearest neighbours then
        flag x for removal
for all x belong to T do
    if x flagged for removal then T=T-{x}
> Iterate until no cases removal:
repeat
    for all x belongs to T do
        compute reachable(x)
        compute coverage(x)
    progress=false
    for all x belongs to T do
        if |reachable(x)|>|coverage(x)| then
            flag x for removal
            progress=true
    for all x belongs to T do
        if x flagged for removal then T=T-{x}
until not progress
return T
```

Algorithm 3

SVMCl

Require: For Server: encrypted test data (Epk (Di')) and Epk (qi') encrypted weight vectors w∗;

```
1. compute a = ∑mal=1 (wl · Epka (xl))
2. compute b = ∑ml=ma+1 (wl · Epkb (xl))
3. compute class lable = a + b − i + Epka (ta) +
   Epka (tb),
   where ta = Ta(ida) and
         tb = Tb(idb)
return class lable to C1
```

C. Mathematical Model

Given training vectors in two classes, and a vector. SVC solves the following primal problem:

$$\min_{w,b,\zeta} \frac{1}{2} w^T w + C \sum_{i=1}^{n} \zeta_i$$

Subject to:

$$y_i(w^T \phi(x_i) + b) \geq 1 - \zeta_i, \zeta_i \geq 0, i = 1, \ldots, n$$

Its dual is

$$\min_{\alpha} \frac{1}{2} \alpha^T Q_{\alpha} - e^T \alpha$$

Subject to

$$y^T \alpha = 0$$

$$0 \le \alpha_i \le C, i = 1,\ldots,n$$

Where e is the vector of all ones, C > 0 is the upper bound, Q is an n by n positive semi definite matrix, $Qy = yi\ yj$ K(x_i, x_j) Where K(x_i, x_j) = $\emptyset(x_i)^T\ \emptyset(x_j)$ is the kernel. Here training vectors are implicitly mapped into a higher (maybe infinite) dimensional space by the function$\emptyset$. The decision function is:

$$\mathrm{sgn}(\sum_{i=1}^{n} y_i \alpha_i K(x_i, x) + \rho)$$

TABLE I. State transition table

	W1	W2	W3
S1	S2		
S2		S3	
S3			S3

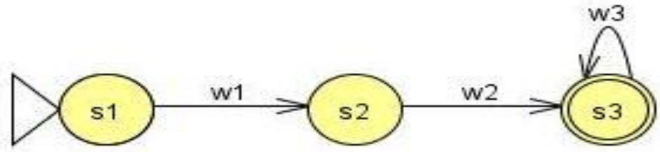

Fig. 2. State transition diagram

D. Experimental Setup

The system is developed using Java framework-version jdk 8 on Windows platform. The Netbeans version 8.1 is used as a development tool. Jung tool is used for the generating the network which contain sensor nodes. The system does not require particular hardware to run, any standard machine is capable of running the application.

IV. RESULT AND DISCUSSION

A. Dataset

To evaluate the performance of proposed system, a car dataset is used. This dataset contain 1728 records with content number of cost management and as an attribute number of features. It contain 6 attributes such as, buying (v-high, high, med, low), maint (v-high, high, med, low), doors (2, 3, 4, 5), persons (2, 4, more), lug_boot (small, med, big) Safety (low, med, high). There is also an individual class feature and the dataset is classified within various classes.

B. Results

TABLE II. Time Comparison

List Size	Existing System	Proposed System
1	130000 ms	110000 ms
2	80000 ms	70000 ms
3	120000 ms	105000 ms
4	150000 ms	130000 ms
5	70000 ms	61000 ms

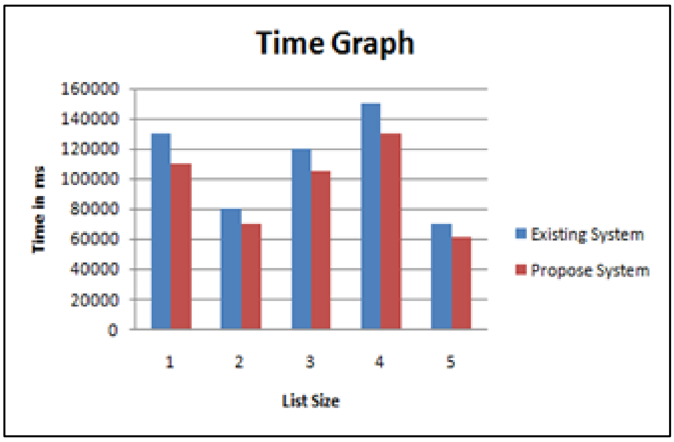

Fig. 3. Time comparison existing and proposed system

The fig. 3 shows the time Comparison Graph between Existing and Propose System. The propose system used SVM classifier than the KNN classifier. It has more classification speed that's why propose system save the time than existing system.

TABLE III. Memory Comparison

List Size	Existing System	Proposed System
1	4700000 bytes	4100000 bytes
2	4200000 bytes	3900000 bytes
3	4900000 bytes	4500000 bytes
4	4100000 bytes	3800000 bytes
5	4300000 bytes	4100000 bytes

The fig. 4 shows the memory comparison Graph between Existing and Propose System. The propose system used SVM classifier than the KNN classifier. It required less memory for storage that's why propose system save the memory than existing system.

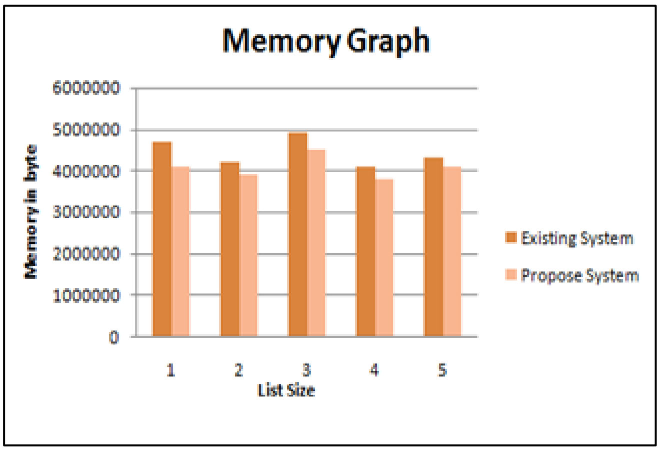

Fig. 4. Memory Comparison between Existing and Propose System

V. CONCLUSION

Various privacy preserving classification techniques have been proposed from many years. But these techniques are not efficiently classified the data stored on cloud server. Because the data stored on cloud servers are in encrypted format. To overcome this issue, a new system is proposed which makes use of SVM classifier for classification purpose. This system also provide the security and confidentiality to user’s data and their requested queries and also hide the data access patterns. The performance of system is evaluated by using car dataset. Also system performance is compared with KNN classifier. Experimental results prove that the classification on encrypted data on cloud is better in terms of time, memory and accuracy, with SVM rather than KNN classifier.

REFERENCES

1. M. Kumar, J. Meena, R. Singh and M. Vardhan, "Data outsourcing: A threat to confidentiality, integrity, and availability," Green Computing and Internet of Things (ICGCIoT), 2015 International Conference on, Noida, 2015, pp. 1496-1501.
2. B. K. Samanthula, Y. Elmehdwi and W. Jiang, "k-Nearest Neighbor Classification over Semantically Secure Encrypted Relational Data," in IEEE Transactions on Knowledge and Data Engineering, vol. 27, no. 5, pp. 1261-1273, May 1 2015.
3. S. Thurner, M. Grün, S. Schmitt and H. Baier, "Improving the Detection of Encrypted Data on Storage Devices," IT Security Incident Management & IT Forensics (IMF), 2015 Ninth International Conference on, Magdeburg, 2015, pp. 26-39.
4. Y. Fu; H. Xiong; X. Lu; J. Yang; C. Chen, "Service Usage Classification with Encrypted Internet Traffic in Mobile Messaging Apps," in *IEEE Transactions on Mobile Computing* , vol.PP, no.99, pp.1-1
5. M. D. Singh, P. R. Krishna and A. Saxena, "A privacy preserving Jaccard similarity function for mining encrypted data," TENCON 2009 - 2009 IEEE Region 10 Conference, Singapore, 2009, pp. 1-4.
6. F. Liu, W. K. Ng and W. Zhang, "Encrypted SVM for Outsourced Data Mining," *Cloud Computing (CLOUD), 2015 IEEE 8th International Conference on*, New York City, NY, 2015, pp. 1085-1092.
7. Y. Rahulamathavan, R. C. W. Phan, S. Veluru, K. Cumanan and M. Rajarajan, "Privacy-Preserving Multi-Class Support Vector Machine for Outsourcing the Data Classification in Cloud," in IEEE Transactions on Dependable and Secure Computing, vol. 11, no. 5, pp. 467-479, Sept.-Oct. 2014.
8. J. Xu, W. Zhang, C. Yang, J. Xu and N. Yu, "Two-Step-Ranking Secure Multi-Keyword Search over Encrypted Cloud Data," Cloud and Service Computing (CSC), 2012 International Conference on, Shanghai, 2012, pp. 124-130.
9. H. Hu, J. Xu, C. Ren and B. Choi, "Processing private queries over untrusted data cloud through privacy homomorphism," Data Engineering (ICDE), 2011 IEEE 27th International Conference on, Hannover, 2011, pp. 601-612.
10. Y. Huang, J. Katz and D. Evans, "Quid-Pro-Quo-tocols: Strengthening Semi-honest Protocols with Dual Execution," Security and Privacy (SP), 2012 IEEE Symposium on, San Francisco, CA, 2012, pp. 272-284.