K-NEAREST NEIGHBOR CLASSIFICATION OVER ENCRYPTED RELATIONAL DATA

Ms.Gadekar R.R.[1], Mr.R.S.Bhosale[2]

[1] *Ms.Gadekar R.R.,Student,M.E.,Information Technology,AVCOE,Sangamner,Maharashtra,India*
[2] *Mr.R.S.Bhosale,Assistant Professor,Information Technology,AVCOE,Maharashtra,India*

ABSTRACT

Data Mining is widely used in bank, medical, scientific areas. In Data mining, classification is used. Many solutions are derived for classification problem under different models, because many problems occurred regarding privacy issues. Users can outsource their data, in encrypted form, as well as the data mining tasks to the cloud because cloud computing becomes popular. Previous privacy-preserving classification techniques can not be used, because data on cloud is encrypted. In this paper classification over encrypted data is concerned firstly. In this paper a secure and simple k-NN classifier over encrypted data with paillier cryptosystem in the cloud is determined. Paillier cryptosystem is secured as compare to RSA.This protocol protects the confidentiality, privacy and hides the data access patterns.

Keyword: - *Security, privacy, k-NN classifier, databases, encryption*

1.INTRODUCTION

The Internet becomes progressively useful tool in our day to day life, both professional and personal, as its users are becoming more and more everyday. One of the best revolutionary concept is Cloud Computing. The Cloud constitutes hardware and software that are provided as a service over the Internet. Extraction of hidden extrapolative information from large databases, is a powerful new technology with high potential to help companies concentration on the most information in their data warehouses, like data mining . Techniques and applications related to data mining are very much needed in the cloud computing model. Data Mining in cloud computing is the method of extracting structured information from unstructured or semi-structured web data sources. Ignoring these tremendous advantages that the cloud offers, privacy and security issues in the cloud tells companies not to utilize those advantages. When data are highly sensitive, the data need to be encrypted earlier outsourcing to the cloud. However, when data are encrypted, irrespective of the basic encryption scheme, performing any data mining tasks becomes very challenging without ever decrypting the data.

1.1. K-NN Algorithm:-

It is the nearest neighbour algorithm. The *k*-nearest neighbour's algorithm is a technique for classifying objects based on the next training data in the feature space. It is among simplest of all mechanism learning algorithms [30]. The algorithm operates on a set of d-dimensional vectors, $D = \{\mathbf{xi} \mid i = 1. . . N\}$, where $\mathbf{xi} \in kd$ denotes the *i* th data point. The algorithm is initialized by selection *k* points in *kd* as theinitial *k* cluster representatives or "centroids". Techniques for select these primary seeds include sampling at random from the dataset, setting them as the solution of clustering a small subset of the data or perturbing the global mean of the data *k* times. Then the algorithm iterates between two steps till junction:

Step 1: Data Assignment each data point is assign to its adjoining centroid, with ties broken arbitrarily. This results in a partitioning of the data.

Step 2: Relocation of "means". Each group representative is relocating to the center (mean) of all data points assign to it. If the data points come with a possibility measure (Weights), then the relocation is to the expectations (weighted mean) of the data partitions.

"Kernelize" *k*-means though margins between clusters are still linear in the embedded high-dimensional space, they can become non-linear when projected back to the original space, thus allowing kernel *k*-means to deal with more

complex clusters. The *k*-medoid algorithm is similar to *k*-means except that the centroids have to belong to the data set being clustered. Fuzzy c-means is also similar, except that it computes fuzzy membership functions for each clusters rather than a hard one.

1.2 Paillier Cryptosystem

The scheme is an perfect example of an Homomorphic encryption which means given only the public-key and the encryption of m1 and *m*2, one can compute the encryption of *m*1 + *m*2.

Key-Gen Algorithm

1) Choose two prime numbers p & q and calculation =p*q and λ = lcm(p-1,q-1) such that gcd(p*q,(p-1)*(q-1)) = 1
2) Select g Є Z*n^2 and calculate μ = (L(g^λ mod n^2))^(-1) mod n where L(x) = x-1/n
3) n, g acts as a public key
4) λ, μ acts as a private key

Encryption Algorithm:

1) Let m Є Zn be the message
2) Choose r Є Z*n
3) Required Cipher text is c = g^m *r^n mod n^2

Decryption Algorithm:

1) Compute m = L(c^λ mod n^2)*μ mod n

The homomorphic properties is used for secure electronic voting and electronic cash.
This algorithm provides security against chosen-plain text attack.

2. Literature Survey

In [2], Pascal Paillier propose a new trapdoor mechanism and derive from this technique three encryption schemes : a trapdoor permutation and two homomorphic probabilistic encryption schemes computationally comparable to RSA. This cryptosystems, based on usual modular arithmetics, are provably secure under appropriate assumptions in the standard model.

In [3], Clifton presents a method for privately computing *k* − *nn* classification from distributed sources without revealing any information about the sources or their data, other than that revealed by the final classification result.

In [4], Mikhail J. Atallah adapt their techniques to also solve the general multi-step k-NN search, and describe a specific embodiment of it for the case of sequence data. The protocols and correctness proofs can be extended to suit other privacy-preserving data mining tasks, such as classification and outlier detection.

In [5], The novelty of Sven Laur's solution is in the choice of the secret sharing scheme and the design of the protocol suite. They have made many practical decisions to make large-scale share computing feasible in practice. The protocols of SHAREMIND are information-theoretically secure in the honest-but-curious model with three computing participants. Although the honest-but-curious model does not tolerate malicious participants, it still provides significantly increased privacy preservation when compared to standard centralised databases.

2.1 Survey of papers

No	Author	Advantages	Disadvantages	Technique Used
1.	Bharath K. Samanthulla, Yousef Elmehdwi and Wei Jiang	Protects data confidentiality, user's query privacy, and hides data access patterns.	Computation Cost is quite high and poor run time performance.	Privacy Preserving k-NN Classification
2	Peter Williams, Radu Sion and Bogdan	Efficiency and privacy was improved.	Computational complexity and privacy leak.	Oblivious data access protocol.

	Carbunar			
3	Craig Gentry	Security level is good and ciphertext scheme for use keys.	Computational complexity and untrusted server scheme.	Fully homomorphic encryption scheme.
4	Craig Gentry and Shai Halevi	space-efficient and running-time advantage of the optimization.	Ciphertext size is high.	Key-generation method for the homomorphic Encryption.
5	Dan Bogdanov, Sven Laur, and Jan Willemson	Easy to use application development interface and performance improved.	Did not providing security guarantees against active adversaries and application programmer's interface.	Sharemind protocol
6	Haibo Hu, Jianliang Xu, Chushi Ren, Byron Choi	Improve the efficiency of the query processing and performance.	Did not provide Mutual privacy protection for queries on senior Unstructured datasets.	Processing Private Queries over Untrusted Data Cloud through Privacy Homomorphism

3.Architectural View

Here we are using Simple KNN,KNN with distance and paillier cryptosystem for implementation.

3.1. Simple KNN:-

1.When patient is going to find out nearest blood bank in which required blood is present,he can find out it on the basis of Rh Factor.Rh factor is related to blood.

2.To Find out blood group,from Rh value simple knn algorithm is used.

3.2. KNN with distance:-

1.Patient find out nearest location from it's original location by this algorithm.

3.3.Paillier cryptosystem:-

1.This algorithm is used for encryption.

2.When patient request for data,his request is executed by TPA.

3.TPA send data in encrypted form,to the patient by using this algorithm.

So here patient and TPA are working node each time.and Blood bank is only activated for adding updated data of particular bank.Architecture is shown in following figure.

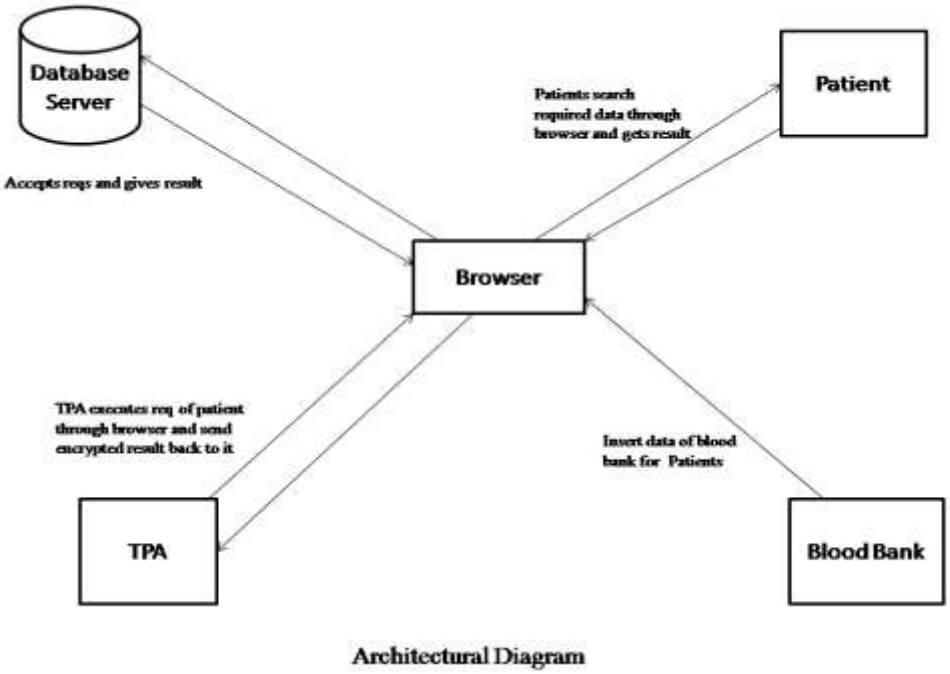

Figure 1: Architecture diagram

4. CONCLUSIONS

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 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.we will investigate and extend our research to other classification algorithms.

5. REFERENCES

[1] B. K. Samanthula, Y. Elmehdwi, and W. Jiang, "k-nearest neighbor classification over semantically secure encrypted relational data." eprint arXiv:1403.5001, 2014.

[2]. P. Paillier, "Public key cryptosystems based on composite degree residuosity classes," in *Eurocrypt*, pp. 223–238, 1999.

[3]. M. Kantarcioglu and C. Clifton, "Privately computing a distributed k-nn classifier," in *PKDD*, pp. 279–290, 2004.

[4]. Y. Qi and M. J. Atallah, "Efficient privacy-preserving k-nearest neighbor search," in *IEEE ICDCS*, pp. 311–319, 2008.

[5] D. Bogdanov, S. Laur, and J. Willemson, "Sharemind: A framework for fast privacy-preserving computations," in *ESORICS*, pp. 192–206, Springer, 2008.

[6]P. Mell and T. Grance, ―The NIST definition of cloud computing (draft),‖ NIST Special Publication, vol. 800, p. 145, 2011.

[7] S. De Capitani di Vimercati, S. Foresti, and P. Samarati, ―Managing and accessing data in the cloud: Privacy risks and approaches,‖ in Proc. 7th Int. Conf. Risk Security Internet Syst., 2012, pp. 1–9. 1272 IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 27, NO. 5, MAY 2015

[8] P. Williams, R. Sion, and B. Carbunar, ―Building castles out of mud: Practical access pattern privacy and correctness on untrusted storage,‖ in Proc. 15th ACM Conf. Comput. Commun. Security, 2008, pp. 139–148

[9] C. Gentry, "Fully homomorphic encryption using ideal lattices," in Proc. 41st Annu. ACM Sympos. Theory Comput., 2009, pp. 169– 178.

[10] R. Agrawal and R. Srikant, "Privacy-preserving data mining," ACM Sigmod Rec., vol. 29, pp. 439–450, 2000.

[11]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.

[12]Y. Lindell and B. Pinkas, ―Privacy preserving data mining,‖ in Proc. 20th Annu. Int. Cryptol. Conf. Adv. Cryptol., 2000, pp. 36–54.