Scientific Journal of Impact Factor(SJIF): 3.134

e-ISSN(O): 2348-4470
p-ISSN(P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 2,Issue 11, November -2015

k-Nearest Neighbor Classification over Semantically Secure Encrypted Relational Data

Miss. Rani Ramchandra Sandbhor[1], Prof. Nilesh Mali[2]

[1,2]Siddhant Colleges of Engineering, Sudumbare, pune

ABSTRACT: *Data Mining has wide applications in many areas such as banking, medicine, scientific research and among government agencies. 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 this paper, we focus on solving the classification problem over encrypted data. In particular, we propose a secure k-NN classifier over encrypted data in the cloud. The proposed protocol protects the confidentiality of data, privacy of user's input query, and hides the data access patterns. To the best of our knowledge, our work is the first to develop a secure k-NN classifier over encrypted data under the semi-honest model. Also, we empirically analyze the efficiency of our proposed protocol using a real-world dataset under different parameter settings.*

I. INTRODUCTION

As of late, the cloud computing view is altering the associations' method for working their information especially in the way they store, get to and process information. As a developing computing worldview, cloud computing pulls in numerous associations to consider truly with respect to cloud potential regarding its expense productivity, adaptability, and offload of authoritative overhead. Regularly, associations appoint their computational operations notwithstanding their information to the cloud. In spite of colossal points of interest that the cloud offers, protection and security issues in the cloud are forestalling organizations to use those preferences. At the point when information are profoundly touchy, the information should be encoded before outsourcing to the cloud. Be that as it may, when information are scrambled, independent of the basic encryption plan, performing any information mining errands turns out to be exceptionally testing while never decoding the information. There are other protection concerns, showed by the accompanying illustration.

As developing processing world view, distributed computing draws in numerous associations to consider genuinely with respect to database potential as far as its cost efficiency adaptability and offload of regulatory overhead. Regularly, associations designate their computational operations notwithstanding their information to the cloud. Notwithstanding enormous preferences that the cloud offers, protection and security issues in the database are counteracting organizations to use those points of interest. At the point when information is exceedingly delicate, the information
should be encoded before outsourcing to the database. Nonetheless, when information is scrambled, independent of the fundamental encryption plan, performing any information mining undertakings turns out to be extremely testing while never decoding the information.

Assume an insurance agency outsourced its scrambled clients database and applicable information mining undertakings to a database. At the point when an operators from the organization needs to focus the danger level of a potential new client, the specialists can utilize an order strategy to focus the danger level of the client. Initially, the operators needs to create an information record q for the client containing certain individual data of the client, e.g., FICO assessment, age, conjugal status, and so on. At that point this record can be sent to the database, and the database will figure the class mark for q. All things considered, since q contains delicate data, to secure the client's protection, q ought to be encoded before sending it to the database. The above case demonstrates that information mining over encrypted information (indicated by DMED) on a database additionally needs to ensure a client's record when the record is a piece of an information mining procedure. In addition, database can likewise determine helpful and sensitive data about the genuine information things by watching the information access examples regardless of the fact that the information is encoded. Along with these, the protection/security necessities of the DMED issue on a database are triple:

1.Classification of the encoded information

2.Privacy of a client's question record
3.Concealing information access designs.

Literature Survey

Building Castles out of Mud: Practical Access Pattern Privacy and Correctness on Untrusted Storage

We present another reasonable instrument for remote information capacity with efficient access design security and accuracy. A stockpiling customer can convey this system to issue encoded peruses, composes, and supplements to a conceivably inquisitive what's more, vindictive stockpiling administration supplier, without uncovering data or access designs. The supplier is incapable to set up any relationship between's progressive gets to, or indeed, even to recognize a read and a compose. In addition, the customer is given with solid accuracy certifications to its operations – unlawful supplier conduct does not go undetected. We assembled a first down to earth framework – requests of greatness quicker than existing usage – that can execute more than a few questions for every second on 1Tbyte+ databases with full computational security.

Fully Homomorphic Encryption Using Ideal Lattices

We propose a completely homomorphic encryption plan – i.e., a plan that permits one to assess circuits over scrambled information without having the capacity to unscramble. Our answer comes in three stages. In the first place, we give a general result – that, to build an encryption plot that allows assessment of self-assertive circuits, it suffices to develop an encryption plan that can assess (marginally enlarged adaptations of) its own particular decoding circuit; we call a plan that can assess its (expanded) decoding circuit bootstrappable. Next, we depict an open key encryption plan utilizing perfect cross sections that is practically bootstrappable. Cross section based cryptosystems ordinarily have decoding calculations with low circuit many-sided quality, regularly ruled by an internal item calculation that is in NC1. Likewise, perfect cross sections give both added substance and multiplicative homomorphisms (modulo a open key perfect in a polynomial ring that is spoken to as a cross section), as expected to assess general circuits.

Implementing Gentry's Fully-Homomorphic Encryption Scheme

We portray a working usage of a variation of Gentry's completely homomorphic encryption plan (STOC 2009), like the variation utilized as a part of a prior usage effort by Smart what's more, Vercauteren (PKC 2010). Keen and Vercauteren executed the basic "fairly homomorphic" plan, yet were not ready to execute the bootstrapping usefulness that is expected to get the complete plan to work. We demonstrate various enhancements that permit us to execute all parts of the plan, including the bootstrapping usefulness. Our fundamental advancement is a key-era technique for the basic to some degree homomorphic encryption, that does not oblige full polynomial reversal. This lessens the asymptotic many-sided quality from ˜O(n 2:5) to ˜O(n 1:5) when working with measurement n cross sections (and for all intents and purposes diminishing the time from numerous hours/days to a few moments/minutes). Different improvements incorporate a clumping system for encryption, a watchful examination of the unscrambling's level polynomial, and some space/time exchange offs for the completely homomorphic plan. We tried our usage with cross sections of a few measurements, comparing to a few security levels.

Managing and Accessing Data in the Cloud: Privacy Risks and Approaches

Guaranteeing fitting security and assurance of the data put away, conveyed, prepared, and scattered in the cloud and in addition of the clients getting to such a data is one of the amazing difficulties of our present day society. As an issue of reality, the headways in the Information Technology and the dispersion of novel standards, for example, information outsourcing and cloud registering, while permitting clients and organizations to effortlessly get to excellent applications and administrations, present novel protection dangers of despicable data divulgence and dispersal. In this paper, we will portray distinctive parts of the protection issue in developing situations. We will represent dangers, arrangements, what's more, open issues identified with guaranteeing security of clients getting to administrations or assets in the cloud, delicate data put away at outside gatherings, and gets to such a data.

Public-Key Cryptosystems Based on Composite Degree Residuosity Classes

This paper researches a novel computational issue, specifically the Composite Residuosity Class Problem, and its applications to open key cryptography. We propose another trapdoor system and get from this system three encryption plans : a trapdoor change furthermore, two homomorphic probabilistic encryption conspires computationally practically identical to

RSA. Our cryptosystems, taking into account normal measured mathematics, are provably secure under fitting suspicions in the
standard model.

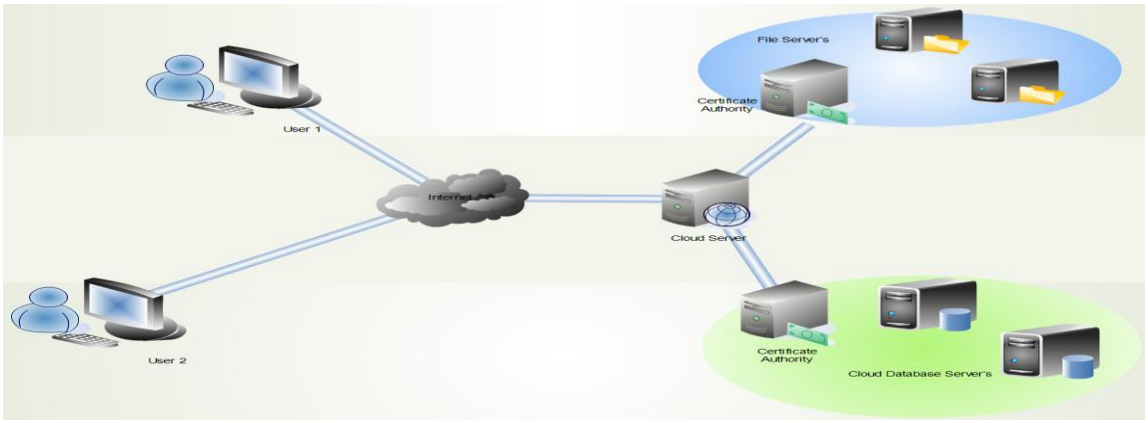
Objective, Aim, Problem Defination

Objective:
To implement securek-NN classifier over semantically secure encrypted data. In our protocol, once the encrypted data are outsourced to the cloud, sender does not participate in any computations. Therefore, no information is revealed to receiver. In addition, our protocol meets the following privacy requirements.

Aim:
The proposed convention means to ensure the secrecy of information, protection of client's info question, and conceals the information access designs. and also to build up a protected k-NN classifier over scrambled information under the semi-genuine model. Likewise, we experimentally break down the productivity of our proposed convention utilizing a certifiable data-set under diverse parameter settings.

Problem Definition:
Using encryption as a way to achieve the data confidentiality may cause another issue at the cloud during the query evaluation. The question here is "how can the cloud perform computations over encrypted data while the data stored are in encrypted form?" Along this direction, various techniques related to query processing over encrypted data. However, PPkNN is a more complex problem , so we are focused to design a framework with secure kNN queries over encrypted data

Architecture:



Algorithm/ Procedure/ Mathematical Model:

Algorithm:

The training examples are vectors in a multidimensional feature space, each with a class label. The training phase of the algorithm consists only of storing the feature vectors and class labels of the training samples.

In the classification phase, *k* is a user-defined constant, and an unlabeled vector (a query or test point) is classified by assigning the label which is most frequent among the *k* training samples nearest to that query point.

A commonly used distance metric for continuous variables is Euclidean distance. For discrete variables, such as for text classification, another metric can be used, such as the **overlap metric** (or Hamming distance). In the context of gene expression microarray data, for example, *k*-NN has also been employed with correlation coefficients such as Pearson and Spearman.[3] Often, the classification accuracy of *k*-NN can be improved significantly if the distance metric is learned with specialized algorithms such as Large Margin Nearest Neighbor or Neighbourhood components analysis.

Steps:

1. Store the output values of the M nearest neighbors to query scenario q in vector $r = \{r^1, \dots, r^M\}$ by repeating the following loop M times:
   a. Go to the next scenario s^i in the data set, where i is the current iteration within the domain $\{1, \dots, P\}$
   b. If q is not set or $q < d(q, s^i)$: $q \leftarrow d(q, s^i), t \leftarrow o^i$
   c. Loop until we reach the end of the data set (i.e. $i = P$)
   d. Store q into vector c and t into vector r
2. Calculate the arithmetic mean output across r as follows:

$$\bar{r} = \frac{1}{M}\sum_{i=1}^{M} r_i$$

3. Return $\bar{r}$ as the output value for the query scenario q

Mathematical Model:

System Description:

Let S be a system that de scribes k-NN Query Processing and classification over relational data base

S= {…..}

Identify input as I

S= {I,..}

Let I = {i1, i2, i3…, i_d}

The inputs will the files with different parameters and attributes.

Identify output as O

S = {I, O,…}

O= the user is able to search data when he is authenticated

Identify the processes as P

S= {I, O, P,…..}

P={C, S}

C= {parameter, id, Files}

S= {parameter, Skid, TreeIndex}

Identify the initial condition as Ic

S= {I,O,P,F,s,Ic,…}

Ic=user should always be online and authorized.

Success condition:

If ({ X,Y,Z})=({d0|Φ}∈D)
Then Success.

Failure condition:

If ({ X,Y,Z})≠({d0|Φ}∈D)

Then Failure.

Advantages:-

1.The cost of the learning process is zero.
2. No assumptions about the characteristics of the concepts to learn have to be done
3. Complex concepts can be learned by local approximation using simple procedures.

Limitations:

1. The model can not be interpreted (there is no description of the learned concepts.
2. techniques are not applicable to outsourced database environments where the data resides in encrypted form on a third-party server**.**

Future Scope:

Since enhancing the effectiveness of SMINn is an imperative initial step for enhancing the execution of our PPkNN convention, we plan to examine option and more proficient answers for the SMINn issue in our future work. Likewise, we will examine and extend our exploration to other characterization calculations.

CONCLUSION:

To ensure client security, different protection safeguarding arrangement systems have been proposed over the previous decade. The current strategies are not relevant to outsourced database situations where the information lives in encoded structure on an outsider server. This system will give novel protection safeguarding k-NN characterization convention over scrambled information in the database. Our system will secure the information's privacy, client's info inquiry, and conceals the information access designs. We likewise assessed the execution of our convention under distinctive parameter settings. Since enhancing the effectiveness of SMINn is a critical first stride for enhancing the execution of our PPkNN convention, we plan to research elective and more proficient answers for the SMINn issue in our future work. Additionally, we will examine and extend our exploration to other order calculations.

References:

[1] P.williams,R. Sion and B.Carbunar,"Building Castles out of Mud: Practical Access Pattern Privacy and Correctness on Untrusted Storage", in ACM CCS,pp.139-148,2008.

[2] C.Gentry,"Fully Homomorphic Encryption Using Ideal Lattices",in ACM STOC,pp. 169-178,2009.

[3] C.Gentry and S.Halevi,"Implementing Gentry's Fully- Homomorphic Encryption Scheme", in Eurocrypt, pp.129-148,Springer,Feb 2011.

[4] S.De Capitani di Vinercati, S.Foresti, and P.Samarati," Managing and Accessing Data in the Cloud: Privacy Risks and Approaches", inCRiSIS , pp.1-9,2012.

[5] P.Paillier," Public-Key Cryptosystems Based on Composite Degree Residuosity Classes",in Eurocrypt,pp.223-238,1999.

