Classification of Encrypted Data with Elliptic Curve Cryptography

Brindha.M[1], Dr.R.Nedunchezhiyan[2]
PG scholar, Department of Computer Science and Engineering, KIT-KalaignarKarunanidhi Institute of Technology, Coimbatore, India[1]
Vice Principal, KIT- KalaignarKarunanidhi Institute of Technology, Coimbatore, India[2]

***Abstract* - Data mining, the mining of hidden predictive information from huge databases, is a powerful new technology with grand latent to help companies focus on the most vital information in their data warehouses. Data mining tools expect future trends and behaviors, allowing businesses to make positive, knowledge-driven decisions. Text classification is the method of conveying text documents based on assured categories. Due to the rising trends in the field of internet and computers, billions of text data are processed at a known time and so there is a require for systematize these data to offer easy storage and accessing .Many text classification approaches were developed for efficiently solving the difficulty of identifying and classifying these data. During the data retrieval of the classified data, privacy, security, accuracy and time consuming is the challenging task. In this paper, classification after encryption has been applied for security issue and support vector machine classification has been compared with k Nearest Neighbor classification for accuracy and time consumption issue. A classifier is used to define the suitable class for each text document based on the input algorithm used for classification. Encryption is the procedure of encoding messages or information in such a way that only approved parties can read it, which provides high security and accuracy.**

***Keywords* - privacy of data, classifier, encryption, k-NN, text document**

.

I.INTRODUCTION

Today's digital infrastructure supports innovative ways of storing, processing, and disseminating data. In fact, we can store our data in remote servers, access reliable and efficient services provided by third parties, and use computing power available at multiple locations across the network. Furthermore, the growing adoption of portable devices together with the diffusion of wireless connections in home and work environments has led to a more distributed computing scenario. These advantages come at a price of higher privacy risks and vulnerabilities as a huge amount of information is being circulated and stored, often not under the direct control of its owner.

Ensuring proper privacy and protection of the information stored, communicated, processed, and disseminated in the cloud as well as of the users accessing such an information is one of the grand challenges of our modern society. As a matter of fact, the advancements in the Information Technology and the diffusion of novel paradigms such as data outsourcing and cloud computing, while allowing users and companies to easily access high quality applications and services, introduce novel privacy risks of improper information disclosure and dissemination.

Classification is one of the commonly used tasks in data mining applications. For the past decade, due to the rise of various privacy issues, many theoretical and practical solutions to the classification problem have been proposed under different security models. However, with the recent popularity of cloud computing, users now have the opportunity to outsource their data, in encrypted form, as well as the data mining tasks to the cloud. Since the data on the cloud is in encrypted form, existing privacy-preserving classification techniques are not applicable.

In cryptography, encryption is the process of encoding messages or information in such a way that only authorized parties can read it. Encryption does not of itself prevent interception, but denies the message content to the interceptor. In an encryption scheme, the intended communication information or message, referred to as plaintext, is encrypted using an encryption algorithm, generating cipher text that can only be read if decrypted. For technical reasons, an encryption scheme usually uses a pseudorandom encryption key generated by an algorithm. It is in principle possible to decrypt the message without possessing the key, but, for a well-designed encryption scheme, large computational resources and skill are required. An authorized recipient can easily decrypt the message with the key provided by the originator to recipients, but not to unauthorized interceptors.

1.1 Advantages

Cloud computing relies on sharing of resources to achieve coherence and economies of scale similar to a utility (like the electricity grid) over a network. At the foundation of cloud computing is the broader concept of converged infrastructure and shared services.

The cloud also focuses on maximizing the effectiveness of the shared resources. Cloud resources are usually not only shared by multiple users but are also dynamically re-allocated per demand. This can work for allocating resources to users. For example, a cloud computer facility, which serves European users during European business hours with a specific application (e.g. email) while the same resources are getting reallocated and serve North American users during North America's business hours with another application (e.g. web server). This approach should maximize the use of computing powers thus reducing environmental damage as well since less power, air conditioning, rackspace, etc. is required for a variety of functions.

The term "moving to cloud" also refers to an organization moving away from a traditional CAPEX model (buy the dedicated hardware and depreciate it over a period of time) to the OPEX model (use a shared cloud infrastructure and pay as you use it).

Proponents claim that cloud computing allows companies to avoid upfront infrastructure costs, and focus on projects that differentiate their businesses instead of infrastructure. Proponents also claim that cloud computing allows enterprises to get their applications up and running faster, with improved manageability and less maintenance, and enables IT to more rapidly adjust resources to meet fluctuating and unpredictable business demand.

II.OVERVIEW

Existing work on Privacy-Preserving Data Mining (either perturbation or secure multi-party computation based approach) cannot solve the DMED problem. Perturbed data do not possess semantic security, so data perturbation techniques cannot be used to encrypt highly sensitive data. Also the perturbed data do not produce very accurate data mining results. Secure multi-party computation (SMC) based approach assumes data are distributed and not encrypted at each participating party.

- Computation Cost is quite high because we need to compute distance of each query.
- Distance based learning is not clear which type of distance to use and which attribute to use to produce the best results.
- They have poor run-time performance when the training set is large.
- It is very sensitive to irrelevant or redundant features because all features contribute to the similarity and thus to the classification.
- Data perturbation techniques cannot be used to encrypt highly sensitive data.
- Do not produce very accurate data mining results.

III.PROPOSED SYSTEM

The *k*-nearest neighbor algorithm is amongst the simplest of all machine learning algorithms. A novel secure and efficient scheme for k-NN query on encrypted cloud data in which the key of data owner to encrypt and decrypt outsourced data will not be completely disclosed to any query user. Therefore, our scheme can efficiently support the secure k-NN query on encrypted cloud data even when query users are not trustworthy enough. We emphasize that the intermediate results seen by the cloud in our protocol are either newly generated randomized encryptions or random numbers. Thus, which data records correspond to the k-nearest neighbors and the output class label are not known to the cloud. To achieve this, we first present a two-party protocol for the exponential mechanism. This protocol can be used as a sub protocol by any other algorithm that requires the exponential mechanism in a distributed setting. Furthermore, we propose a two-party algorithm that releases differentially-private data in a secure way according to the definition of secure multiparty computation.

3.1 Data collection and Encryption

The data owner has a collection of n files to outsource onto the cloud server in encrypted form and expects the cloud server to provide keyword retrieval service to data owner himself or other authorized users. To achieve this, the data owner needs to build a searchable index from a collection of keywords extracted out of files, and then outsources both the encrypted index and encrypted files onto the cloud server.

Each file which is to be uploaded is encrypted with encryption key. Once file is encrypted, next step is to upload it to the storage system along with data decryption key. Owner specifies the set of attributes for access structure, it then encrypts the file. Finally, owner uploads encrypted file and encryption key and set of attributes to the storage system. While data owner uploading the encrypted file, they also upload set of attributes. The data owner gives the attributes of the receiver while sending the file to the receiver; the file gets encrypted as per the given attributes. Thus the attributes for specified file is to be distributed and decryption key for decrypting the file are to be distributed to the data users.

The encryption has been done using the elliptic curve cryptography using paillier cryptosystem. In that algorithm the key has been generated and encryption is done with that key.

Key generation

- Choose two large prime numbers *p* and *q* randomly and independently of each other such that $\gcd(pq,(p-1)(q-1))=1$. This property is assured if both primes are of equal length.
- Compute $n=pq$ and $\lambda=\operatorname{lcm}(p-1,q-1)$.
- Select random integer g where $g\in\mathbb{Z}^*_{n^2}$
- Ensure n divides the order of g by checking the existence of the following modular multiplicative inverse: $\mu=(L(g^{\lambda}\bmod n^2))^{-1}\bmod n$,

where function L is defined as $L(u) = \frac{u-1}{n}$.

- The public (encryption) key is (n, g).
- The private (decryption) key is (λ, μ).

If using p,q of equivalent length, a simpler variant of the above key generation steps would be to set $g = n+1, \lambda = \varphi(n)$, and $\mu = \varphi(n)^{-1} \bmod n$, where $\varphi(n) = (p-1)(q-1)$.

Encryption

- Let m be a message to be encrypted where $m \in \mathbb{Z}_n$.
- Select random r where $r \in \mathbb{Z}_n^*$.
- Compute ciphertext as: $c = g^m \cdot r^n \bmod n^2$.

3.2 User Control

The main idea of this module is to design the user interface for users in the project. The login page is to design for data owner and data user. After the data owner logins into the system, the page displayed which allows the data owner to achieve the encrypted file upload to the system. When the user logins to the system, the system allows the user to input the decryption key and attributes for retrieval of specified file. Before accessing the file from system, the user must register into the system.

This module can be also used to register users for custom modules that support personalization and user specific handling. If the users wish to create their own user accounts, i.e. register, then registration checks for the username availability and assign unique ID. User Control means controlling the login with referring the username and password which are given during the registration process.

3.3 Classification Process

This module is for collecting the data records with splitting of data from Index. The support vector machine is a training algorithm for learning classification and regression rules from data. SVMs are currently among the best performers for a number of classification tasks ranging from text to genomic data. SVMs can be applied to complex data types by designing kernel functions for such data.

3.4 Data Retrieval

The data user is authorized to process keyword retrieval over the outsourced data. The computing power on the user side is limited, which means that operations on the user side should be simplified. For privacy consideration, which keywords the data user has searched must be concealed. In our process, we are mining the data from the database very efficiently. In the database, first it will classify all the data based on the contexts and if the user requests for any queries then the mining process will be done using SVM. The resulted output will produce accurate data.

IV.CONCLUSIONS

To protect user privacy, various privacy-preserving classification techniques have been proposed over the past decade. The existing techniques are not applicable to outsourced database environments where the data resides in encrypted form on a third-party server. This paper proposed a non-preserving k-NN classification protocol over encrypted data in the cloud. Our protocol protects the confidentiality of the data, user's input query, and hides the data access patterns. We also evaluated the performance of our protocol under different parameter settings. Since improving the efficiency of SMINn is an important first step in improving the performance of our PPkNN protocol we plan to investigate alternative and more efficient solutions to the SMINn problem in our future work. Also, we will investigate and extend our research to other classification algorithms.

V.FUTURE ENHANCEMENTS

This paper proposed a novel privacy-preserving k-NN classification protocol over encrypted data in the cloud. Our protocol protects the confidentiality of the data, user's input query, and hides the data access patterns. We also evaluated the performance of our protocol under different parameter settings. Since improving the efficiency of SMINn is an important first step in improving the performance of our PPkNN protocol, we plan to investigate alternative and more efficient solutions to the SMINn problem in our future work. Also,we will investigate and extend our research to other classification algorithms.

REFERENCES

[1] Bharath K. Samanthula, Yousesf Elmehdwi, and Wei Jiang, "k-Nearest Neighbor Classification Over Semantically Secure Encrypted Relational Data", ieee transaction, vol 27,no 5,p 1261-1273, 2015.
[2] A. Shamir, "How to share a secret," Commun. ACM, vol. 22,pp. 612–613, 1979.
[3] C. Gentry, "Fully homomorphic encryption using ideal lattices," in Proc. 41st Annu. ACM Sympos. Theory Comput., 2009, pp. 169–178.
[4] C. Gentry and S. Halevi, "Implementing gentry's fully-homomorphic encryption scheme," in Proc. 30th Annu. Int. Conf. Theory Appl. Cryptographic Techn.: Adv. Cryptol., 2011, pp. 129–148.
[5] D. Bogdanov, S. Laur, and J. Willemson, "Sharemind: A framework for fast privacy-preserving computations," in Proc. 13th Eur.Symp. Res. Comput. Security: Comput. Security, 2008, pp. 192–206.
[6] H. Hu, J. Xu, C. Ren, and B. Choi, "Processing private queries over untrusted data cloud through privacy homomorphism," in Proc. IEEE 27th Int. Conf. Data Eng., 2011, pp. 601–612.
[7] J. Camenisch and M. Michels, "Proving in zero-knowledge that a number is the product of two safe primes," in Proc. 17th Int. Conf. Theory Appl. Cryptographic Techn., 1999, pp. 107–122.

[8] M. Bohanec and B. Zupan. (1997). The UCI KDD Archive[Online]. Available: http://archive.ics.uci.edu/ml/datasets/Car+Evaluation.
[9] P. Paillier, "Public key cryptosystems based on composite degree residuosity classes," in Proc. 17th Int. Conf. Theory Appl. CryptographicTechn., 1999, pp. 223–238.
[10] P. Williams, R. Sion, and B. Carbunar, "Building castles out of mud: Practical access pattern privacy and correctness on untrustedstorage," in Proc. 15th ACM Conf. Comput. Commun. Security, 2008, pp. 139–148.
[11] P. Mell and T. Grance, "The NIST definition of cloud computing (draft)," NIST Special Publication, vol. 800, p. 145, 2011.
[12]R. Agrawal and R. Srikant, "Privacy-preserving data mining," ACM Sigmod Rec., vol.29, pp. 439–450, 2000.
[13] S. De Capitani di Vimercati, S. Foresti, and P.Samarati, "Managing and accessing data in the cloud: Privacy risks and a approaches," in Proc. 7th Int.Conf. Risk Security Internet Syst., 2012, pp. 1–9.
[14] Y. Huang, J. Katz, and D. Evans, "Quid-Pro-Quo-tocols: Strengthening semi honest protocols with dual execution," in Proc. IEEE Symp. Security Privacy, 2012, pp. 272–284.
[15] Y. Huang, D. Evans, J. Katz, and L. Malka, "Faster secure twoparty computation using garbled circuits," in Proc. 20th USENIX Conf. Security, 2011, pp. 35–35.