Privacy-Preserving and K- Nearest Means Clustering over Relational Data

Kaipa Swetha, R.Suresh, Professor

Kaipa Swetha, PG Scholar, Dept. of CSE, CREC, Tirupati, AP, India
R.Suresh, .Professor, Dept. of CSE, CREC, Tirupati, AP, India

Abstract : Data Mining has wide use in many fields such as financial, medication, medical research and among govt. departments. Classification is one of the widely applied works in data mining applications. For the past several years, due to the increase of various privacy problems, many conceptual and realistic alternatives to the classification issue have been suggested under various protection designs. On the other hand, with the latest reputation of cloud processing, users now have to be able to delegate their data, in encoded form, as well as the information mining task to the cloud. Considering that the information on the cloud is in secured type, current privacy-preserving classification methods are not appropriate. In this paper, we concentrate on fixing the classification issue over encoded data. In specific, we recommend a protected k-classifier over secured data in the cloud. The suggested protocol defends the privacy of information, comfort of user's feedback query, and conceals the information access styles. To the best of our information, our task is the first to create a protected k-classifier over secured data under the semi-honest model. Also, we empirically evaluate the performance of our suggested protocol utilizing a real-world dataset under various parameter configurations. To secure user privacy, numerous privacy-preserving category methods have been suggested over the past several years. The current methods are not appropriate to contracted database surroundings where the information exists in secured form on a third-party server

Keywords: Security, k- classifier, outsourced databases, encryption

I. INTRODUCTION

Recently, the cloud computing paradigm [1] 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. There are other privacy concerns, demonstrated by the following example.

Example 1.

Suppose an insurance company outsourced its encrypted customers database and relevant data mining tasks to a cloud. When an agent from the company wants to determine the risk level of a potential new customer, the agent can use a classification method to determine the risk level of the customer. First, the agent needs to generate a data record q for the customer containing certain personal information of the customer, e.g., credit score, age, marital status, etc. Then this record can be sent to the cloud, and the cloud will compute the class label for q. Nevertheless, since q contains sensitive information, to protect the customer's privacy, q should be encrypted before sending it to the cloud.

The above example shows that data mining over encrypted data (denoted by DMED) on a cloud also needs to protect a user's record when the record is a part of a data mining process. Moreover, cloud can also derive useful and sensitive information about the actual data items by observing the data access patterns even if the data are encrypted [2], [3]. Therefore, the privacy/security requirements of the DMED problem on a cloud are threefold: (1) confidentiality of the encrypted data, (2) confidentiality of a user's query record, and (3) hiding data access patterns.

Existing work on privacy-preserving data mining (PPDM) (either perturbation or secure multi-party computation (SMC) 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 based approach assumes data are distributed and not encrypted at each participating party. In addition, many intermediate computations are performed based on non-encrypted data. As a result, in this paper, we proposed novel methods to effectively solve the DMED problem assuming that the encrypted data are outsourced to a cloud. Specifically, we focus on the classification problem since it is one of the most common data mining tasks. Because each classification technique has their own advantage, to be concrete, this paper concentrates on executing the k- nearest neighbor classification method over encrypted data in the cloud computing environment.

1.1ProblemDefinition

Suppose Alice owns a database D of n records $t_1, . . . ,t_n$ andm + 1 attributes. Let $t_{i,j}$denote the jth attribute value of record t_i. Initially, Alice encrypts her database attribute wise, that is, she computes $E_{pk}(t_{i,j})$, 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 [4]. 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 $q =<q_1, . . . ,q_m>$ by applying the k-NN 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:

$PPkNN(D', q) \rightarrow c_q$,

wherec_q denotes the class label for q after applying k- NN classification method on D' and q.

1.2OurContributions

In this paper, we propose a novel PPkNN protocol, a secure k-NN classifier over semantically secure encrypted data. In our protocol, once the encrypted data are outsourced to the cloud, Alice does not participate in any computations. Therefore, no information is revealed to Alice. In addition, our protocol meets the following privacy requirements:

- Contents of D or any intermediate results should not be revealed to the cloud.
- Bob's query q should not be revealed to the cloud.
- c_q should be revealed only to Bob. Also, no other information should be revealed to Bob.
- Data access patterns, such as the records corresponding to the k-nearest neighbors of q, should not be revealed to Bob and the cloud (to prevent any inference attacks).

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. In addition, after sending his encrypted query record to the cloud, Bob does not involve in any computations. Hence, data access patterns are further protected from Bob (see Section 5 for more details).

The rest of the paper is organized as follows. We discuss the existing related work and some concepts as a background in Section 2. A set of privacy-preserving protocols and their possible implementations are provided in Section 3. The formal security proofs for the mentioned privacy-preserving primitives are provided in Section 4. The proposed PPkNN protocol is explained in detail in Section 5. Section 6 discusses the performance of the proposed protocol under different parameter settings. We conclude the paper along with future work in Section 7.

II. RELATED WORK AND BACKGROUND

Due to space limitations, here we briefly review the existingrelated work and provide some definitions as a background. Please refer to our technical report [5] for a more elaborated related work and background. At first, it seems fully homomorphic cryptosystems (e.g., [6]) can solve the DMED problem since it allows a thirdparty(that hosts the encrypted data) to execute arbitrary functions over encrypted data without ever decrypting them. However, we stress that such techniques are very expensive and their usage in practical applications have yet to be explored. For example, it was shown in [7] that even for weak security parameters one -bootstrapping‖ operation of the homomorphic operation would take at least 30 seconds on a high performance machine. It is possible to use the existing secret sharing techniques in SMC, such as Shamir's scheme [8], to develop a PPkNNprotocol. However, our work is different from the secret sharing based solution in the following aspect. Solutions based on the secret sharing schemes require at least three parties whereas our work require only two parties. For example, the constructions based on Sharemind [9], a wellknown SMC framework which is based on the secret sharing scheme, assumes that the number of participating parties is three. Thus, our work is orthogonal to Sharemind and other secret sharing based schemes.

2.1Privacy-Preserving Data Mining

Agrawal and Srikant [4], Lindell and Pinkas [6] werethe first to introduce the notion of privacy-preserving under data mining applications. The existing PPDM techniques can broadly be classified into two categories: (i) data perturbation

2.2Query Processing over Encrypted Data

Various techniques related to query processing overencrypted data have been proposed, e.g., , [8]. However, we observe that PPkNN is a more complex problem

than the execution of simple kNN queries over encrypted data [2], [3]. For one, the intermediate k- nearest neighbors in the classification process, should not be disclosed to the cloud or any users. We emphasize that the recent method in [3] reveals the k- nearest neighbors to the user. Second, even if we know the k-nearest neighbors, it is still very difficult to find the majority class label among these neighbors since they are encrypted at the first place to prevent the cloud from learning sensitive information. Third, the existing work did no addressed the access pattern issue which is a crucial privacy requirement from the user's perspective.

In our most recent work [4], we proposed a novelsecure k-nearest neighbor query protocol over encrypted data that protects data confidentiality, user's query privacy, and hides data access patterns. However, as mentioned above, PPkNN is a more complex problem and it cannot be

and (ii) data distribution. Agrawal and Srikant [4] proposed the first data perturbation technique to build a decision-tree classifier, and many other methods were proposed later (e.g., [7], [4], [5]). However,

as mentioned earlier in Section 1, data perturbation techniques cannot be applicable for semantically secure encrypted data. Also, they do not produce accurate data mining results due to the addition of statistical noises to the data. On the other hand, Lindell and Pinkas [6] proposed the first decision tree classifier under the two- party setting assuming the data were distributed between them. Since then much work has been published using SMC techniques (e.g., [6], [7]). We claim that the PPkNNproblem cannot be solved using the data distribution techniques since the data in our case is encrypted and not distributed in plaintext among multiple parties. For the same reasons, we also do not consider secure k-NN methods in which the data are distributed between two parties (e.g., [8]).

solved directly using the existing secure k-nearest neighbor techniques over encrypted data. Therefore, in this paper, we extend our previous work in [4] and provide a new solution to the PPkNN classifier problem over encrypted data.

More specifically, this paper is different from our preliminary work [4] in the following four aspects. First, in this paper, we introduced new security primitives, namely secure minimum (SMIN), secure minimum out of n numbers (SMINn), secure frequency (SF), and proposed new solutions for them. Second, the work in [4] 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. Additionally, we discuss various techniques through which the proposed PPkNN protocol can possibly be extended to a protocol that is secure under the malicious setting. Third, our preliminary work in [4] addresses only secure kNNquery which is similar to Stage 1 of PPkNN. However, Stage 2 in PPkNN is entirely new.
Finally, our empirical analyses in Section 6 are based on a real dataset whereas the results in are based on a simulated dataset Furthermore, new experimental results are include in this paper

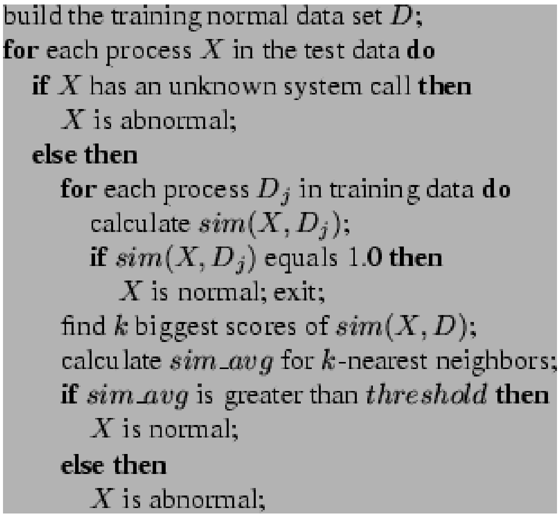
2.3ThreatModel

We adopt the security definitions in the literature of securemulti-party computation , and there are three common adversarial models under SMC: semi- honest, covert and malicious. In this paper, to develop secure and efficient protocols, we assume that parties are semi-honest. Briefly, the following definition captures the properties of a secure protocol under the semi-honest model

Definition 1. Let ai be the input of party P_i, $\prod_i(\pi)$ be P_i's execution image of the protocol p and b_i be the output for party $P_i$computed from p. Then, p is secure if $\prod_i(\pi)$can be simulated from a_i and b_i such that distribution of the simulated image is computationally indistinguishable from$\prod_i(\pi)$.

In the above definition, an execution image generally includes the input, the output and the messages communicated during an execution of a protocol. To prove a protocol is secure under semi-honest model, we generally need to show that the execution image of a protocol does not leak any information regarding the private inputs of participating parties .

KnnAlgorithm



```
build the training normal data set $D$;
for each process $X$ in the test data do
    if $X$ has an unknown system call then
        $X$ is abnormal;
    else then
        for each process $D_j$ in training data do
            calculate $sim(X, D_j)$;
            if $sim(X, D_j)$ equals 1.0 then
                $X$ is normal; exit;
        find $k$ biggest scores of $sim(X, D)$;
        calculate $sim\_avg$ for $k$-nearest neighbors;
        if $sim\_avg$ is greater than $threshold$ then
            $X$ is normal;
        else then
            $X$ is abnormal;
```

III. EXPERIMENT RESULTS

More specifically, this paper is different from our preliminary work [4] in the following four aspects. First, in this paper, we introduced new security primitives, namely secure minimum (SMIN), secure minimum out of n numbers (SMINn), 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. Additionally, we discuss various techniques through which the proposed PPkNN protocol can possibly be extended to a protocol that is secure under the malicious setting. Third, our preliminary work in addresses only secure kNNquery which is similar to Stage 1 of PPkNN. However, Stage 2 in PPkNN is entirely new. Finally, our empirical analyses in Section 6 are based on a real dataset whereas the results in are based on a simulated dataset. Furthermore, new experimental results are included in this paper

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 novel privacy-preserving k-NN classification protocol over encrypted data in the cloud. Our protocol protects the confidentialityof 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 for 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] P. 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 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

[3] 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.

[4] P. Paillier, -Public key cryptosystems based on composite degree residuosity classes,|| in Proc. 17thInt.Conf.Theory Appl. Cryptographic Techn., 1999, pp. 223–238.

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

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

[7] 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.

[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.