ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 3, March 2016

K-Nearest Neighbor Classification over Semantically Secure Encrypted Relational Data

Chetan Wankhede[1], Nikhil Daundkar[2], Rohan Wadmare[3], Tushar Hinge[4], Prof. M.A.Ansari[5]

Student, Dept. of IT, Rajarshi Shahu College of Engineering, Tathawade, Savitribai Phule Pune University, Pune, India[1234]

Prof. Dept. of IT, Rajarshi Shahu College of Engineering, Tathawade, Savitribai Phule Pune University, Pune, India[5]

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 k-NN protocol protects the confidentiality of the data, user's input query, and 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 standard semi-honest model. Also, we empirically analyze the efficiency of our solution through various experiments.

KEYWORDS: Security, k-NN Classifier, Outsourced Databases, Encryption, privacy preserving.

I. INTRODUCTION

Recently, the cloud computing paradigm is revolutionizing the organizations' way of operating their data particularly in the way they store, access and process data. As an emerging computing paradigm, cloud computing attracts many organizations to consider seriously regarding cloud potential in terms of its cost-efficiency, flexibility, and offload of administrative overhead. Most often, organizations delegate their computational operations in addition to their data to the cloud. Despite tremendous advantages that the cloud offers, privacy and security issues in the cloud are preventing companies to utilize those advantages. When data are highly sensitive, the data need to be encrypted before outsourcing to the cloud. However, when data are encrypted, irrespective of the underlying encryption scheme, performing any data mining tasks becomes very challenging without ever decrypting the data.The data owner outsources his/her database and DBMS functionalities (e.g., kNN query) to an untrusted external service provider which manages the data on behalf of the data owner where only trusted users are allowed to query the hosted data at the service provider. By outsourcing data to an untrusted server, many security issues arise, such as data privacy (protecting the confidentiality of the data from the server as well as from query issuer). To achieve data privacy, data owner is required to use data Anonymization models (e.g., k-anonymity) or cryptographic (e.g., encryption and data perturbation) techniques over his/her data before outsourcing them to the server. Encryption is a traditional technique used to protect the confidentiality of sensitive data such as medical records. Due to data encryption, the process of query evaluation over encrypted data becomes challenging. Along this direction, various techniques have been proposed for processing range and aggregation queries over encrypted data. Using encryption as a way to achieve data confidentiality may cause another issue during the query processing step in the cloud. In general, it is very difficult to process encrypted data

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 3, March 2016

without ever having to decrypt it. The question here is how the cloud can execute the queries over encrypted data while the data stored at the cloud are encrypted at all times.

II. GOALS AND OBJECTIVE

1. Improving the efficiency of SMINn is an important first step for improving the performance of our PPkNN protocol.
2. Our protocol protects the confidentiality of the data, user's input query, and hides the data access patterns.
3. We also evaluated the performance of our protocol under different parameter settings.

III. MOTIVATION OF THE PROJECT

We motivated the PPKNN over encrypted data to achieve economies of scale for Cloud Computing. Then we introduced new security primitives, namely secure minimum (SMIN), secure frequency (SF), and proposed new solutions for them. Second, the work in did not provide any formal security analysis of the underlying sub-protocols. On the other hand, this paper provides formal security proofs of the underlying sub-protocols as well as the PPkNN protocol under the semi-honest model. We show that our proposed solution is secure and privacy-preserving, while correctly realizing the goal of PPKNN.

IV. EXISTING SYSTEM PROBLEM

Suppose Alice owns a database D of n records t1. . . tnand m + 1 attributes. Let ti,jdenote the jthattribute value of record ti. Initially, Alice encrypts her database attribute-wise, that is, she computes Epk(ti,j), for $1 \leq i \leq n$ and $1 \leq j \leq m+1$, where column (m+1) contains the class labels. We assume that the underlying encryption scheme is semantically secure . Let the encrypted database be denoted by D′. We assume that Alice outsources D′ as well as the future classification process to the cloud.Let Bob be an authorized user who wants to classify his input record q = hq1, . . . ,qmiby applying the k-Classification method based on D′. We refer to such a process as privacy-preserving k-NN (PPkNN) classification over encrypted data in the cloud. Formally, we define the PPkNN protocol as:
PPkNN(D′, q) → cq
wherecqdenotes the class label for q after applying k-NN classification method on D′ and q.

V. LITERATURE SURVEY

1. Project Title: Survey on Privacy Preserving Data Mining

From this paper We Referred

Data mining is the extraction of interesting patterns or knowledge from huge amount of data. In recent years, with the explosive development in Internet, data storage and data processing technologies, privacy preservation has been one of the greater concerns in data mining. A number of methods and techniques have been developed for privacy preserving data mining. This paper provides a wide survey of different privacy preserving data mining algorithms and analyses the representative techniques for privacy preserving data mining, and points out their merits and demerits. Finally the present problems and directions for future research are discussed.

2. Project Title: Proving in Zero-Knowledge that a Number Is the Product of Two Safe Primes.

From this paper We Referred

We present the reticent statistical zero-knowledge protocols to prove statements such as:

- A committed number is a prime.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 3, March 2016

- A committed (or revealed) number is the product of two safe primes, i.e., primes p and q such that (p − 1) =2and (q − 1) =2areprime.
- A given integer has large multiplicative order modulo a composite number that consists of two safe prime factors.

3. Project Title: Secure k-Nearest Neighbor Query over Encrypted Data in Outsourced Environments.

From this paper We Referred

For the past decade, query processing on relational data has been studied extensively, and many theoretical and practical solutions to query processing have been proposed under various scenarios. With the recent popularity of cloud computing, users now have the opportunity to outsource their data as well as the data management tasks to the cloud. However, due to the rise of various privacy issues, sensitive data (e.g., medical records) need to be encrypted before outsourcing to the cloud. In addition, query processing tasks should be handled by the cloud; otherwise, there would be no point to outsource the data at the first place. To process queries over encrypted data without the cloud ever decrypting the data is a very challenging task. In this paper, we focus on solving the k-nearest neighbor (kNN) query problem over encrypted database outsourced to a cloud: a user issues an encrypted query record to the cloud, and the cloud returns the k closest records to the user. We first present a basic scheme and demonstrate that such a naive solution is not secure. To provide better security, we propose a secure kNN protocol that protects the confidentiality of the data, user's input query, and data access patterns. Also, we empirically analyze the efficiency of our protocols through various experiments. These results indicate that our secure protocol is very efficient on the user end, and this lightweight scheme allows a user to use any mobile device to perform the kNN query.

4. Project Title: Managing and Accessing Data in the Cloud Privacy Risks and Approaches

From this paper We Referred

Ensuring proper privacy and protection of the information stored, communicated, processed, and disseminated in the cloud as well as of the users accessing such 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. In this paper, we will characterize different aspects of the privacy problem in emerging scenarios. We will illustrate risks, solutions, and open problems related to ensuring privacy of users accessing services or resources in the cloud, sensitive information stored at external parties, and accesses to such information.

5. Project Title: Privacy-preserving data mining in the malicious model

From this paper We Referred

Most of the cryptographic work in privacy-preserving distributed data mining deals with semi-honest adversaries, which are assumed to follow the prescribed protocol but try to infer private information using the messages they receive during the protocol. Although the semi-honest model is reasonable in some cases, it is unrealistic to assume that adversaries will always follow the protocols exactly. In particular, malicious adversaries could deviate arbitrarily from their prescribed protocols. Secure protocols that are developed against malicious adversaries require utilization of complex techniques. Clearly, protocols that can withstand malicious adversaries provide more security. However, there is an obvious trade-off: protocols that are secure against malicious adversaries are generally more expensive than those secure against semi-honest adversaries only. In this paper, our goal is to make an analysis of trade-offs between performance and security in privacy-preserving distributed data mining algorithms in the two models. In order to make a realistic comparison, we enhance commonly used sub protocols that are secure in the semi-honest model with zero knowledge proofs to be secure in the malicious model. We compare the performance of these protocols in both models.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 3, March 2016

VI. METHODOLOGY USED

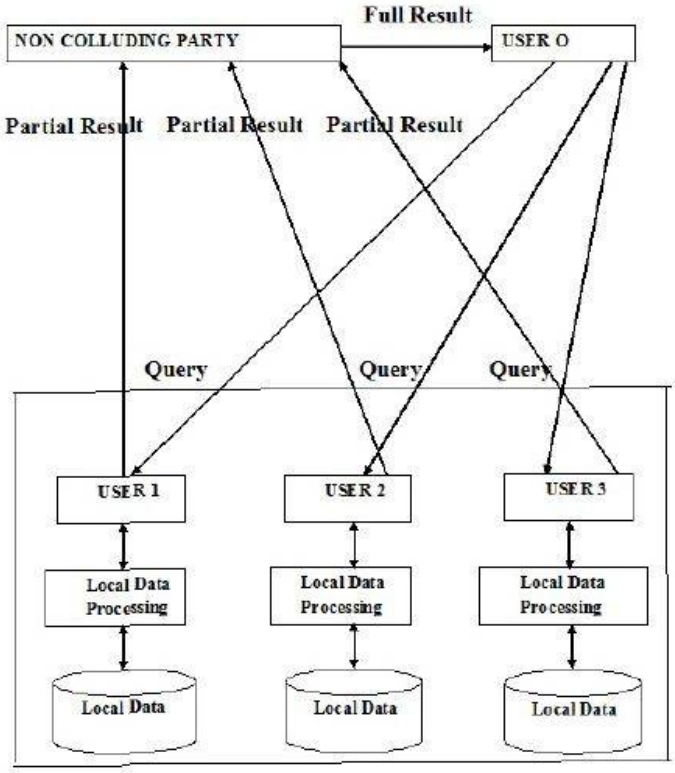


Fig No 01 System Architecture

1. Data confidentiality - Contents of T or any intermediate results should not be revealed to the cloud.
2. Query privacy - Bob's input query Q should not be revealed to the cloud.
3. Correctness - The output (t'1……t'k) should be revealed only to Bob. In addition, no information other thant'1…..t'kshould be revealed to Bob.
4. Low computation overhead on Bob - After sending his encrypted query record to the cloud, Bob involves only in a little computation compared with the existing works. More details are given in Section.
5. Hidden data access patterns - Access patterns to the data, such as the records corresponding to the k-nearest neighbors of Q, should not be revealed to Alice and the cloud (to prevent any inference attacks).

VII. ALGORITHM AND TECHNIQUE USED

- **Search algorithm.**

The search process of our DMRS scheme starts from the root node with a recursive procedure upon the tree in a special depth-first manner, which is called as "Greedy Depth first Traverse Strategy". Specifically, if the node's similarity score is less than or equal to the minimum similarity score of the currently selected top-k documents, search process returns to the parent node, otherwise, it goes down to examine the child node. The similarity score of each node u is calculated as Formula (1), i.e., the inner product of query vector Q and data vectorDu . This procedure is executed recursively until the objects with top- k scores are selected. The search can be done very efficiently, since only part of

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 3, March 2016

the index tree is visited due to the relatively accurate maximum score prediction. Algorithm 1 shows the process of our proposed search scheme.

- **Secure algorithm.**

As soon as the plaintext index tree is built, secure encryption scheme needs to be executedto prevent information leakage. We adopt the encryption scheme in [4] to secure our indextree, the whole process is described as follows:

- Setup
- GenIndex
- GenTrapdoor
- SimEvaluation

VIII. APPLICATIONS

It is used in cloud when we store the data on cloud in encrypted format. And access the data using secret key in decrypted form. We know the k-NN classifier and developed a privacy-preserving protocol for it over encrypted data.

- Increased website ROI

When users can find what they are looking for on a website easily they are more likely to take the desired action, whether that be a product purchase, information request, or simply learn what they wanted to know.

- Reduced customer service costs

Providing a self-service means to access common information on a website can reduce the number of calls or emails to customer service. In addition customer service can use the same search when answering questions.

- Increased productivity

If you are like many companies you have file shares full of documents, but aren't sure exactly what is there or where it is. By providing a quality search you can quickly locate both the documents you are looking for, as well as related documents that may already exist, preventing duplicate effort.

IX. EXPERIMENTAL SETUP AND RESULT

Here discussed some experiments demonstrating the performance of Privacy Preserving k-Nearest Neighbor (PPKNN) classification method with some parameter settings. The Partial Homomorphic encryption scheme is used as the underlying additive homomorphic encryption scheme and implemented the proposed PPkNN protocol in JAVA.

1 Dataset Details and Experimental Setup

This protocol used the Car Evaluation dataset from the UCI KDD archive[12]. This dataset consists of 1728 records (n = 1728) and 6 attributes (m = 6). There is a separate class attribute and the dataset is categorized into four different classes (w = 4). Encrypted this dataset attribute wise, using the Homomorphic encryption the key size is varied in experiments, and the encrypted data were stored on server machine. Based on PPkNN protocol, executed a randomquery over this encrypted data.

2 Performance of Privacy Preserving k-Nearest Neighbor Classification with Partial Homomorphic Encryption The encryption key size K is either 512 or 1024 bits if is K=512

bits, the computation cost varies from 9.98 to 46.16 minutes whenk is changed from 5 to 25, respectively. When K=1024 bits, the computation cost varies from 66.97 to 309.98 minutes when k is changed from 5 to 25, respectively. For K=512 bits, the computation time for Stage 2 to generate the final class label corresponding to the input query varies from 0.118 to 0.285 seconds when k is changed from 5 to 25. For K=1024 bits, Stage 2 took 0.789 and 1.89 seconds when k = 5 and 25, respectively. Here observed that the computation time of Stage 1 accounts for at least 99

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

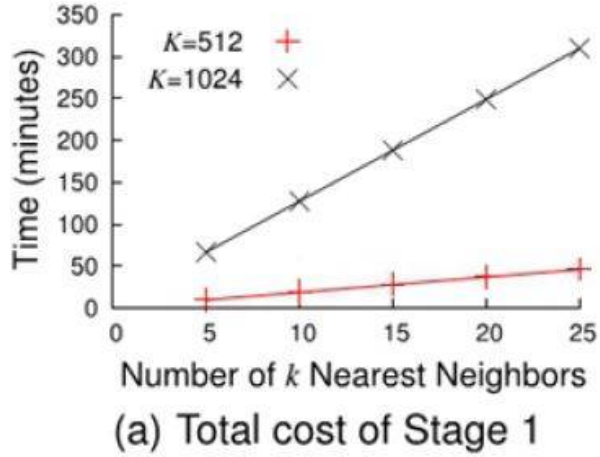
(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 3, March 2016

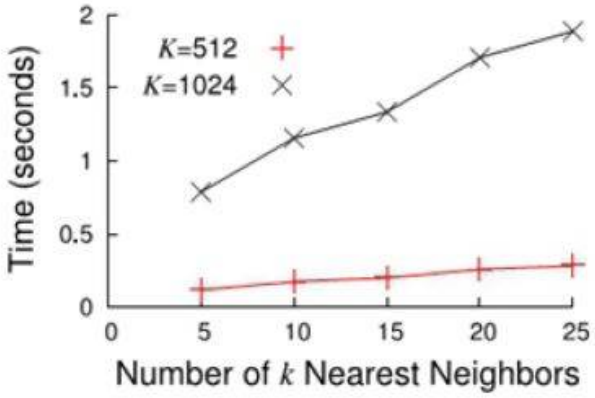
percentage of the total time in PPkNN. For example, when k = 10 and K=512bits, the processing costs of Stage 1 and 2 are 19.06 minutes and 0.175 seconds, respectively. Under this scenario, cost of Stage 1 is 99.98 percentage of the total cost ofPPkNN. The total computation time of PPkNN grows almost linearly with n and k.

X. RESULTS

As mentioned in Section 2.3, to formally prove that SMIN is secure under the semi-honest model, we need to show that the simulated image of SMIN is computationally indistinguishable from the actual execution image of SMIN. An execution image generally includes the messages exchanged and the information computed from these messages.



(a) Total cost of Stage 1



(b) Total cost of Stage 2

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 3, March 2016

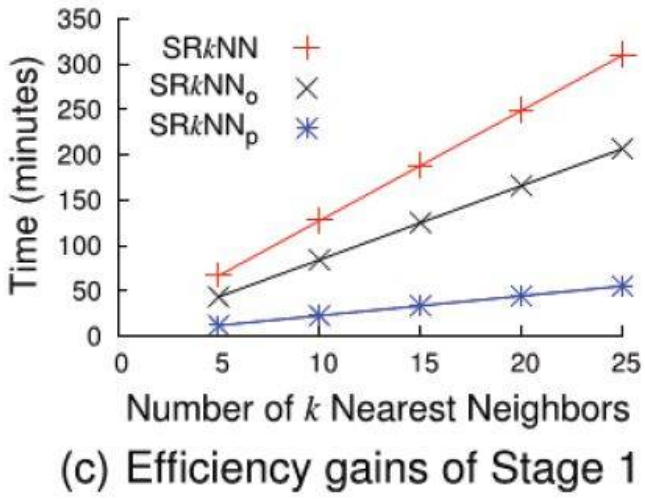


Fig No 02 Computation costs of PPkNN for varying number of k nearest neighbors and encryption key size K

XI. CONCLUSION

To protect user privacy, various privacy- preserving classification techniques have been proposed over the past decade. The existing techniques are not applicable to out sourced database environment s where the data resides in encrypted form on a third- party server. 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 pat terns. We also evaluated the performance of our protocol under different parameter set tings. Since improving the efficiency of SMIN n is an import ant first step f or improving the -performance of our PPkNN protocol, we plan to investigate alternative and more efficient solute- ions t o t he SMIN n problem in our future work. Also, we will investigate and extend our research to other classification algorithms.

REFERENCES

[1] Mell and T. Grance, "The nist definition of cloud computing(draft)," NIST special publication , vol. 800, p. 145, 2011.
[2] S. De Capitani di Vimercati, S. Foresti, and P. Samarati, "Managing and accessing data in the cloud: Privacy risks and approaches," in CRiSIS, pp. 1 –9, 2012.
[3] 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.
[4] P. Paillier, "Public key cryptosystems based on composite degree residuosity classes," in Eurocrypt , pp. 223–238, 1999.
[5] B. K. Samanthula, Y. Elmehdwi, and W. Jiang, "k-nearest neighbor classification over semantically secure encrypted re-lational data." eprint arXiv:1403.5001, 2014.
[6] C. Gentry, "Fully homomorphic encryption using ideal lattices," in ACM STOC , pp. 169–178, 2009.
[7] C. Gentry and S. Halevi, "Implementing gentry's fully- homomorphic encryption scheme," in EUROCRYPT , pp. 129– 148, Springer, 2011.
[8] A. Shamir, "How to share a secret,"Commun.ACM, vol. 22, pp. 612–613, 1979.
[9] 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.
[10] R. Agrawal and R. Srikant, "Privacy-preserving data mining,"ACMSigmod Rec., vol. 29, pp. 439–450, 2000.